Defense Threat Reduction Agency
8725 John J. Kingman Road, MS
6201 Fort Belvoir, VA 22060-6201

DTRA-TR-23-10

TECHNICAL REPORT

Characterizing and Modeling Radiation Effects in Neuromorphic Computing Paradigm

Distribution Statement A. Approved for public release; distribution is unlimited.

November 2022

HDTRA1-17-1-0038

Prepared by: Hugh J. Barnaby

Arizona State University, 975 S. Myrtle Ave,Tempe AZ 85281

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
11-2022	Final Report	20200420	20210930

4. TITLE AND SUBTITLE

Characterizing and Modeling Radiation Effects in Neuromorphic Computing Paradigm

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	HDTRA1-17-1-0038	
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)

Hugh J. Barnaby

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

Arizona State University,
975 S. Myrtle Ave,
Tempe, AZ 85281

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)
Defense Threat Reduction Agency 8725 John J Kingman Rd #6201 Fort Belvoir, VA 22060	DTRA	DTRA-TR-23-10

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A. Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The objective of this program was to perform basic research on radiation effects in chip-scale cognitive computers inspired by neuromorphic architectures

15. SUBJECT TERMS

Radiation Effects, Neuromorphic, Computing, Paradigm, Total Ionizing Dose, TID, CMOS, non-volatile resistive memory, RRAM, radiation response, modeling

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	150

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Jacob Calkins	571-616-5946

PREVIOUS EDITION IS OBSOLETE.

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE.
Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE.
State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED.
Indicate the time during which the work was performed and the report was written.

4. TITLE.
Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER.
Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER.
Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER.
Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER.
Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER.
Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER.
Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/ FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

DTRA Basic Research Final Report

Grant/Award #: HDTRA1-17-1-0038
PI Name: Hugh J. Barnaby
Organization/Institution: Arizona State University
Project Title: Characterizing and Modeling Radiation Effects in Neuromorphic Computing Paradigm

What are the major goals of the project?
List the major goals of the project as stated in the approved application or as approved by the agency. If the application lists milestones/target dates for important activities or phases of the project, identify these dates and show actual completion dates or the percentage of completion. Generally, the goals will not change from one reporting period to the next. However, if the awarding agency approved changes to the goals during the reporting period, list the revised goals and objectives. Also explain any significant changes in approach or methods from the agency approved application or plan.

Objectives

Overview

The objective of this program was to perform basic research on radiation effects in chip-scale cognitive computers inspired by neuromorphic architectures. Total ionizing dose and effects are characterized using experimentally validated and radiation-enabled device and circuit models. System-level radiation effects are investigated in several architectures that utilize a variety of neuro-inspired learning algorithms. The sensitivity of the neuromorphic system to radiation effects are quantified by experiments on existing neuromorphic hardware parts and large-scale modeling of the neuromorphic architectures.

The program focuses on two neuromorphic platforms: 1) fully CMOS systems and 2) hybrid systems composed of both CMOS and non-volatile resistive memory (RRAM) devices. Radiation response characterization, modeling and hardening of these platforms was achieved through a hierarchical device, circuit, and architecture-based co-design methodology, which incorporates experimentally validated models for radiation effects. Hybrid neuromorphic architectures consist of CMOS neurons communicating with each other in a parallel system across RRAM based synaptic devices. These systems offer the promise of reduced power consumption, increased density, as well as the potential for increased radiation hardness.

The research team, composed of investigators at Arizona State University (ASU) and Sandia National Laboratories (SNL), has utilized its capabilities to support radiation effects research on neuromorphic computing systems, which has been leveraged to characterize and model radiation effects in neuromorphic systems via a hierarchical device-circuit-architecture co-design methodology, as illustrated in Fig. 1.

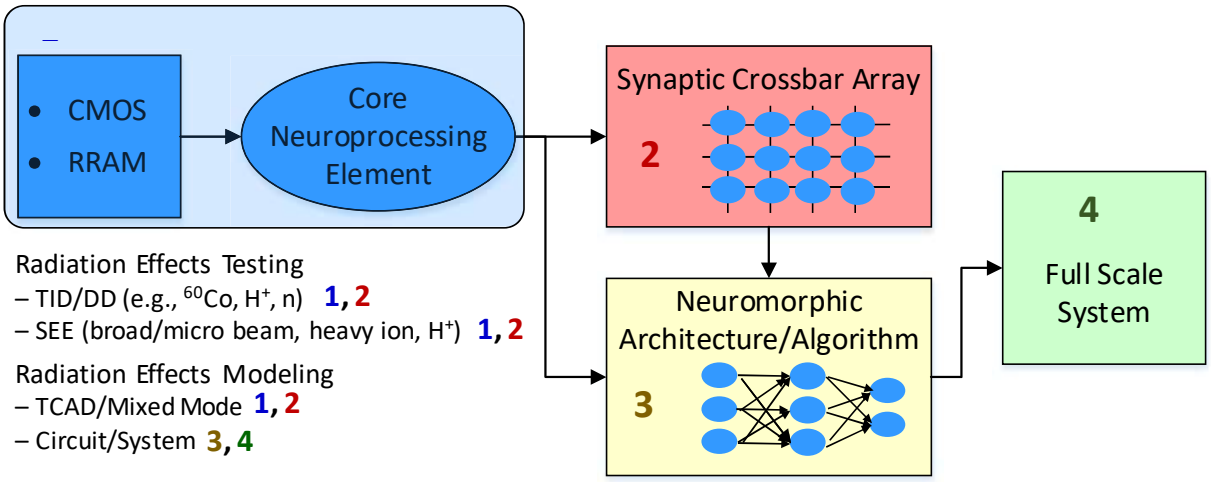


Fig. 1. Hierarchical device-circuit-architecture co-design approach for characterizing and modeling radiation effects in neuromorphic computing systems.

In this program, we endeavored to address the following questions:

1. What are the primary radiation threats to the neuromorphic computing systems?
2. What are the better hardware platforms to implement the synaptic arrays, SRAM-based or RRAM based, in terms of radiation hardness?
3. Would the neuron nodes or the synaptic weights be more susceptible to the radiation effects?
4. Would the radiation effects impact more on the training stage or on the testing (read) stage of the neuro-inspired learning algorithm?

In order to answer these questions, the program was organized around four primary tasks:

- **Task 1. Device-level Radiation Effects on Neuro-processing Elements**
- **Task 2. Array-level Testing/Simulation of Radiation Effects on Neuro Crossbar**
- **Task 3. Architecture-level Modeling Radiation Effects in Neuromorphic Systems**
- **Task 4. Radiation Effects on Full-scale Neuromorphic System**

Fig. 2 is a Gantt chart of the program over the three base years.

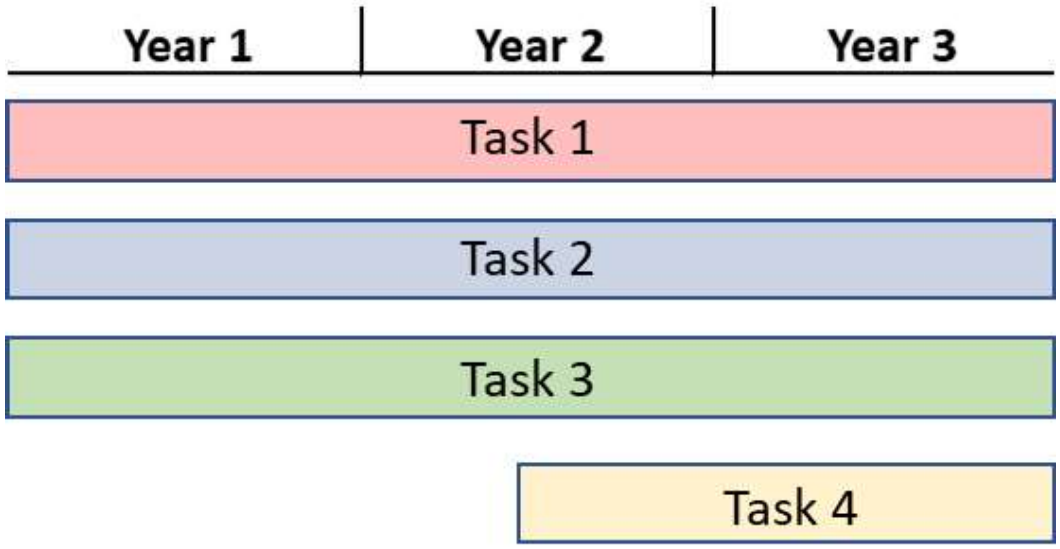


Fig. 2 Program Gantt chart

Milestones and Deliverables

Task 1

Task 1.1 Development of radiation effects models for CMOS devices and circuits

- Design, fabricate, and test CMOS test-structures representing the devices used in core neuro-processing elements.
- Perform TID on CMOS devices.
- Build radiation-enabled compact models for CMOS transistors.

Task 1.2 Characterize radiation effects and build models for synaptic weight tuning in RRAM devices

- Design, fabricate, and test RRAM devices.
- Perform TCAD simulation to model RRAM.
- Perform TID tests on RRAM

Task 2

Task 2.1 Pre-rad testing of neuro-crossbar RRAM array

- Design and fabrication 256×256 1T1R RRAM Array
- Electrical testing of 256×256 1T1R RRAM Array

Task 2.2 Modeling and validation of radiation effects in core neuro-processing elements

- Design full un-hardened RRAM-based integrated neuromorphic system.

Task 2.3 Simulation of radiation effects in SRAM array with peripheral neuron circuits

- See Task 3.3

Task 3

Task 3.1 Implement neuro-inspired algorithms on neuromorphic hardware platforms

- Develop the software codes for evaluating deep feed-forward multi-layer perceptron network.

Task 3.2 Post-rad testing of neuro-crossbar RRAM array (with peripheral neuron circuits)

- Perform TID tests on 256×256 1T1R RRAM Array

Task 3.3 Incorporate device and circuit-level radiation models into deep neural network

- Model TID effects in SRAM-based multi-layer perceptron network.

Task 4

Task 4.1 Fabrication, and test unhardened CMOS neuromorphic system

- Fabricate multi-layer perceptron (vector matric multiply – VMM) neuromorphic system without radiation hardening techniques.
- Perform tests to evaluate susceptibility to TID

What was accomplished under these goals?
For this reporting period describe: 1) major activities; 2) specific objectives; 3) significant results, including major findings, developments, or conclusions (both positive and negative); and 4) key outcomes or other achievements. Include a discussion of stated goals not met. As the project progresses, the emphasis in reporting in this section should shift from reporting activities to reporting accomplishments.

Task 1. Device-level Radiation Effects on Neuro-processing Elements

Task 1.1 Development of radiation effects models for CMOS devices and circuits

Design, fabricate, and test CMOS test-structures representing the devices used in core neuro-processing elements and perform TID on CMOS devices.

Several variants of CMOS structures were designed and fabricated during the three-year program. The CMOS technologies assessed were the 350nm bulk process from ON Semiconductor, the 90nm bulk technology node by Winbond Electronics Corporation, and the Global Foundries 22 nm fully depleted SOI (22FDX) technologies.

ON Semiconductor 0.35 μm bulk CMOS

0.35 μm bulk devices and circuits were fabricated on test chips designed at ASU. Fig. 3 is a picture of the ON test chip layout. The ON ASIC chip includes access to an 8x8 resistor array (discussed later), large NMOS transistors in a chain configuration, diode total ionizing dose (TID) test structures, as well as CMOS circuits (discussed later) designed to characterize the full neuromorphic system. TID characterization was performed at ASU and at the Air Force Research Laboratory in Albuquerque NM. Eight wafers of the CMOS chips were fabricated at the ON-semiconductor plant in Pocatello ID. Fig. 4 is a photomicrograph of the final die.

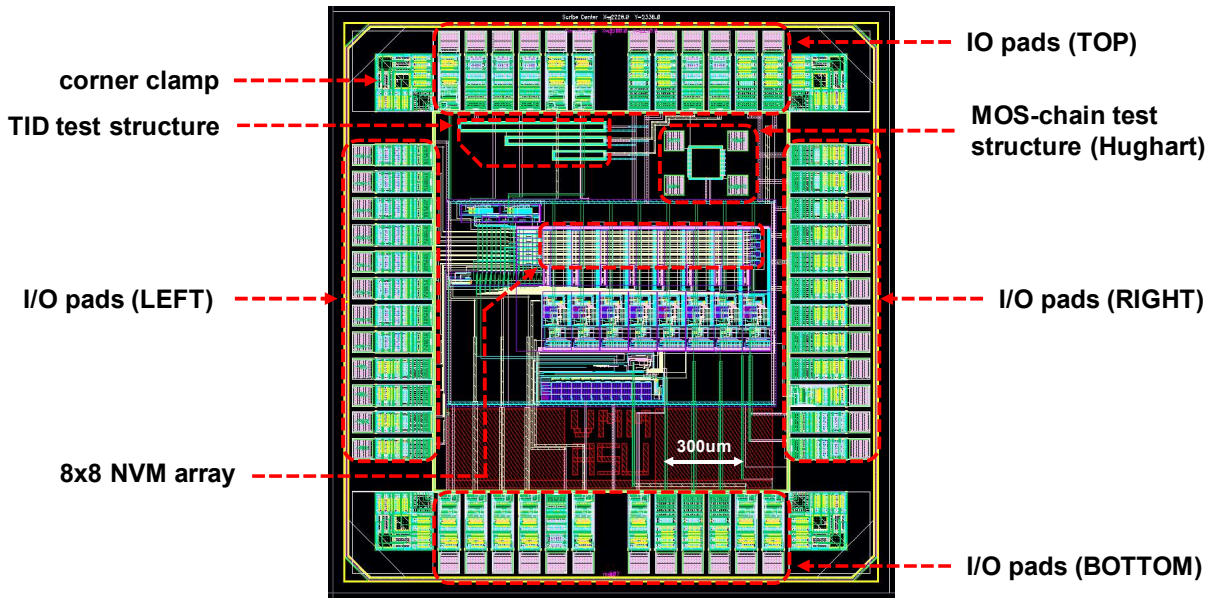


Fig. 3. ON Semiconductor chip layout

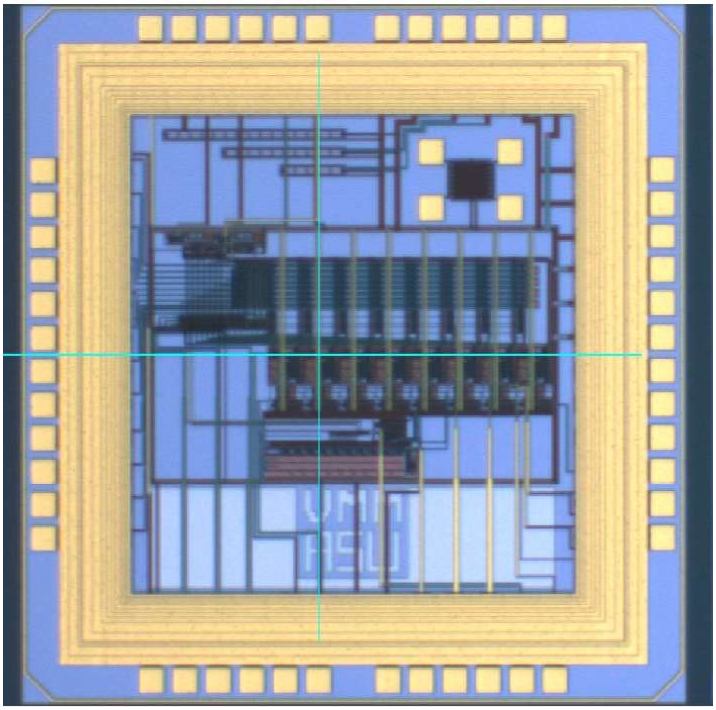


Fig. 4. Die photomicrograph of fabricated ON Semiconductor Chip chip.

Forward curves from the diodes and the IV characteristics of the MOSFETs were characterized on the ON ASIC pre- and post-irradiation. The pre-rad diode curves are shown in Fig. 5. These Fig.s show the forward bias diode from 0V to 3.2V (the current compliance was set to 10mA). The sweep voltage was applied to the anode during the forward bias sweep. During these sweeps the cathode was grounded. The MOSFET pre-rad IV curve is shown in Fig. 6. The curve is a standard I_D vs V_{GS} curve, with the drain connected to 3.2V, the source and body connected to ground, and the gate voltage swept from 0V to 3.2V.

A set of ON ASIC chips were irradiated under bias. We irradiated them using the Cobalt-60 source at ASU. Fig. 7 shows the MOSFET post-rad data. The MOSFETs showed very little change in IV response up to 3.5Mrad(Si). This was attributed to the fact that the MOSFET test structures were extremely wide. Thus, TID-induced edge leakage current in the devices was very small relative to the as-drain device current, which scaled to the width. It is also interesting the threshold voltage shifts for these parts is negligible, which suggest that, even for this older process, the gate oxide is not degraded by the exposures.

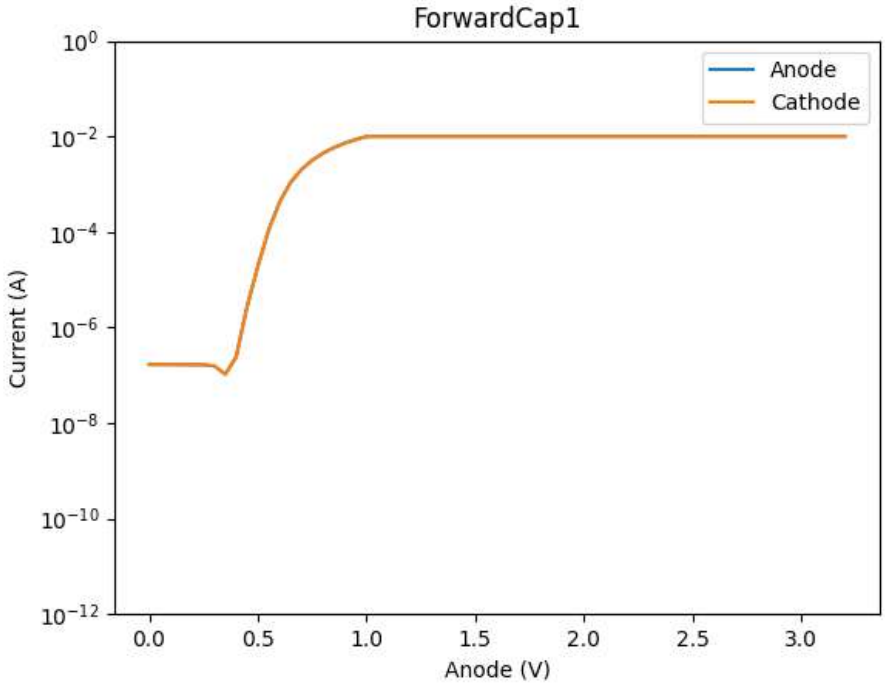


Fig. 5. Forward bias sweep of diode and resulting current. The cathode was held at 0V, and the anode was swept from 0 to 3.2 V. Note that the current levels out and hit a maximum at 10 mA which is the compliance current of the measurement.

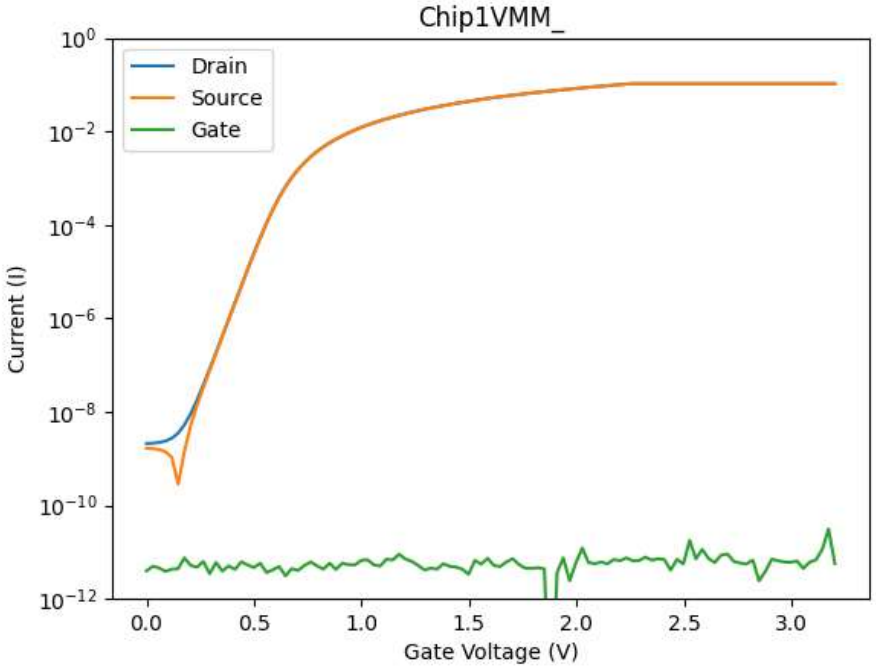


Fig. 7. IV Characteristics of MOSFET on ASIC. The drain is connected to 3.2V, the source and body connected to ground, and the gate voltage swept from 0V to 3.2V.

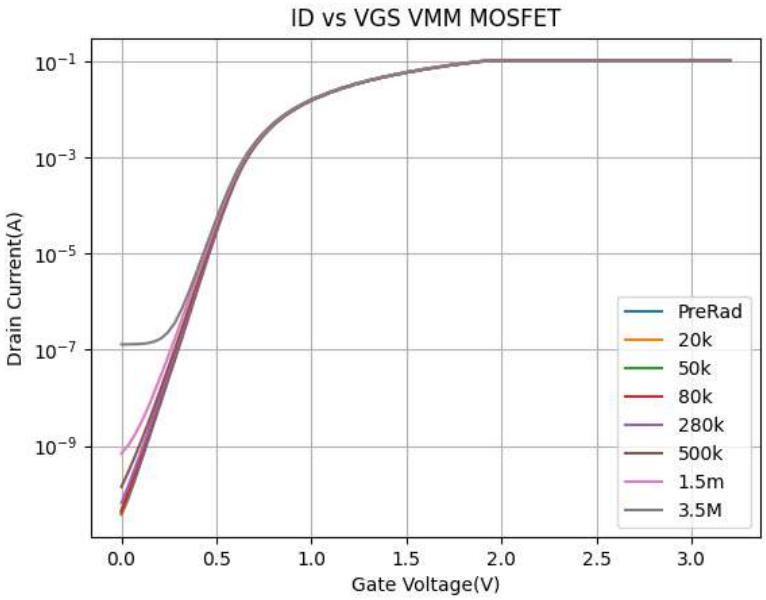


Fig. 8. MOSFET ID vs VGS curve for TID radiation up to a dose level of 3.5Mrad.

Global Foundries - 22nm fully depleted silicon-on-insulator (FDSOI) devices

Test devices for the Global Foundries 22 nm FDSOI process (22FDX) were obtained by ASU in bare coupon form. Fig. 9 shows the rectangular coupons used by the ASU team for this program. The transistor devices tested are wired out to interior probe pads on the coupon as shown in Fig. 10 and Fig. 11 shows the pad layout for each row consisting of seven devices with shared well contacts.

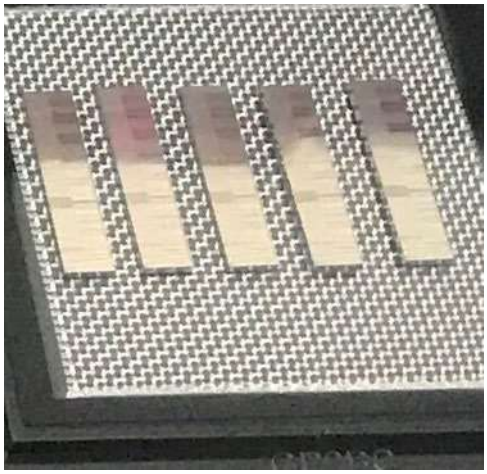


Fig. 9. Rectangular die coupon for 22 FDX transistors

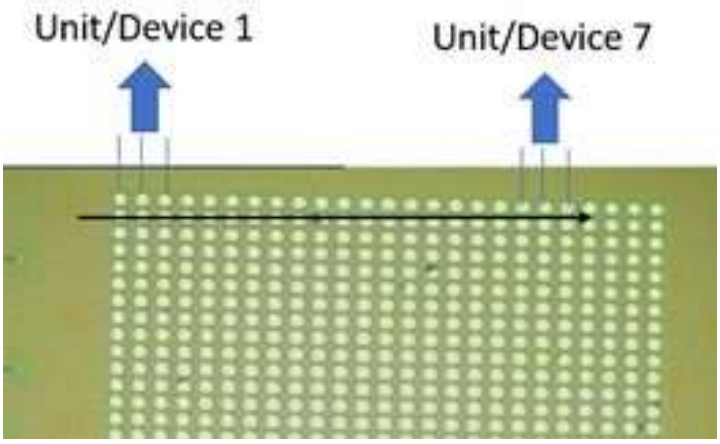


Fig. 10. Probe pad array on 22FDX test coupons.

A notable characteristic of these devices is the buried oxide that isolates the fully depleted silicon under the gate oxide of each device. The expected result of exposure to the ionizing radiation is a generation of holes trapped in relatively thick BOX layer. Charge trapped in the BOX can be electrostatically coupled to the front gate, causing a shift in transistor I-V curves and threshold voltage. The front gate oxide in the devices is sufficiently thin that negligible trapped charges will build up there, so any change in I-V characteristics are likely due to radiation-induced defects in isolating oxides like the BOX.

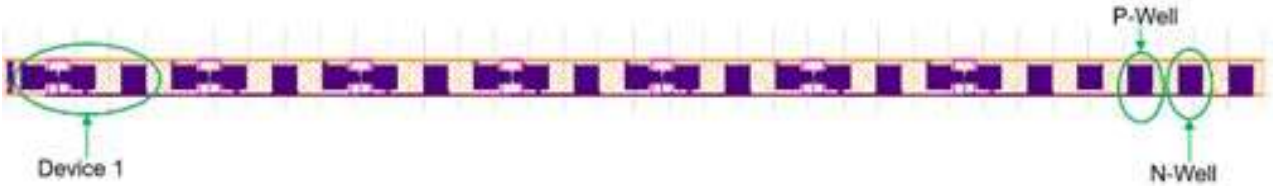


Fig. 11. Device pad layout for each row consisting of seven devices with shared well contacts

Various rows consisting of PFET and NFET devices were tested using a Keithley 4200 (Fig. 12) and a Keysight B1500A parameter analyzer (Fig. 12) connected to a probe card (Fig. 13) . Tables I -III list the different device variants characterized. Note that we used data from one of the variants to build radiation aware compact models for used in neuromorphic circuit simulation (discussed later). Cross-sections of n- and p-channel MOSFET variants. Fig. 14 shows the flipped well devices and Fig. 15 shows the regular well devices. Available devices also have different gate oxide thickness, i.e., thick (enhanced) and standard, with various widths and lengths. Three threshold voltage variants were evaluated: super low threshold voltage (SLVT), high threshold voltage (HVT), and regular threshold voltage (RVT). It may be noted that the SLVT transistors have on average higher threshold voltages than the RVT and HVT variants. This may be due to the fact that the SLVT devices are fabricated in the flipped well configuration and have thick gates oxides, which increase the V_{TH}.

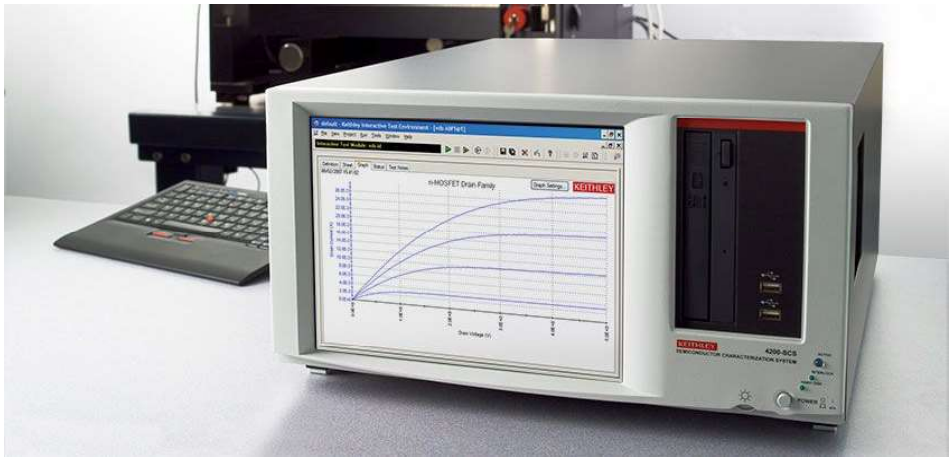


Fig. 12. Keithley 4200-SCS

Fig. 13. Probe card with coaxial connections.

Table I

Device parameters for the SLVT type devices

Parameter	PFET Value	NFET Value
Well Type	Flipped	Flipped
Gate Oxide	Thick	Thick
V_{TH} Type	SLVT	SLVT
V_{TH} Value(V)	-0.448	0.292
Length(µM)	0.15	0.15
Width(µM)	0.16,0.25, 0.5, 1, 2, 5	0.16, 0.25, 0.5

Table III

Device parameters for the HVT type devices

Parameter	PFET Value	NFET Value
Well Type	Regular	Regular
Gate Oxide	Standard	Standard
V_{TH} Type	HVT	HVT
V_{TH} Value(V)	-0.424	0.328
Length(µM)	0.02	0.02
Width(µM)	0.17, 0.3,1,2	0.08, 0.12

Table IIIII

Device parameters for the RVT type devices

Parameter	PFET Value	NFET Value
Well Type	Regular	Regular
Gate Oxide	Standard	Standard
V_{TH} Type	RVT	RVT
V_{TH} Value(V)	-0.256	0.232
Length(µM)	0.02	0.02
Width(µM)	1	0.08, 0.12

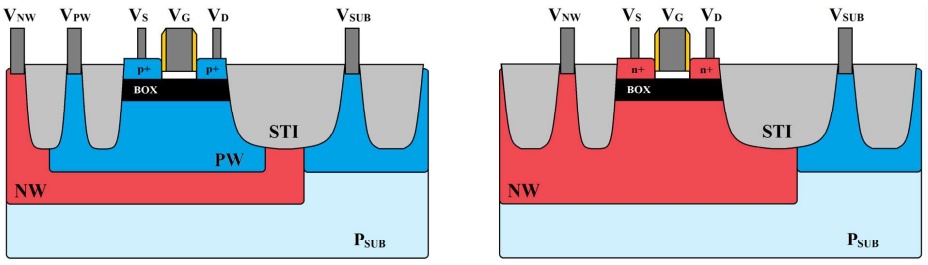


Fig. 14. Typical structure layout of flipped well PFET and NFET devices.

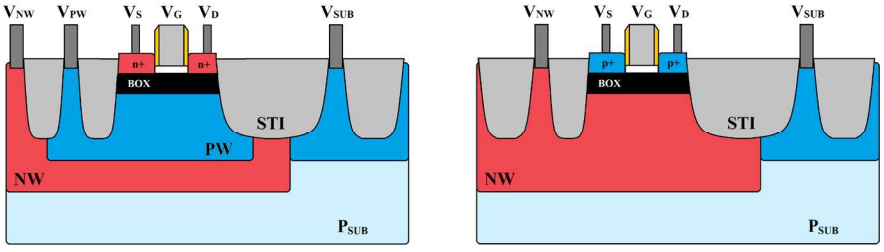


Fig. 15. Typical structure layout of regular well NFET and PFET devices.

For each transistor type a pre-radiation voltage sweep was performed to confirm correct pad landing contact for the probe card, device functionality, and to establish the reference characteristic behavior for each of the devices for the following irradiation exposure and analysis. During parameter measurement the gate voltage was swept from -0.2V to 1V for the NFET devices and 0.2V to -1V for the PFET devices. The drain was biased to 1V for the NFET and -1V for the PFET devices while the source and base were grounded. The drain current vs. gate voltage sweeps were conducted with a semiconductor parameter analyzer. A change in threshold voltage induced by TID exposures can be calculated by taking the difference in gate voltage at a constant drain current.

The bias conditions during TID irradiations consisted of applying -1V to the gate while grounding the drain, source, and bulk of each of the PFET devices, i.e., “ON-state” biasing. The bias conditions during irradiation for the NFET devices consisted of applying 1V to the drain and source while grounding the gate and bulk, i.e., “Transmission Gate” (TG) biasing. These have been shown to be the worst-case biasing for the FD-SOI technology. Each device was irradiated and tested at radiation levels between 20k and 625k rad[Si].

TID testing was performed at the Air Force Research Laboratories (AFRL) facility at the Kirtland Air Force Base in Albuquerque, NM. The ARACOR x-ray system at AFRL was conFig.d with a probe card to enable the characterization of two rows (24 pads/row) of transistors on the GF 22FDX test chips. Fig. 16 shows the setup of the ARACOR outfitted with the probe card used for TID testing.

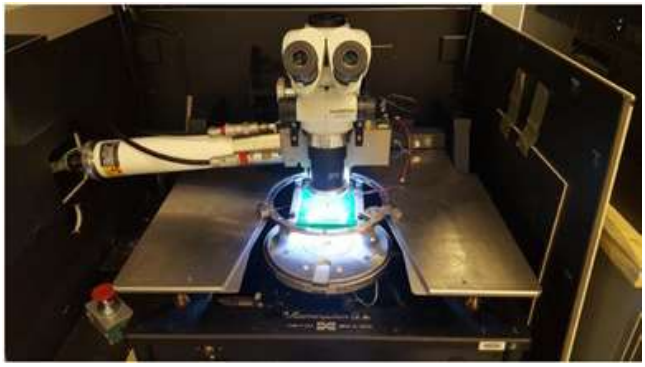


Fig. 16. ARACOR X-Ray system with custom probe card.

A notable characteristic of these devices is the buried oxide that isolates the fully depleted silicon under the gate oxide of each device. The expected result of exposure to the ionizing radiation is a generation of holes trapped in relatively thick buried oxide (BOX) layer. Charge trapped in the BOX can be electrostatically coupled to the front gate, causing a shift in transistor I-V curves and threshold voltage. The front gate oxide in the devices is sufficiently thin that negligible trapped charges will build up there, so any change in I-V characteristics are likely due to radiation-induced defects in isolating oxides like the BOX.

Fig. 17 shows the I-V characteristics of a SLVT device after various TID exposures. The devices show a leftward voltage shift at each irradiation level. The plot on Fig. 17. is typical for all the PFET devices tested on this row of structures. The post exposure yield was 93.3% which

resulted in a well-defined gate voltage shift vs. irradiation dose data capture. Fig. 18 is the I-V characteristic of the same SLVT NFET variant with an equivalent gate length. Although this selection of devices performed well during testing, only three of the NFETs shared the same length which was a critical characteristic for the voltage shift analysis. The overlapping range of irradiation levels shared by the PFET and NFET were between 50k to 200k rad[Si]. Fig. 20 plots the absolute value of gate voltage shift at I_d = 100nA) vs. irradiation for both SLVT polarities. The symbols on Fig. 19 represents the average and error bars the standard deviation for both the PFETs (6 devices tested) and NFETs (3 devices tested).

The second type of devices analyzed were regular well, standard gate thickness, high threshold voltage PFETs and NFETs. Fig. 20 shows the drain current and gate voltage relationship of the PFET structures. The I-V curves for the complimentary HVT NFET variants with equal gate length are plotted in Fig. 21. From the two plots, a leftward shift is shown with increasing ionizing radiation exposure as expected The PFET devices in this category had a 59.2% post-irradiation yield. Similarly, the NFET variants had a 52.4% post-irradiation yield. Only three p-type and two n-type structures fit the criteria for the analysis. As such there was a limited amount of data points available. However, Fig. 22. shows consistent upward trends (average and standard deviation) for both types.

Lastly, standard gate, regular threshold voltage and well devices were characterized. Yielding data for these two types of FD-SOI devices was limited. Only one PFET and two NFET structures fit the criteria and performed without failure during the TID testing. Fig. 23. shows the single device analyzed. An interesting post-irradiation analysis of this PFET was that after the initial dose at 50krad[Si], the device no longer showed noticeable leftward shift as was common with all the other devices tested. Conversely, the NFET variant does show the expected leftward shift on Fig. 24. The voltage shift vs. irradiation relationship for the one p-type and two n-type structures is shown in Fig. 25. The flat line for the PFET demonstrates the negligible shift it experienced during the post-irradiation voltage sweep.

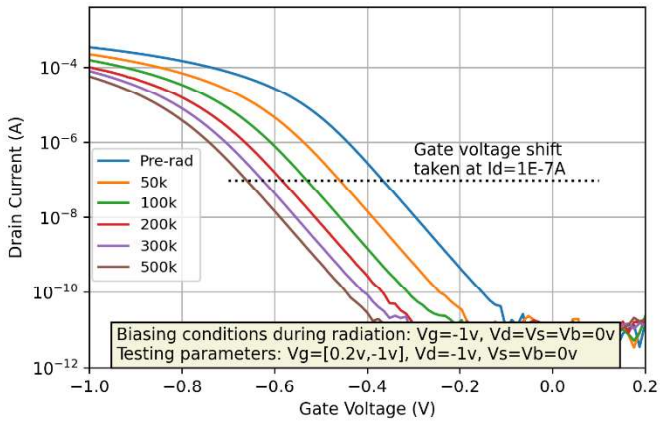


Fig. 17. *Drain current* vs. *gate voltage* characteristics of SLVT, thick gate, flipped well PFET with W/L = 5.0/0.15 (µm/µm).

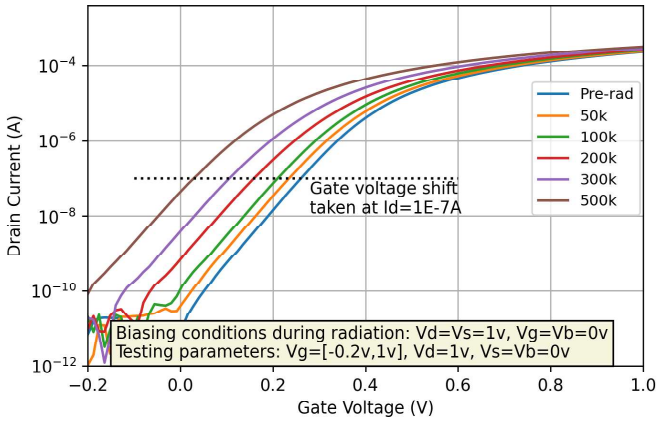


Fig. 18. *Drain current* vs. *gate voltage* characteristics of SLVT, thick gate, flipped well NFET with W/L = 0.25/0.15 (μm/μm).

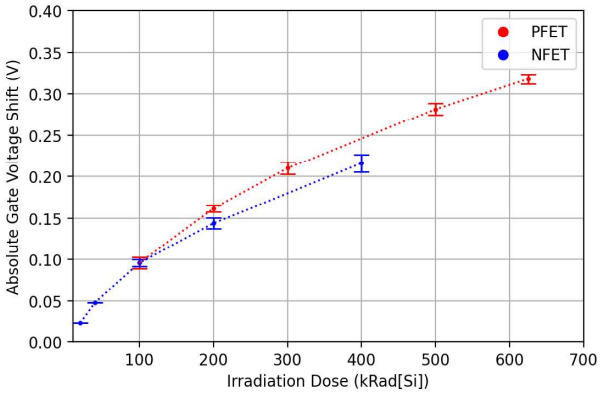


Fig. 19. *Gate voltage shift (@ 100nA)* vs. *irradiation dose* for SLVT, thick gate, flipped well PFET and NFET devices.

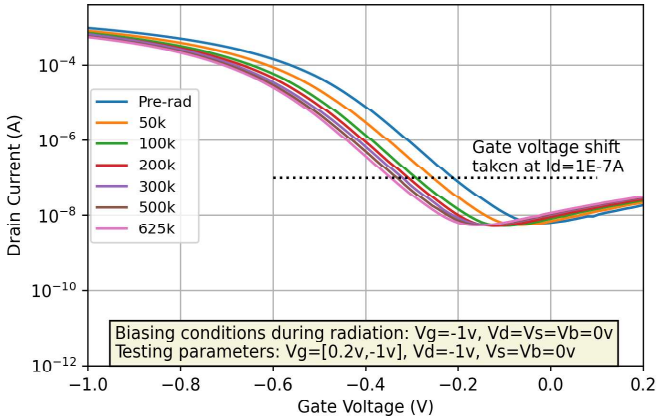


Fig. 20. *Drain current* vs. *gate voltage* characteristics of HVT, standard gate, regular well PFET with W/L = 0.17/0.02 (μm/μm)

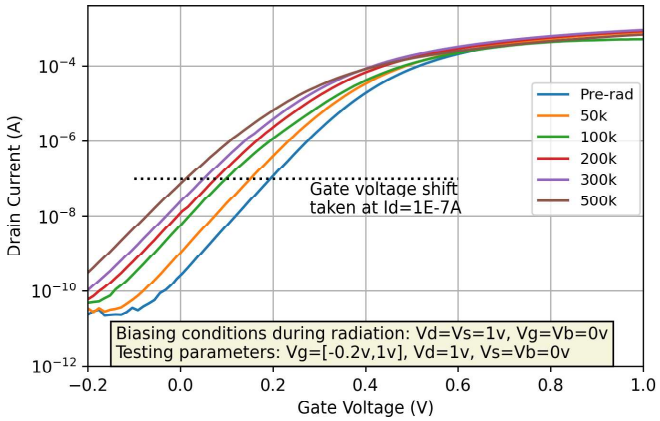


Fig. 21. *Drain current* vs. *gate voltage* characteristics of HVT, standard gate, regular well NFET with W/L = 0.08/0.02 (µm/µm)

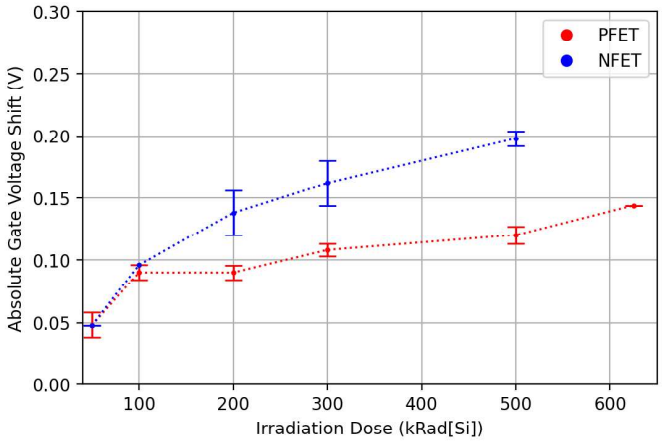


Fig. 22. *Gate voltage shift (@ 100nA)* vs*. irradiation dose* for HVT, standard gate, flipped well PFET and NFET devices

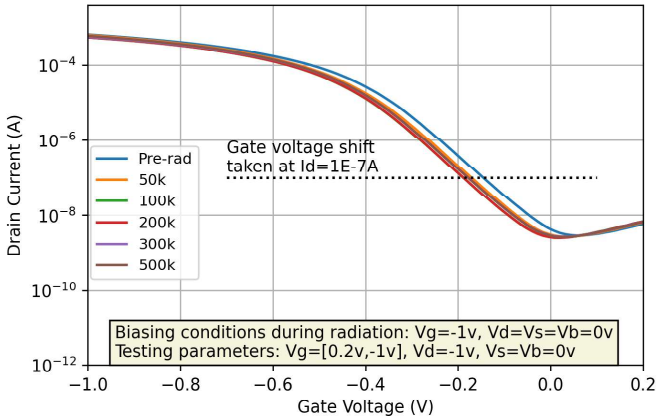


Fig. 23. *Drain current* vs. *gate voltage* characteristics of RVT, standard gate, regular well PFET with W/L = 1/0.02 (µm/µm)

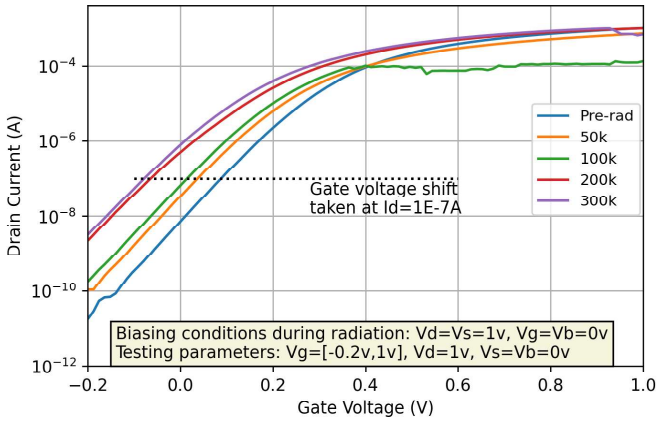


Fig. 24. *Drain current* vs. *gate voltage* characteristics of RVT, standard gate, regular well NFET with W/L = 0.12/0.02 (μm/μm)

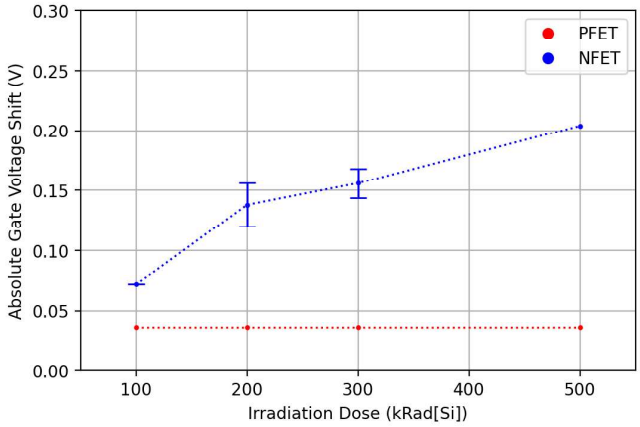


Fig. 25. *Gate voltage shift (@ 100nA)* vs*. irradiation dose* for RVT, standard gate, flipped well PFET and NFET devices

Compact model for 22 FDX

As shown in Figs 17 and 18, TID exposures result in voltage shifts in the I_d vs. V_g characteristics for n- and p-channel SLVT transistors. To perform the radiation response circuit simulations, we build radiation-aware compact models for the transistors. Note that we demonstrate this methodology here on the 22nm FDSOI technology (SLVT variant), but in principle this methodology is general applicable to all CMOS technologies. We model the voltage shifts caused by TID damage in these transistors by connecting a parameterized dependent source in series MOSFET gates. This voltage shift, denoted as the defect potential ϕ_{nt}, is a function of bias and radiation-induced oxide defect densities. For the n-channel transistor (NFET), ϕ_{nt} is set to 96mV, 144mV, and 216mV, which reflect the voltage shifts after 50krad(Si), 100krad(Si), and 200krad(Si), respectively, as shown in Fig. 26a. For the complimentary p-channel transistor (PFET), ϕ_{nt} is set to 96mV, 144mV, and 252mV for the same dose levels (Fig. 26b).

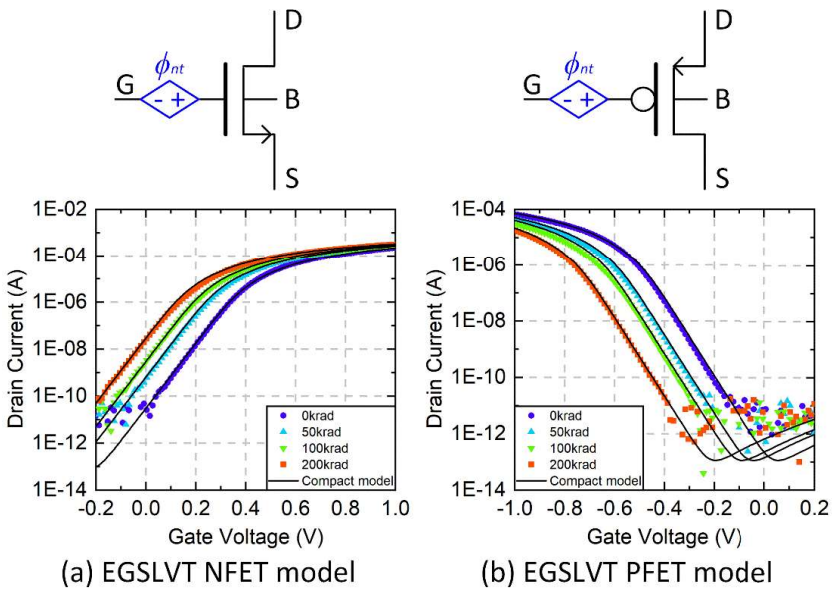


Fig. 26. Radiation-Aware Models for: (a) SLVT NFET and (b) SLVT PFET

Task 1.2 Characterize radiation effects and build models for synaptic weight tuning in RRAM devices

Design, fabricate, and test RRAM devices.

Description

Resistive random access memories (RRAMs) have emerged as the most popular choice for the analog neuromorphic architectures. RRAM devices have shown interesting properties like scalability to sub lithographic limit, multilevel conductance characteristics, low power operation, data retention and endurance. The RRAM devices can be fabricated in the X - bar architecture which helps in the implementation of matrix multiplication functions. The conductance states of the RRAM devices are used to store the synaptic weights within the neural network. The computing power efficiency of an x-point RRAM array has been estimated to be 31000 times than the state-of-the-art microprocessor, with ultra-low power requirement and real time pattern recognition capability.

RRAMs are simple metal-insulator-metal structures that show a controllable resistance change. The resistance change of these devices relies on the movement of ions within the material sandwiched between the electrodes. Depending on the applied electric potential, the devices can be set/programmed into low resistance state (LRS) or reset to a high resistance state (HRS). RRAM are filamentary in nature and divided into two main categories: anion based also known as OxRAM and cation based CBRAM devices.

Conductive Bridge RAM (CBRAM) is a subclass of RRAM. These are cation -based devices that are based on redox reactions leading to the formation of positively charged ions i.e., cations. The migration of such metallic cations from the anode to the cathode enables the growth of a

metallic filament with nanometric dimensions [45]. The anode is made up of an electrochemically active metal which can be easily oxidized or reduced (e.g., Cu or Ag) and the cathode is made of a relatively inert metal (e.g., Ni, W, Pt). The layer sandwiched between the electrodes is a solid-state electrolyte in which the ions can drift/diffuse through it. Examples of solid-state electrolytes used are germanium selenides (Ge_xSe_y), germanium sulfides (Ge_xS_y), SiO_2, Al_2O_3 implemented successfully with Ag or Cu anodes to form ReRAM cells [46]-[52].
In this work, we focus on the Ag-$Ge_{30}Se_{70}$ CBRAM devices. The growth conductive filament across the Ag-$Ge_{30}Se_{70}$ CBRAM devices is illustrated in Fig. 27.

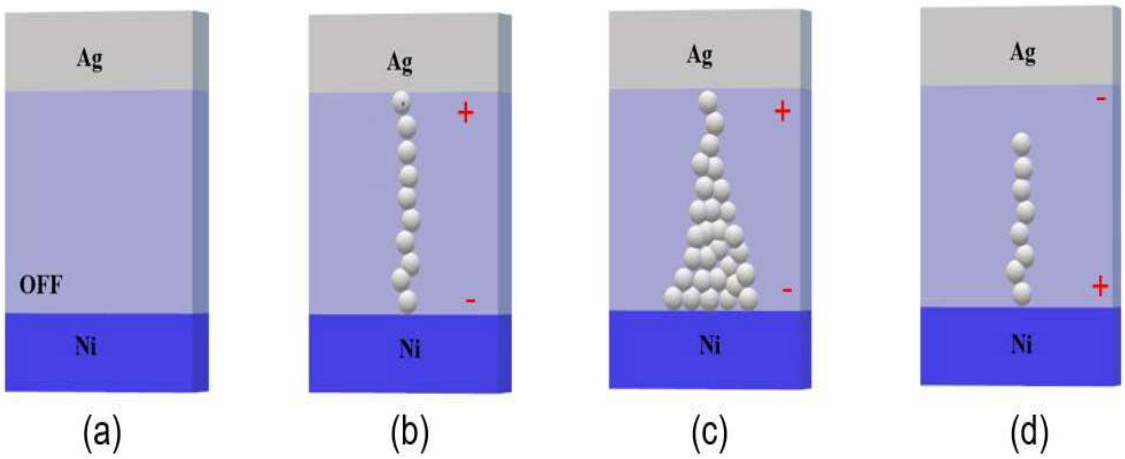


Fig. 27. Schematic representation of mechanism of resistive switching in Ag-$Ge_{30}Se_{70}$ CBRAM devices. (a). Pristine condition –OFF state (b). On application of positive bias, redox reaction takes place leading to formation of conductive filament. (c). On continuous application of positive bias, the conductive filament begins to thicken. (d). Reverse bias leads to removal of the filament – Erase operation.

Referring to Fig. 27, the device is in a high resistance state (HRS) initially. When a positive voltage is applied to the anode, the metal on the anode i.e., Ag gets oxidized (i.e., it loses electron) and Ag^+ cations are generated. The cations start drifting towards to the cathode under the influence of an electric field. As the positive potential on the device is increased, the more Ag^+ ions are generated, and they drift through the solid-state electrolyte. These Ag^+ cations transport across the electrolyte layer and are reduced to Ag at the cathode leading to the formation of a conductive filament. Once the filament bridges the anode and the cathode, the devices can be modulated over a range of "low resistance states" (LRS) by increasing or decreasing the average cross-sectional area (width) of the filament. On application of negative bias on the anode, the Ag which constitutes the filament gets oxidized within the filament on the cathode side. The Ag^+ ions formed start drifting toward the anode under the influence of the electric field. This process continues till the metallic filament is completely broken and the device resistance switches back to HRS. In this work, the use of Ag-$Ge_{30}Se_{70}$ CBRAM devices as an analog synapse for crossbar array type architecture is explored.

CBRAM Device Fabrication

These CBRAM devices were fabricated at the NanoFab class 100 cleanroom facility at the at Arizona State University NanoFab. The CBRAM fabrication process is described in this section.

For the initial processing step, depicted in Fig. 28, a Si wafer is coated with a 170 nm layer of low-pressure chemical vapor deposited (LPCVD) Si_3N_4. This layer serves as a passivation layer between the Si substrate and the CBRAM structure.

Fig. 28. LPCVD deposition of Si3N4.

In the second step, a nickel cathode is created by evaporating 65nm Ni using a Lesker PVD electron-beam (e-beam) evaporator (Fig. 29). The chamber pressure prior to deposition was 10^7 Torr. The deposition rate for nickel was 0.5 Á/s to ensure smooth and even deposition of the metal.

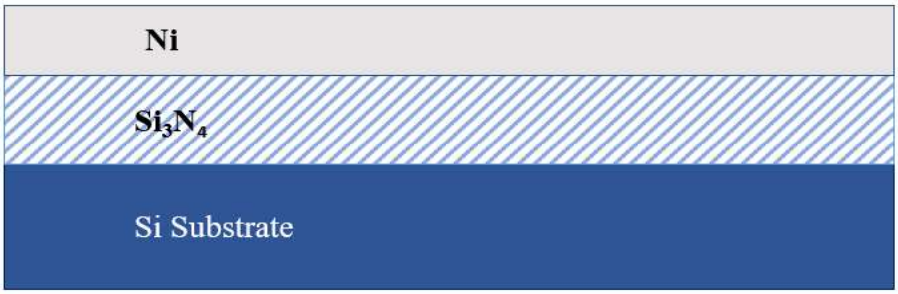


Fig. 29. Deposition of Cathode (Ni) metal.

To pattern the nickel electrode, hexamethyldisilazane (HMDS) is spin coated at 3500 rpm for 30sec and baked at 100 °C for 60 s to facilitate photoresist adhesion. AZ 3312 was then spin coated on the substrate at 3500 rpm for 30 sec and baked at 100 °C for 60 seconds. The photoresist coated wafer is then patterned using EVG 620 lithography aligner set to 45mJ/cm ^{2} to create Ni cathode bars as shown in Fig. 30, which is a top view of the crossbar tile. After exposure, the resist is developed for 60 seconds using Az 300 MIF developer. The resist is hard baked at 110 °C post developing to create an etch mask. The exposed nickel is chemically etched using Nickel Etchant TBF (20 % Nitric Acid solution) for 4 minutes. The remaining resist is removed by soaking the wafer in acetone solution followed by methanol and isopropyl alcohol rinse.

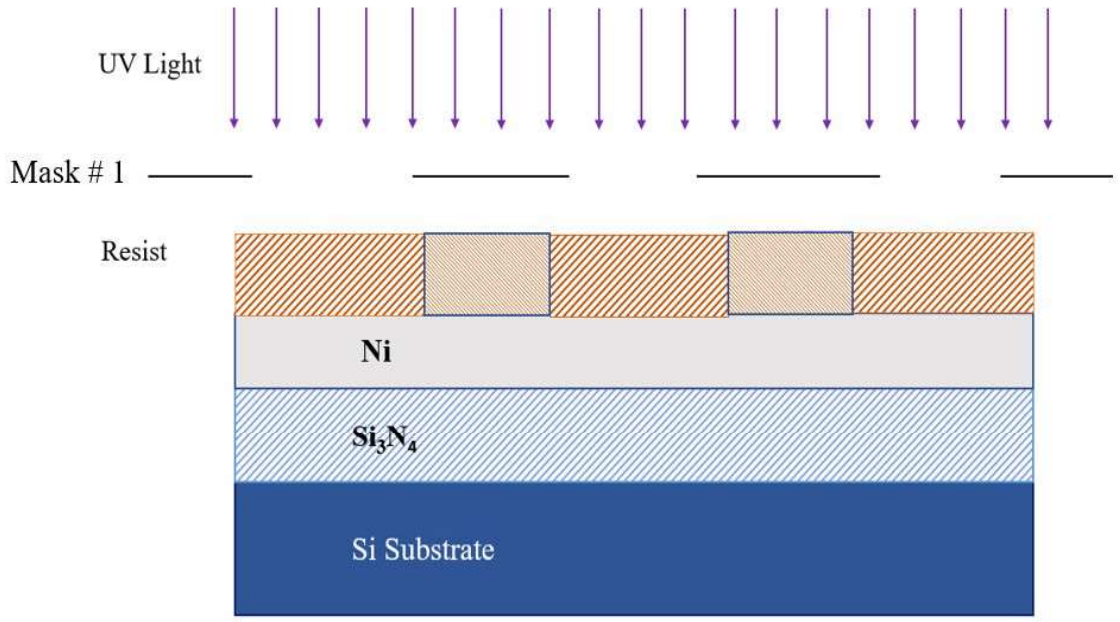


Fig. 30. Mask # 1 creating Nickel cathode bar.

To create isolation between the CBRAM devices, 100nm of SiO_2 is deposited using an Oxford Enhanced Chemical Vapor Deposition (PECVD) tool (Fig. 31). To create the device vias, a double layer resist recipe was used to pattern the SiO_2 layer. First, HMDS is spun on at 4000 rpm for 30 sec followed by a bake at 120 °C for 1 minute. Next, OCG 825 is spun at 4000 rpm for 30sec followed by a bake at 105 °C for 1 minute. The first resist layer is followed by AZ 3312 spun on at 4000 rpm for 30 sec and baked at 95 °C for 1 minute. Mask # 2 was used to create device vias for etching (Fig. 32). The resist is developed at 45 mJ / cm^{-2} and developed in AZ 300 MIF for 1 minute. The device vias are etched using anisotropic reactive ion etching (RIE) for 4 mins 30 secs (Fig. 33).

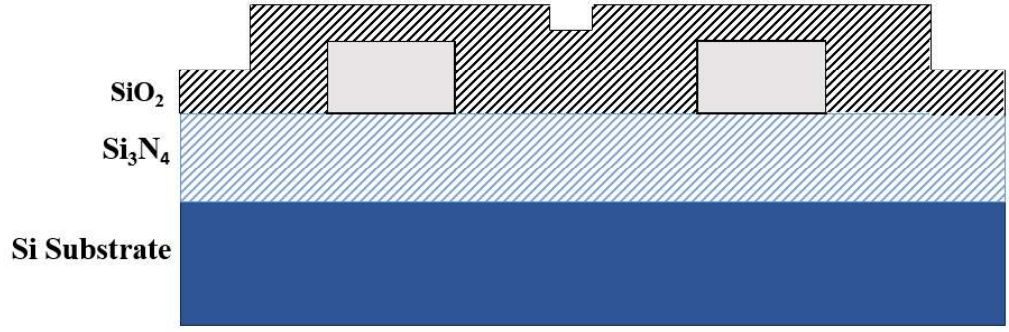


Fig. 31. Ni layer chemically etched to create cathode bar followed by oxide deposition.

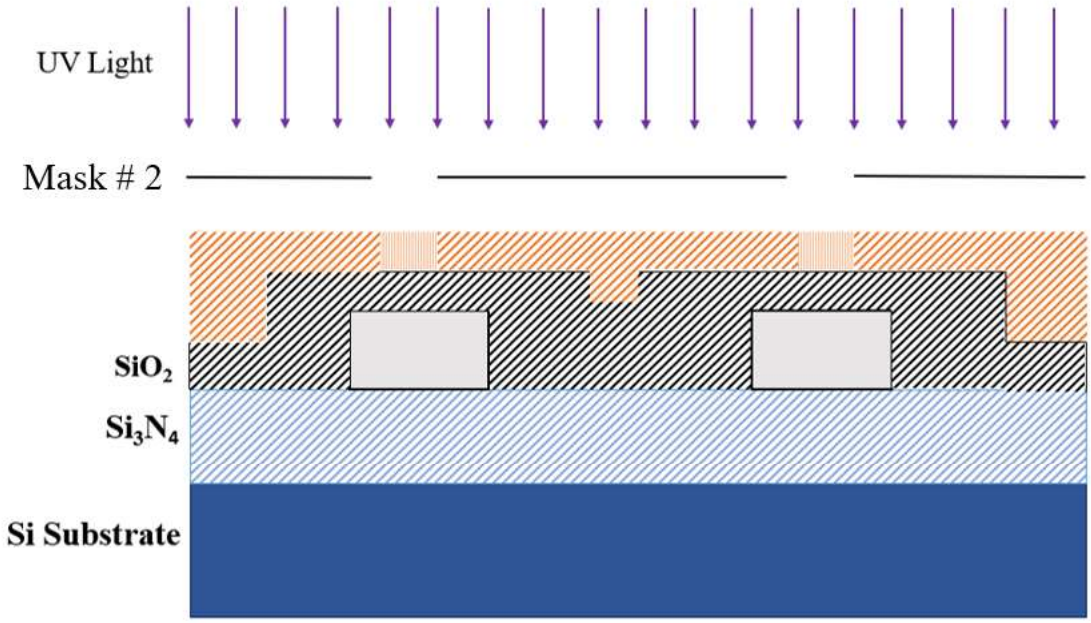


Fig. 32. Mask # 2 used to etch via through SiO_2.

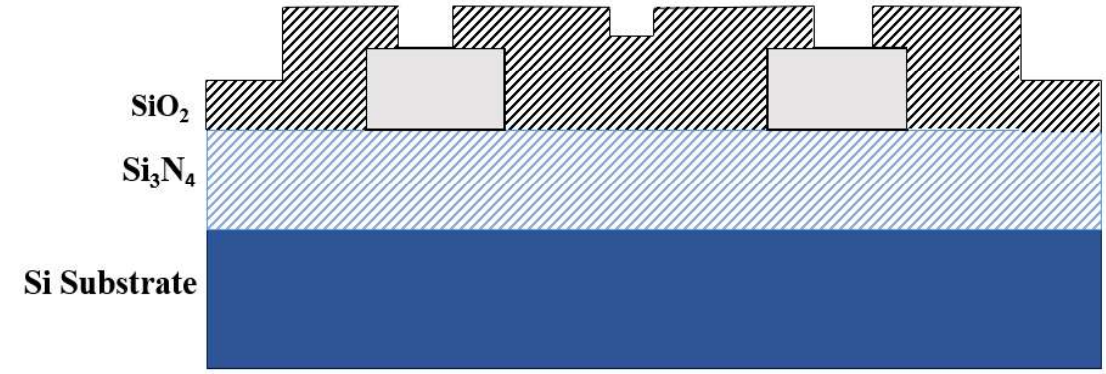


Fig. 33. Dry etching used to create vias through SiO2 to Ni.

Double layer resist is also used as a lift -off layer. The chalcogenide lift-off mask i.e., Mask # 3 is used to pattern the resist with UV exposure to 45 mJ/cm^{-2} and it is developed for 1 minute using AZ 300 MIF (Fig. 34). Next, the wafer is placed in a Cressington 308R thermal evaporator where 60nm of $Ge_{30}Se_{70}$ is thermally evaporated followed by 30nm Ag at a deposition rate of 0.75 Á/s (Fig. 35). The chalcogenide layer ($Ge_{30}Se_{70}$) layer is photodoped with Ag by exposing the wafer to UV light for 1 hour to a dose of 5.3 J/cm^2. The photo-doping process is used to drive the Ag into the chalcogenide layer to attain a saturation concentration of 33 .at% concentration. After the doping process, the wafer is placed back into the Cressington

evaporator where additional 35nm of Ag was deposited to form the top active metal electrode layer. Next, the resist is thermally cracked by placing the wafer on a thermal hot plate at 150 °C. The thermal shock breaks the resist film and allows a clean lift-off (Fig. 36). The wafer is then placed in acetone to remove the metal as well as the resist layers.

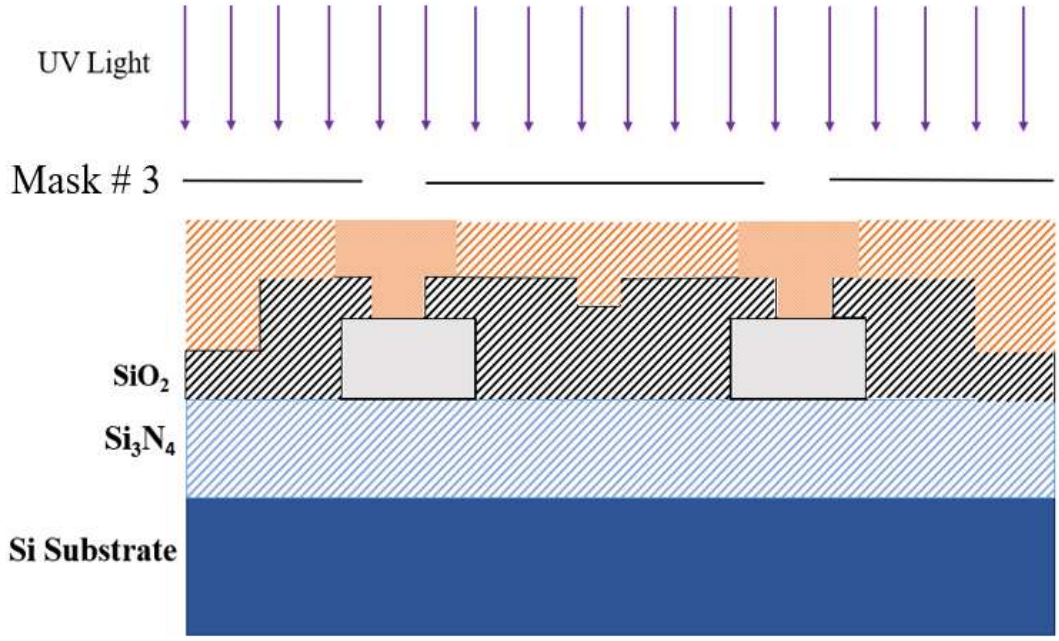


Fig. 34. Mask # 3 used to for chalcogenide -Ag lift off layer.

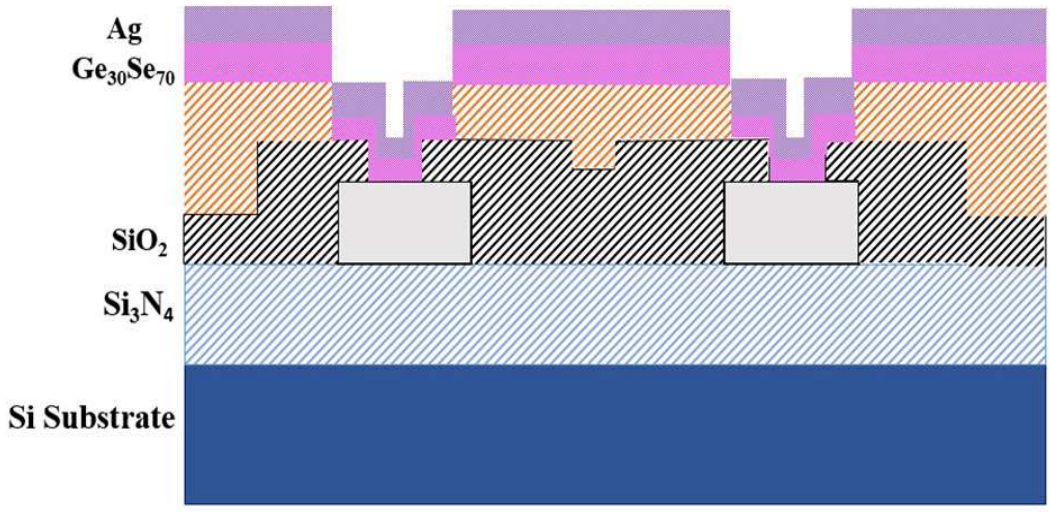


Fig. 35. Active metal electrode deposition.

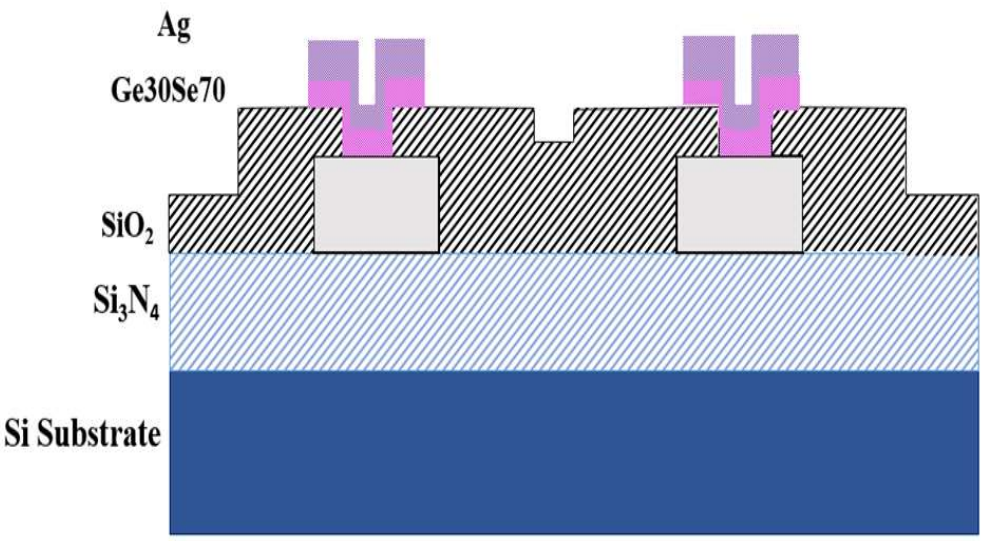


Fig. 36. Lift-off patterning of top active metal electrode layer.

The final step is to create a top Al crossbar electrode. To create the Al metal contact, double layer lift-off resist recipe is used. Mask # 4 is used to pattern the resist with UV exposure to 45 mJ/cm^{-2} and it is developed for 1 minute using AZ 300 MIF (Fig. 37). The wafer is then placed in the Lesker PVD sputtering machine and 250 nm of Al is deposited at a deposition rate of 1.5 Å/sec (Fig. 38). After deposition the resist is thermally cracked to allow clean lift off and placed in acetone. Fig. 39 depicts the final device structure.

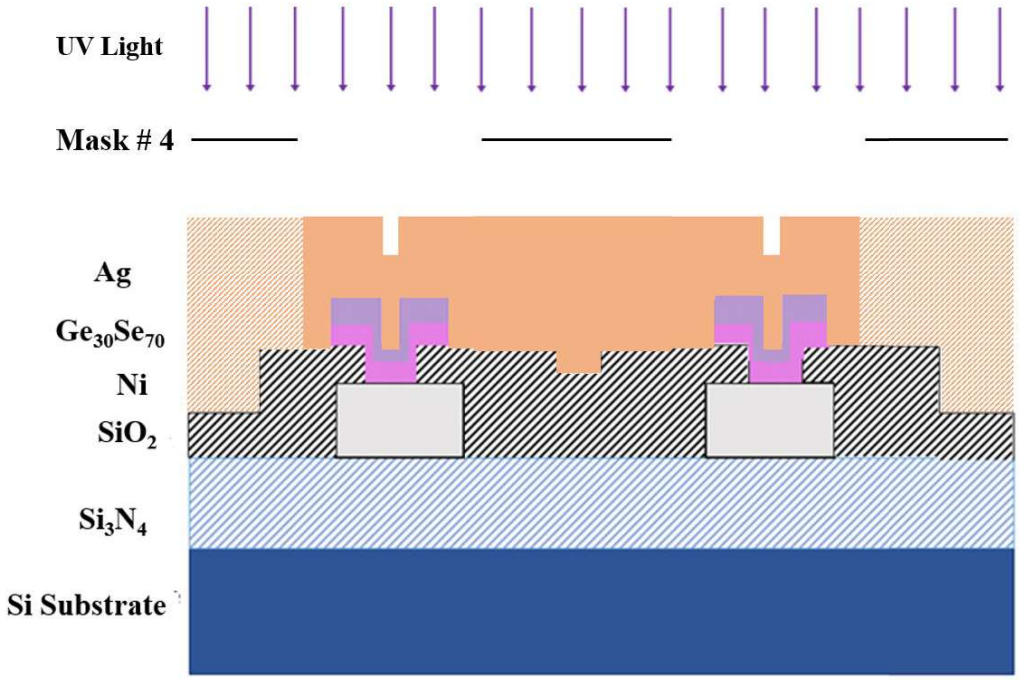


Fig. 37. Mask #4 used for patterning the top contact metal electrode.

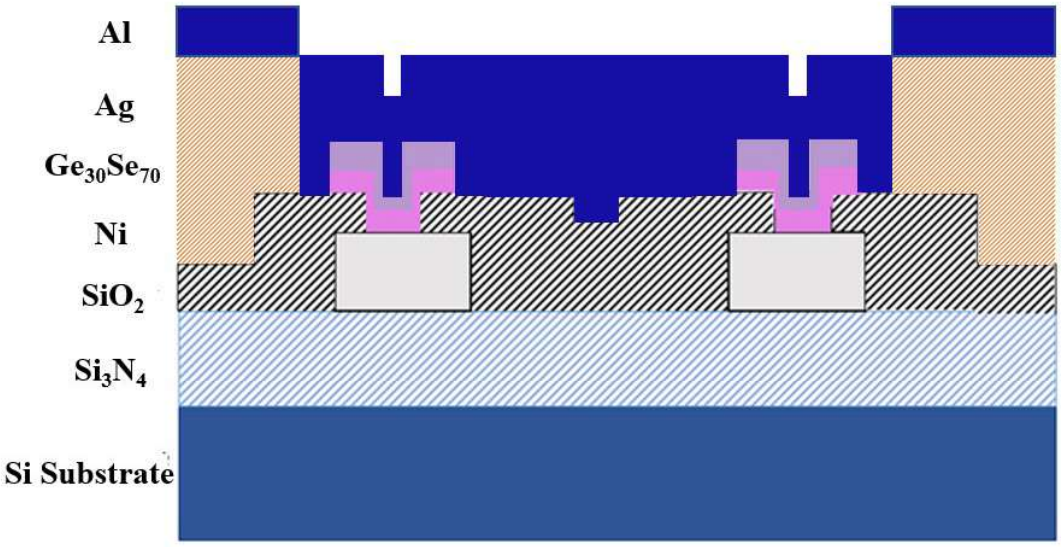


Fig. 38. Contact metal (Al) deposition.

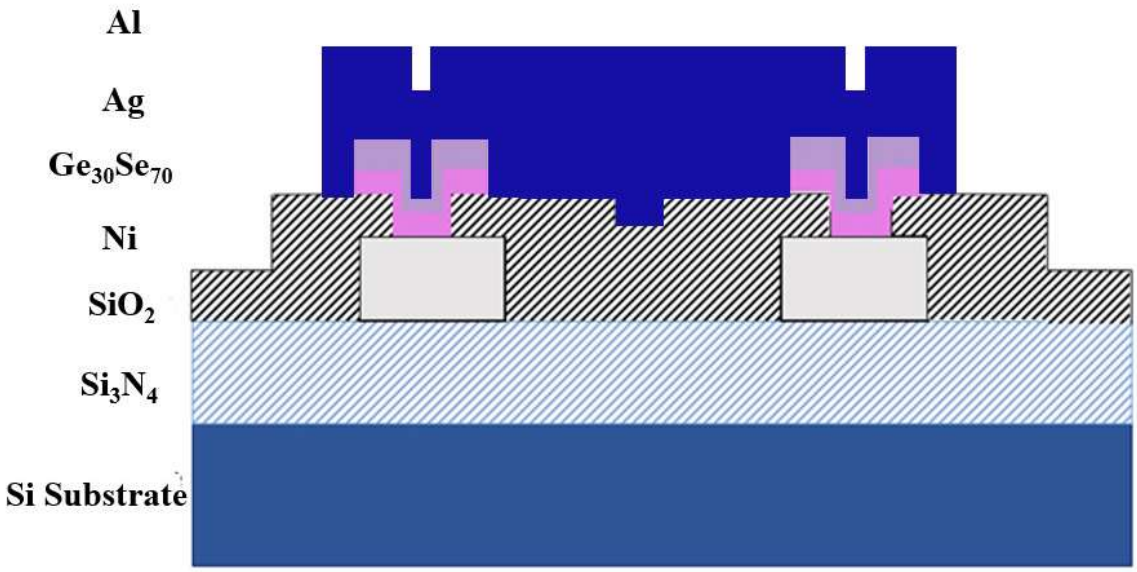


Fig. 39. Final CBRAM device structure.

Study of CBRAM as an analog synapse

The main advantage of the analog crossbar array architecture is that it can perform the inference and training efficiently. A popular choice for an analog crossbar accelerator has been RRAM devices. However, for neuromorphic computing applications, these devices must meet several stringent requirements. For inference applications it is necessary that the devices have

at least two conductance states, low cycle-cycle variability, low read noise as well as good retention. To accelerate training, the RRAM elements should be able to switch between a wide range of analog conductance states. One desirable device characteristic for training is a gradual and symmetric response to different programming pulses which are used for weight update (discussed later). In this section, we investigate the response of Ag-$Ge_{30}Se_{70}$ CBRAM crossbar devices to different programming pulse schemes. We analyse the important synaptic properties like conductance linearity response, device-to-device variability, and conductance on/off ratio of Ag-$Ge_{30}Se_{70}$ CBRAM.

DC I-V Characterization

The current-voltage (I-V) characteristic of a fabricated CBRAM synapse is shown Fig. 40. The Fig. shows the characteristic hysteresis DC switching behaviour over three voltage sweep cycles. It is important to note that Ag-$Ge_{30}Se_{70}$ CBRAM devices are forming-free devices since Ag is introduced in the switching layer by the process of photo dissolution. The DC sweep was performed using Agilent 4156 C analyser. Bipolar switching behaviour was observed reproducibly over 50 cycles. It should be noted that these large-scale DC characterizations which toggle the CBRAM cell between HRS and LRS are performed to assess general operational integrity not the incremental LRS switching required for use as a synaptic element. Cumulative distributions for all HRS and LRS measurements are shown in Fig. 41. The distributions show a minimum HRS to LRS ratio above 20, well above the target minimum of 10.

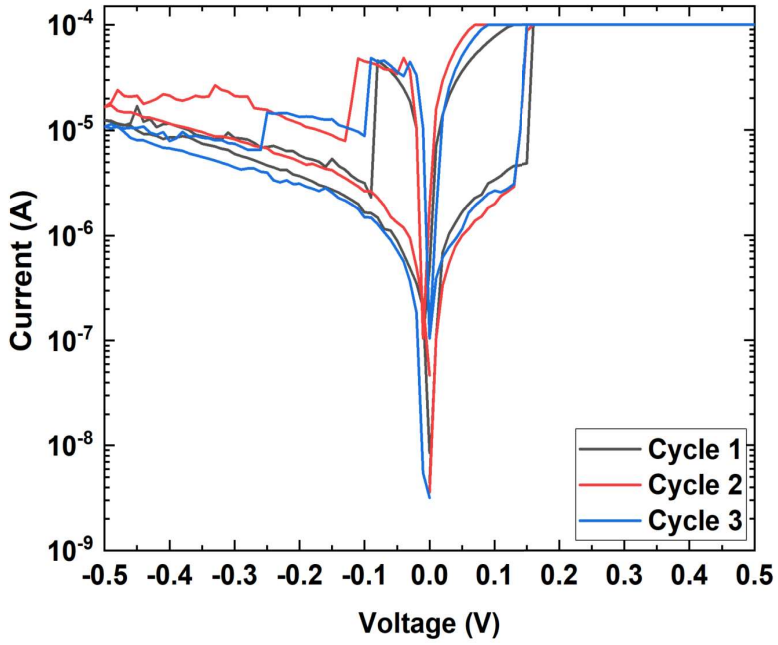


Fig. 40. I-V characteristics of Ag-$Ge_{30}Se_{70}$ CBRAM device cycled at 100μA compliance.

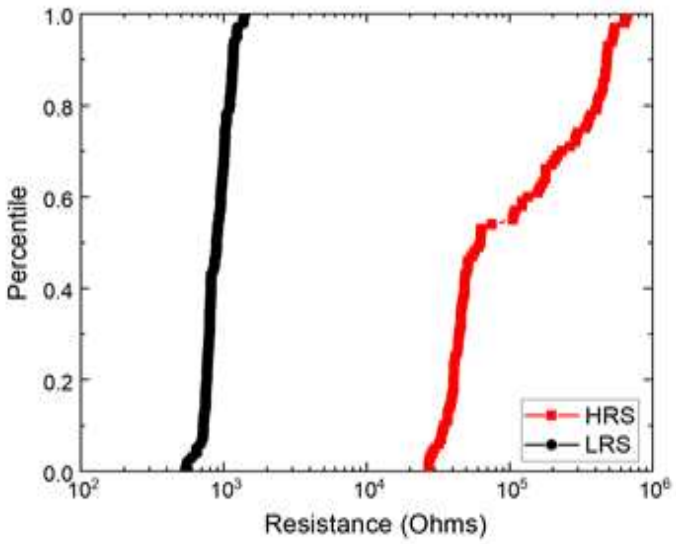


Fig. 41. High Resistance State (HRS) and Low Resistance State (LRS) distribution of the device extracted @0.03V for 50 cycles.

Another important requirement of the synaptic device is to have multilevel conductance characteristics. The multilevel conductance characteristics of crossbar Ag-$Ge_{30}Se_{70}$ CBRAM is demonstrated in Fig. 42 by increasing the current compliance limit for each consecutive DC cycle. The compliance current is the maximum allowable current through the device. A higher the compliance current causes more ion flow through the device leading to a thicker conductive filament and lower resistance states. This characteristic of the CBRAM is well established in earlier works.

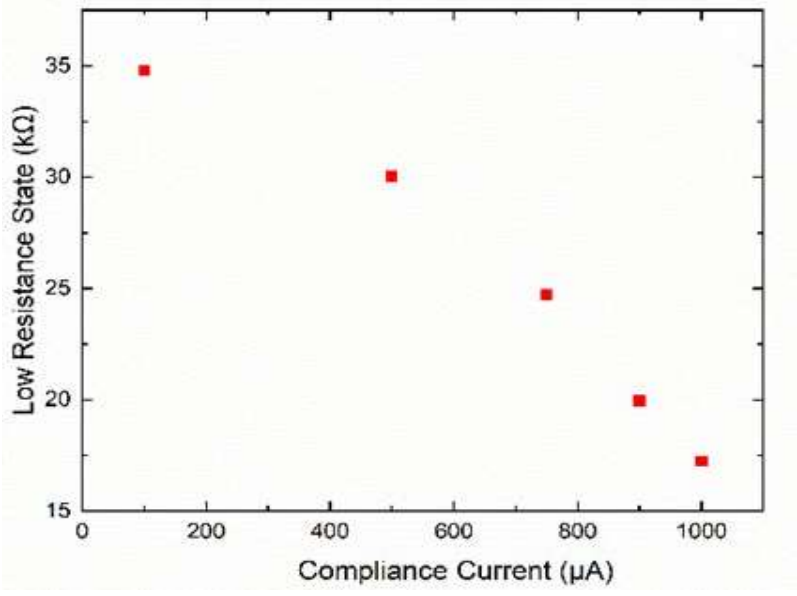


Fig. 42. Multi-level conductance demonstration of Ag-Ge30Se70 CBRAM with increasing compliance current.

Constant Pulse Programming

Multilevel conductance switching is demonstrated by applying voltage pulses to top electrode (Ag anode) while the bottom electrode (Ni cathode) is grounded. The measurements are performed using the Keithley 4200 SCS parameter analyser with a built-in 4225 PMU module. To achieve gradual switching behaviour, we applied 100 consecutive SET pulses (0.35V for 100ns) and 50 consecutive RESET pulses (-0.25V for 100ns). Each write and erase pulse was followed by read voltage of 30 mV to extract conductivity. The change in conductance of the device is plotted as a function of pulse number in Fig. 43 for 20 cycles of SET/RESET operations. For these input pulse parameters, the results indicate a highly non-linear response in device conductance, where the incremental change in conductance is greatest for the early SET/RESET signals in each sequence, saturating quickly to a to a maximum/minimum level. We observe a 10x change in conductivity for the 100ns pulse width.

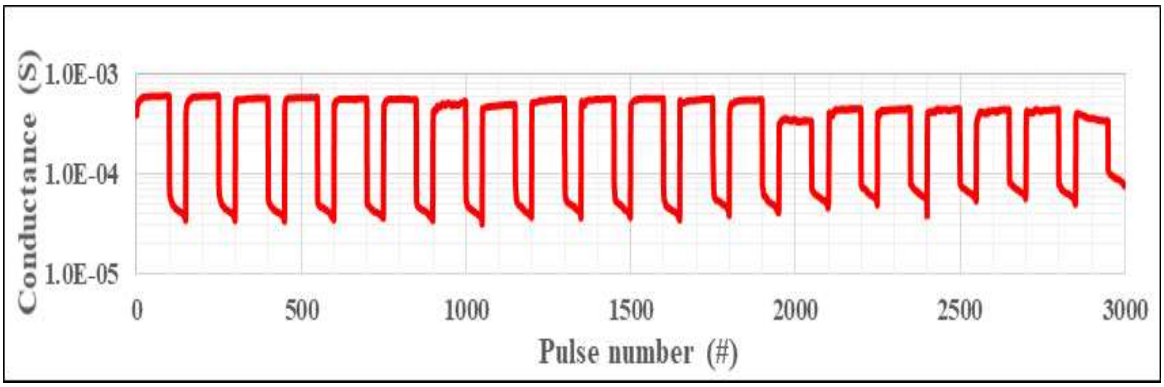


Fig. 43. The potentiation and depression cycling of Ag – $Ge_{30}Se_{70}$ CBRAM with P.W. = 100 ns.

The effect of pulse width on the switching response is observed in Fig. 44. For these data, the SET/RESET pulses have a longer pulse width of 100µs. An increase in maximum and minimum conductance range is observed for higher pulse width, which can be attributed to the greater widening of the conductive filament during the longer voltage stress time.

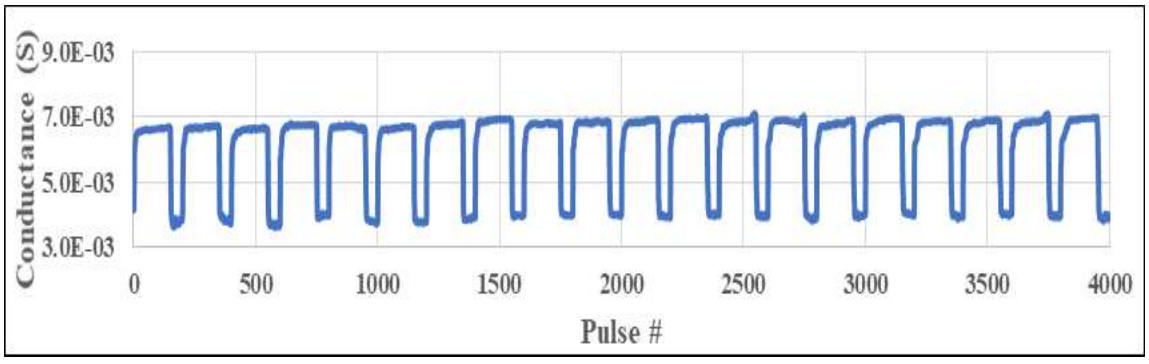


Fig. 44. Repeated pulse cycling Ag-Ge30Se70 CBRAM with P.W =100 µs.

Fig. 45 plots the response for both pulse widths over one cycle and shows that for shorter pulses, i.e., 100ns, we observe a larger G_{MAX} /G_{MIN} ratio and lower conductance levels. When

considering the use of these devices for parallel VMM programming lower conductance levels are preferred since this will reduce the energy required for both VMM programming and inference.

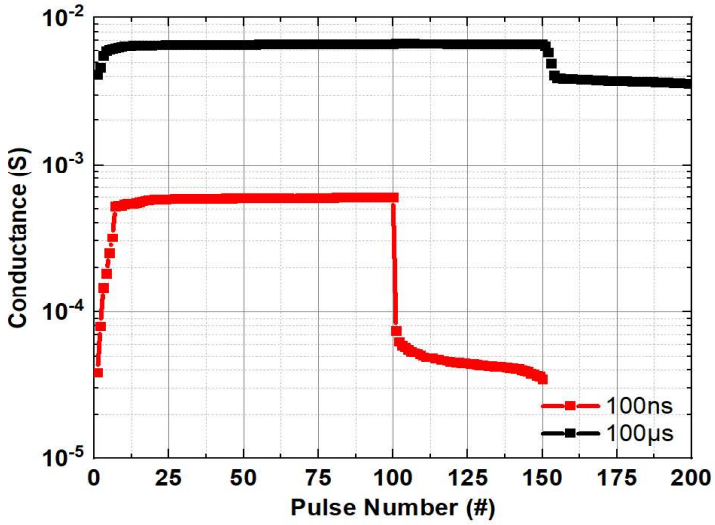


Fig. 45. The analog response of the Ag-Ge30Se70 CBRAM device to different pulse-widths.

The conductance update response for both pulse widths was compared by normalizing the conductance range as depicted in Fig. 26. The conductance update for the potentiation is identical for both the pulse-widths. The depression behaviour for 100ns pulse width seems to be better. A detailed study of the linearity response is provided in the following subsections.

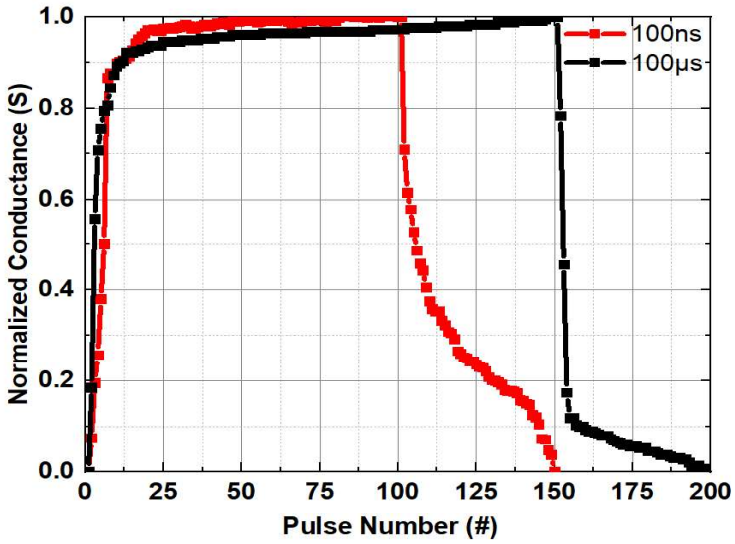


Fig. 46. Normalized conductance of the device to different pulse widths.

Variable Pulse Programming

An increasing amplitude pulses were applied to the anode of the device. Increasing amplitude voltage pulses from 0.2 V to 1.5 V (pulse width = 100 µs and 50 mV step) for potentiation and -0.2 V to -1.2 V (pulse width = 100 µs and 50 mV step) for depression were applied. Each pulse was accompanied by a read pulse of 50 mV with a pulse width of 100 µs. Fig. 47 shows the potentiation and depression characteristics of the device. Fig. 48 displays the response of the devices to multiple pulse cycles. Unlike the characteristics above, these devices are shown to exhibit good linearity over a suitable conductance range. This linear response is realized by careful control of the thickness of the metallic filament. One should however note that using varying amplitude pulses eliminates the advantage of parallelism since it necessitates addressing each analog resistive element in a crossbar, individually.

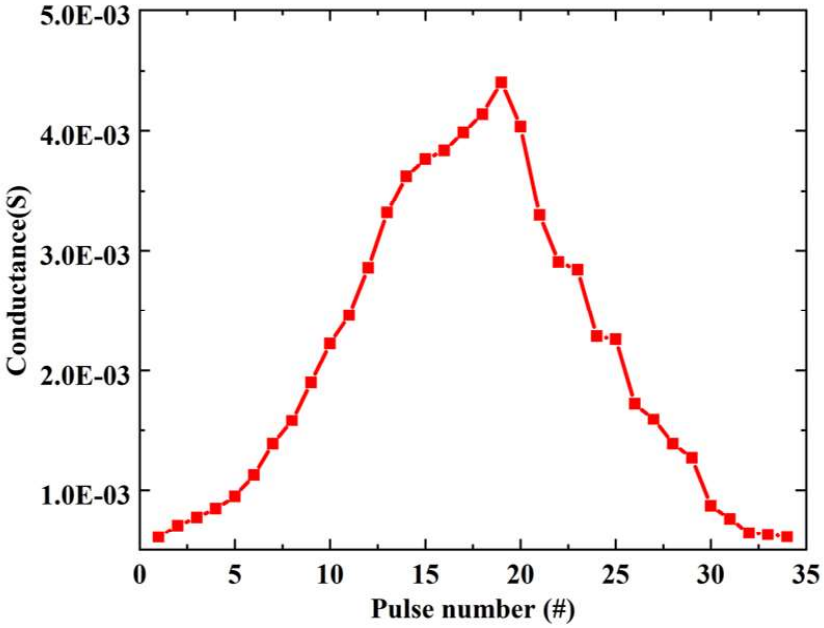


Fig. 47. Conductance response of Ag-$Ge_{30}Se_{70}$ CBRAM to incremental amplitude pulses.

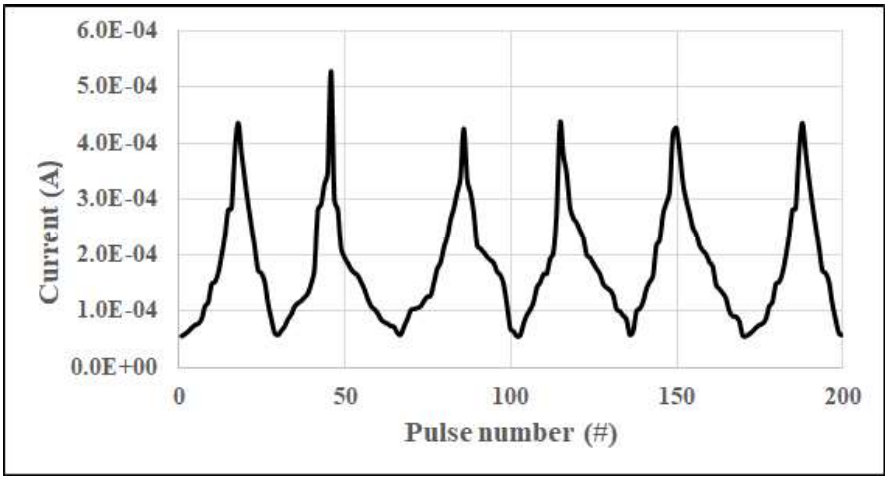


Fig. 48. Potentiation and depression cycling to increasing amplitude pulses.

Non-Linearity Response Study

The linearity metric refers to the linear behavior of the conductance vs. pulse number curve. The linearity metric is important in the weight mapping algorithms. Linear and symmetric behavior is an ideal characteristic. Actual devices typically show some degree of asymmetry and nonlinearity in the weight update characteristics. Asymmetry between potentiation and depression characteristics can degrade pattern recognition and online training by the neural network. The Ag-$Ge_{30}Se_{70}$ CBRAM devices studied in this show an increase in conductance for initial pulse train but tend to saturate at the end. Recent results have shown that this nonlinearity/asymmetry has caused the learning accuracy loss in the neural networks.

To mathematically extract the non-linearity factor, the potentiation and depression response of the device to the 2-pulse scheme response are fit to the following equations:

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

$$G_{LTD} = B\left(1 - e^{\left(P\frac{Pmax}{A}\right)}\right) + G_{max} \quad (2)$$

$$B = \frac{G_{max} - G_{min}}{1 - e^{-\frac{Pmax}{A}}} \quad , \quad (3)$$

where, *G, P*, and *A* are the conductance value, pulse number, and nonlinear behaviour of weight update, respectively. G_{max} and P_{max} are the maximum conductance and pulse widths, respectively, obtained from the experimental data. The different nonlinearity factors for potentiation and depression are obtained by simulating the curves using MATLAB with equations (1) and (2) respectively. The non-linearity factor *A* for the potentiation and depression

are found to be 6.65 and -2.5, respectively, for the constant amplitude (P.W.= 100 ns) as shown in Fig. 49. Similarly, for the increasing amplitude pulses, the non-linearity factors the potentiation and depression was found to be 1 and -0.96 respectively as depicted in Fig. 50.

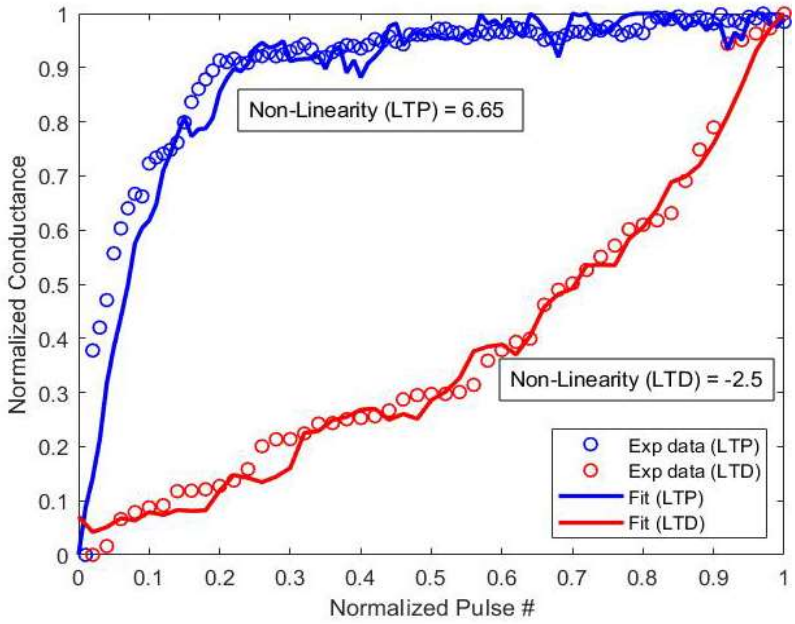


Fig. 49. Non-linearity fitting factors -Constant Amplitude pulses (100 ns).

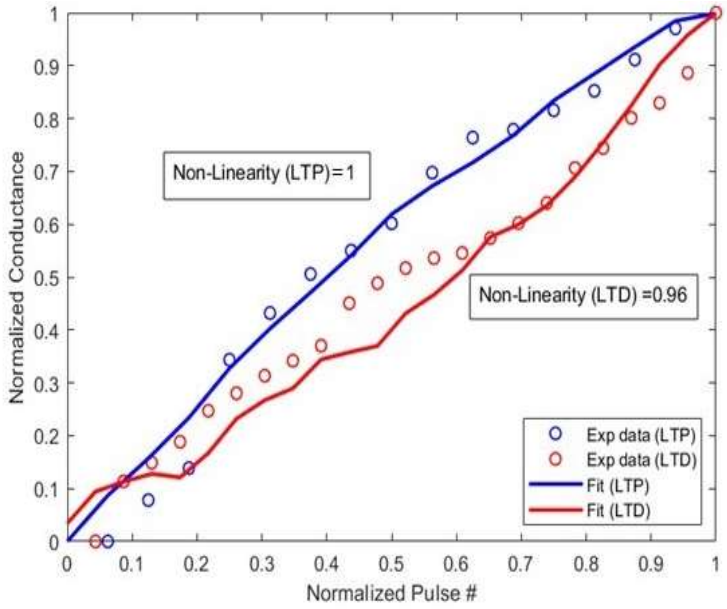


Fig. 50. Non-linearity fitting factors - Increasing Amplitude pulses (100 µs).

Device - Device Variation

The operation of Ag-$Ge_{30}Se_{70}$ CBRAM devices involves the drift and diffusion of metal ions to form a filament and hence these devices are prone to variation in conductance update from

device-to-device across the crossbar array. Different devices will tend to have different non-linearity baselines. The device-device variability is studied by measuring the weight update behaviour of 10 devices in the crossbar array. Here the variability is calculated as the deviation in the non-linearity factor for different devices. The device-device variability is seen to be ±1.05 as shown in Fig. 51.

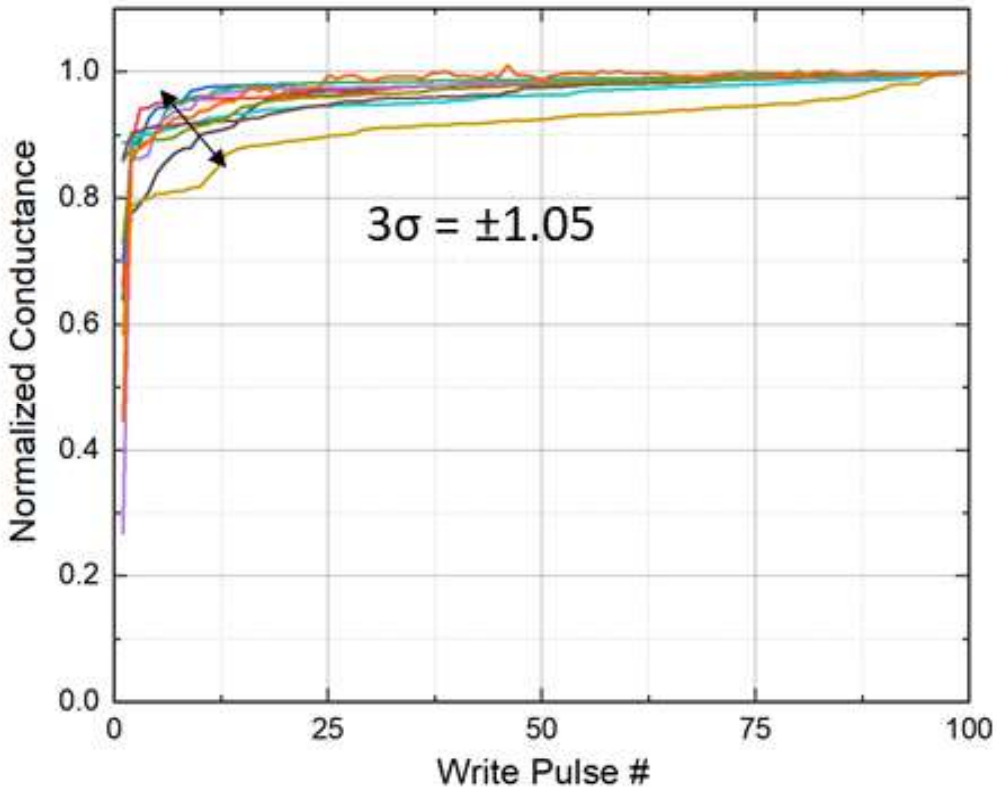


Fig. 51. Device to device weight update variation for Ag-Ge30Se70 CBRAM devices.

On/Off Conduction Ratio

Ideally the weight values in the training algorithms are represented as normalized conductance, ranging from 0 to 1. However, the minimum conductance can be zero only if the ratio of maximum and minimum conductance i.e. On/Off ratio approached infinity which is practically not feasible. The learning accuracy has been observed to degrade with a small on/off conductance ratio because for calculations on a small weight update value will lead to significant distortions. The on/off conduction ratio for Ag-$Ge_{30}Se_{70}$ CBRAM devices for identical pulse application i.e., Fig. 32, is 16.

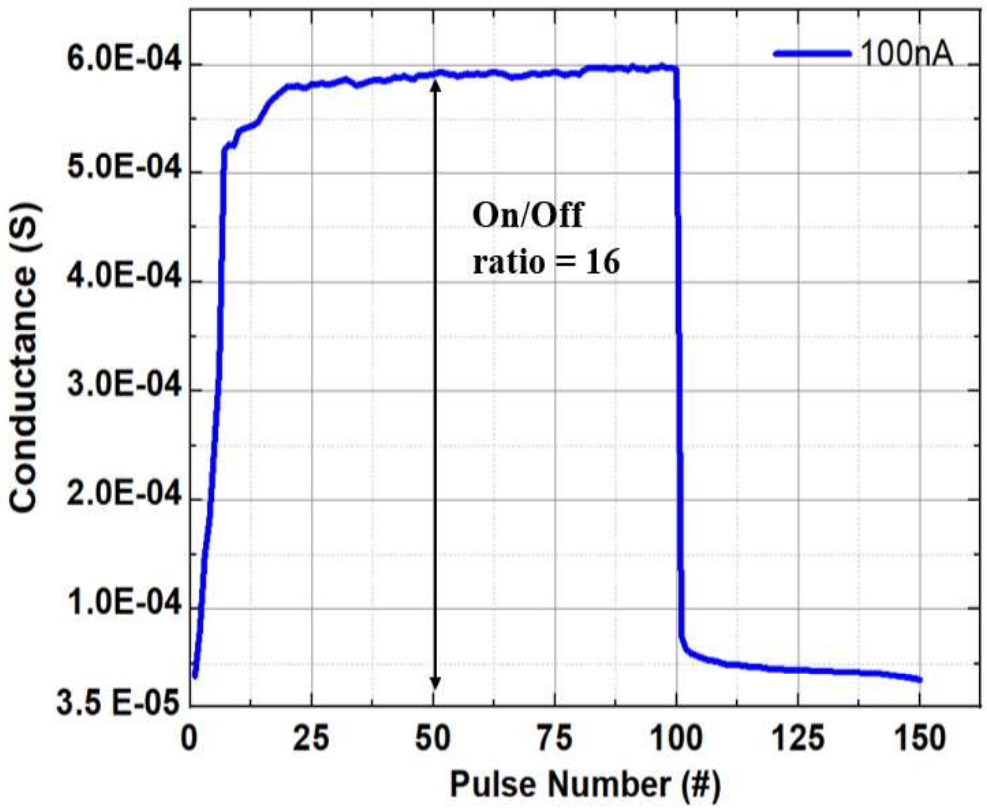


Fig. 52. Conductance on/off ratio for Ag-Ge30Se70 CBRAM device for constant amplitude pulses (100ns).

Modeling CBRAM

TCAD Description

A two-dimensional TCAD structure representative of a CBRAM cell is shown in Fig. 53. The TCAD device simulator that implements our new CBRAM model is Silvaco's Victory Device Simulator with the Electrochemistry Module. For the technology considered in this work, the active metal, (top) anode is silver (Ag). The inert metal, (bottom) cathode is nickel (Ni). The electrolyte layer is a silver doped chalcogenide glass (Ag-$Ge_{30}Se_{70}$), with a thickness of 70 nm, length of 1μm and width of 10 μm. Ag doping is performed to facilitate ion migration through the glass. The doping profile in the chalcogenide glass (ChG) follows a Gaussian distribution with a maximum concentration at the anode/ChG interface. A 2 nm width seed element is added to the top of the Ni cathode. This facilitates the formation of the filament from the seed. Without this seed a conductive metal sheet will be formed instead of the filament during transient simulation. These choices for material type and dimension were chosen to match the features of the experimental devices fabricated at Arizona State University.

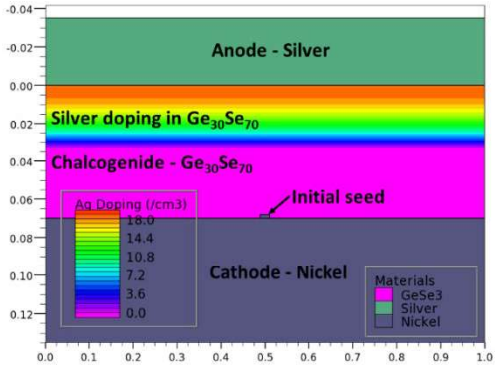


Fig. 53. CBRAM device structure used in the modeling effort.

CBRAM operates via oxidation-reduction reactions. The information stored in the device is set by the resistance between the anode and cathode contacts. For binary storage the resistance can be switched between a High Resistance State (HRS) and a Low Resistance State (LRS). Synapses for neuromorphic applications can be two-state (binary), multi-state, or analog. In an initial HRS condition, there exists no low resistance pathway between the contacts. However, when a positive bias is applied to the anode relative to the cathode, Ag cations (Ag^+) are created through oxidation at the anode/ChG interface (Eq. 4). The cations drift through the electrolyte in the direction of the applied electric field and are subsequently reduced by gaining an electron at the cathode (Eq. 5).

$$Ag \rightarrow Ag^+ + e^- \tag{4}$$

$$Ag^+ + e^- \rightarrow Ag \tag{5}$$

On continued application of positive bias, a conductive filament grows from cathode to anode. The proper combination of time and voltage applied to the anode will create a filament which bridges the electrolyte layer, switching the device into LRS. When the bias is reversed, Ag ions will drift back to the anode, dissolving the filament and switching the device back to HRS. Fig. 54. shows TCAD simulations of the forming and dissolving of Ag filament at different biases during the I-V hysteresis sweep. It can be observed that the concentration of Ag is being modulated in the center of the device where the seed element was added.

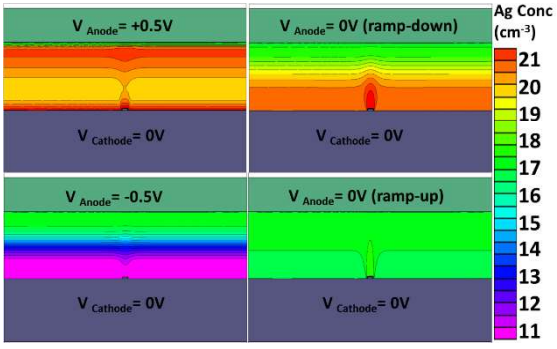


Fig. 54. Concentration of silver at different biases during the I-V hysteresis sweep.

Physical Model

Field Dependent Ionic Mobility Model - The field dependent ion mobility of a chemical species is given by the Mott-Gurney model [2]. The diffusivity, DS of the chemical species can be expressed as

$$D_S = d^2{}_{hop} * v_a * exp\left(\frac{-E_A}{kT_L}\right) * \left(\frac{sinh(\eta_S)}{\eta_S}\right) \quad (6)$$

where dhop and va is the hopping distance and the hopping frequency of the species in the ChG layer, respectively. The exponential term corresponds to the probability that the attempted hopping of the species succeeds in overcoming the energy barrier represented by the activation energy EA. In Eq. 6, k is the Boltzmann constant and TL is the lattice temperature of the ChG layer. The term ηs can be expressed as

$$\eta_S = \left(\frac{d_{hop}}{2}\right) * z_S * \frac{q}{kT_L} * |E|, \quad (7)$$

where, |E| is the magnitude of the electric field and zS corresponds to the magnitude and sign of the electric charge on the chemical species S. In our model two chemical species are present: Ag with neutral charge and Ag+ with a single positive charge. The diffusivity of Ag in ChG layer is negligible compared to Ag+. The diffusivity parameters of Ag+ in ChG layer are obtained in literature.

It should be noted that at very high fields, the Mott-Gurney model becomes invalid because the probability of ion hopping cannot exceed unity. Therefore at high electric fields the ionic drift velocity saturates at a fraction fvsat, of the thermal velocity. This full ionic mobility model is thus expressed as

$$D_S = d^2{}_{hop} * v_a * exp\left(\frac{-E_A}{kT_L}\right) * \{[\frac{Sinh(\eta_S)}{\eta_S}]^{-2} + [f_{vsat} * \frac{Sinh\left(\frac{E_A}{kT_L}\right)}{|\eta_S|}]^{-2}\}^{-1/2} \quad . \quad (8)$$

Electron Mobility Model for Filament

The primary electrochemical reactions occurring in the CBRAM device are given by Eqs. 1 and 2. When the filament starts to form between the cathode and anode, the conductivity of the device increases. As the concentration of the Ag in the filament increases, the electron mobility increases. Therefore the total effective electron mobility (μeff) depends on the electron mobility in ChG layer, electron mobility in the filament and the concentration of the silver in the filament. This electron mobility model is

$$\mu_{eff} = \mu_s + w * (\mu_c - \mu_s), \quad (9)$$

where μs is the electron mobility in the host layer (ChG), μc is the electron mobility in the conductive material (Ag filament) and w is a weight factor that depends on the concentration of the Ag in the conductive filament that can vary between 0 to 1. The weight factor is

$$w = \varepsilon + \frac{\left\{(f - f_0) + \delta * ln\left\{cosh\left[\frac{(f - f_0)}{\delta}\right]\right\}\right\}}{2 * (1 - f_0)} \tag{10}$$

where f is the ratio of local concentration of the conductive bridging species relative to its maximum concentration in the ChG, f0 is the bridging species fraction at which the conductivity starts to increase, δ is the half width of the transition region around f0, ε is a constant required to make w = 0 when f = 0.

SIMULATION RESULTS - Fig. 55. compares experimental data and simulation results of the I-V hysteresis sweep performed on a Ag-Ge30Se70 CBRAM device. The filled black squares and unfilled blue circle symbols correspond to the experimental data and TCAD simulation results, respectively. The sweep direction is numbered in Fig. 55 as follows: step 1 corresponds to ramping up of voltage from 0V to 0.5; step 2 corresponds to ramping down of voltage from 0.5V to 0V; step 3 corresponds to ramping down of voltage from 0V to -0.5; and step 4 corresponds to ramping up of voltage from -0.5V to 0V. The silver filament is formed in step 1, i.e., switched from HRS to LRS. The filament is dissolved in step 3, i.e., switched from LRS to HRS.

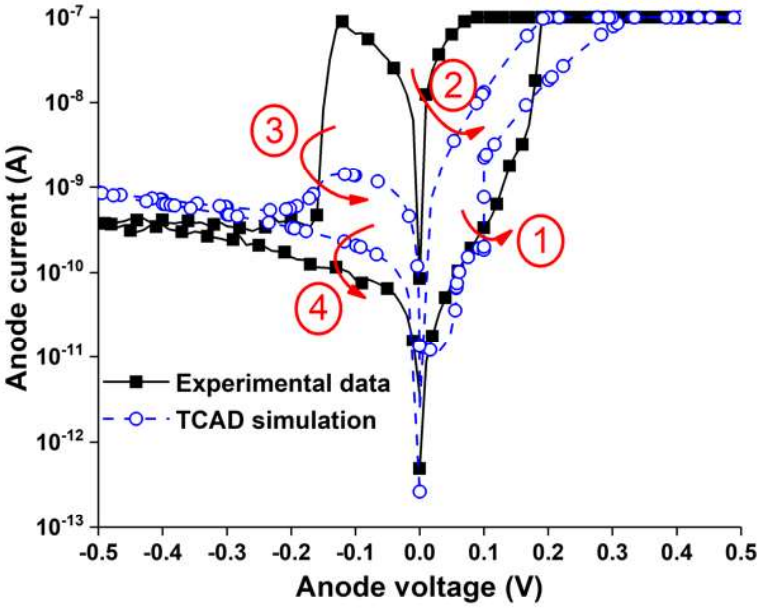


Fig. 55. Comparison of experimental data and simulation results of an I-V hysteresis sweep. The compliance current is set at 100 nA.

Fig. 56. shows the set voltage for different ramp-rates, simulated using the TCAD model. Set voltage is defined as the voltage at which the current reaches 100 nA during I-V hysteresis sweep. This relationship between ramp-rate and set voltage was reported experimentally, providing further validation of the model.

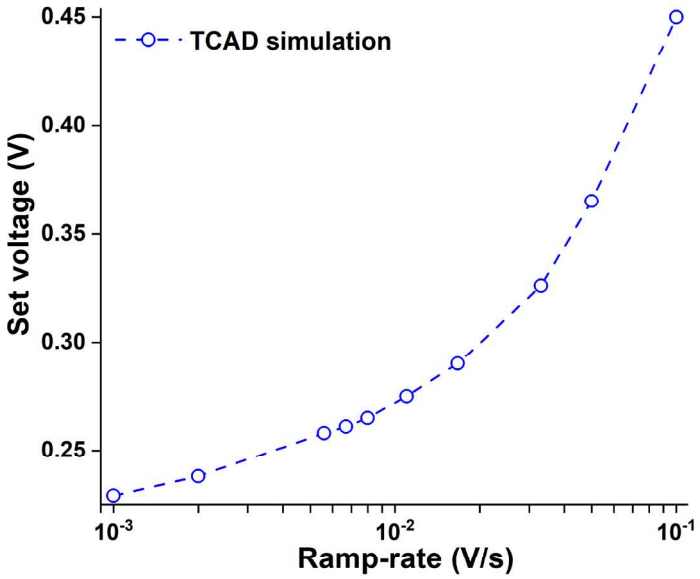


Fig. 56. Simulation results of set voltage vs. ramp-rate.

Radiation Effects on RRAM synapse

TID Effects on RRAM Analog Synapse

RRAM arrays can be used as synaptic weight elements in DNN vector-matrix multiplication. In the configuration that we are considering here, the output of one column in the vector-matrix operation is an output current (I_j), which is the sum of the products of row voltage (V_i) and the synaptic weight matrix elements (G_{ij}). Mathematically, this is represented as in Eq. (11).

$$I_j = \sum V_i * W_{ij} \tag{11}$$

The synaptic weight corresponds to the conductance across a two-terminal RRAM device. Conductance can be modified with an applied voltage pulse. One of the most desired features of an analog synapse for high accuracy pattern recognition is for the change in conductance with each programming pulse to be linear and symmetric. An ideal conductance change is depicted in Fig. 57. The increase in conductance is called potentiation and the decrease is termed as depression.

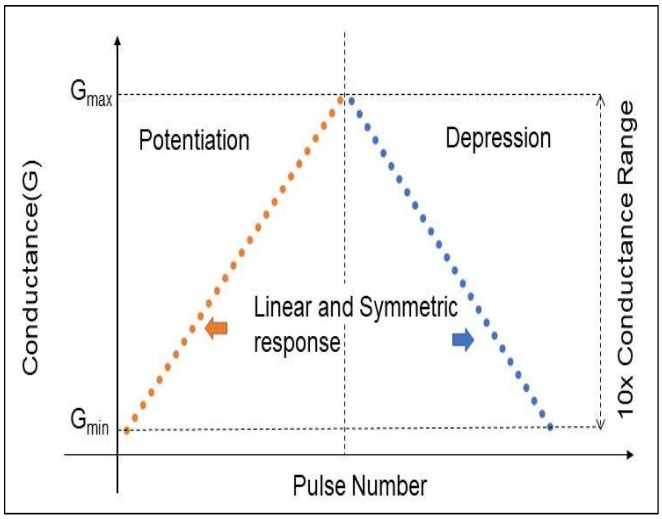


Fig. 57. Ideal conductance characteristics of a synaptic device.

Unfortunately, most RRAM device technologies exhibit highly non-linear weight update behaviors. The non-linearity factors α_p (for potentiation) and α_d (for depression) are parameters for the functions in the following eqns.

$$G_{norm_p} = \frac{1}{\alpha_p} * ln\left(w * e^{\alpha_p} + (1 - w)\right), \tag{12}$$

$$G_{norm_d} = 1 - \frac{1}{\alpha_d} * ln\left(w * e^{\alpha_d} + (1 - w)\right), \tag{13}$$

where

$$G_{norm_{p,d}} = \frac{G - G_{min}}{G_{max} - G_{min}}, \tag{14}$$

and

$$w = \frac{P}{P_{max}}, \tag{15}$$

In Eq. 14, G is the conductance variable, and G_{min} is the minimum and G_{max} is the maximum conductance in a potentiation or depression cycle. For the variable w in Eqs. (12), (13), and (15), P = pulse number corresponding to conductance G and P_{max}= maximum number of pulses. As the values of α_p and α_d approach zero, the conductance change becomes more linear. The aim of our current research was to investigate how the analog operation and linearity of Ag-$Ge_{30}Se_{70}$ CBRAM is impacted by TID. To do this, we focus on TID effects on the four parameters: G_{min}, G_{max}, and non-linearity factors, α_p and α_d.

A cross-section of the Ag-$Ge_{30}Se_{70}$ CBRAM device fabricated, tested, and analyzed for this paper is shown in Fig. 39. These devices have shown excellent radiation tolerance as binary memory elements. However, as will be shown, the analog switching characteristics and linearity of these new devices are deleteriously impacted by ionizing radiation exposure.

An optical micrograph of the fabricated CBRAM device is shown in Fig. 58.

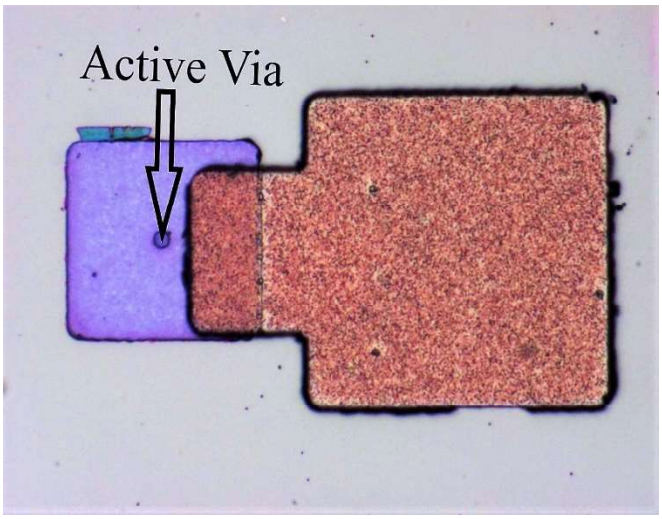


Fig. 58. Optical top-view image of Ag-$Ge_{30}Se_{70}$ CBRAM device.

Test Setup

CBRAM die were cut from whole 4-inch wafer and two dies were selected for the work. One acted as the control sample and the other as a test die. Before placing the test die in the radiation chamber, the devices were subjected to dc cycling using an Agilent 4156 C to ensure their functionality. For this experiment, 4 devices were selected each from the control as well as test dies All devices tested were 2 μm in diameter. Both the control and the test devices were cycled at 1mA of compliance current. After the dc cycling, the devices were subjected to pulse testing. The pulse test setup is shown in Fig.59.

The test die was irradiated with ^{60}Co gamma-rays in the Gamma cell 220 irradiator at ASU. The sample electrodes were left floating during exposure. This testing condition was chosen keeping in mind the application of such devices as weight elements in a DNN network with offline training capability. In offline training, the training is carried out using software and the trained weights are loaded to the synaptic arrays of the neuromorphic hardware by one-time programming and then only the inference or classification is performed on the hardware. The dose level inside the radiation chamber was 227 rad (SiO_2)/min for $Ge_{30}Se_{70}$. The devices were removed from the radiation chamber after the set dose was reached to be tested at room temperature. For pulse testing, increasing amplitude pulses were applied to the anode of the test device. Increasing amplitude voltage pulses from 0.2 V to 1.5V (pulse width=100μs and 50mV pulse step) for potentiation and -0.2 V to -1.2 V (pulse width=100μs and 50mV pulse step) for depression were applied. Each pulse was accompanied by a read pulse of 100mV with a pulse width of 100μs. The cathode was connected in series with a 2.2 kΩ resistor. The amplitude of the applied voltages (V_1) was measured on channel 1 of the Agilent DSOX2024A oscilloscope. The voltage across the resistor (V_2) was measured on channel 2 of the oscilloscope. The resistance of the device was calculated using a simple voltage divider equation as

$$R_{CBRAM} = R_L * \frac{V1-V2}{V2} \ , \tag{16}$$

The devices were tested for TID of up to 1Mrad (SiO_2). A summary of the dose step and exposure time is shown Table II. After each dose step, the test device, as well as the control sample, were cycled multiple times at room temperature. The conductance values at each TID level were collected in the form of look up table for processing in CrossSim. CrossSim is an open source software developed by Sandia National Laboratories for simulating crossbar simulators using RRAM devices [2]. It helps in determining a device's applicability to neuromorphic computing by taking a statistical device conductance data of multiple Potentiation and Depression cycles and then convert it into a lookup table. The data is used to simulate a training of a neural network with weights based on the conductance of RRAM device.

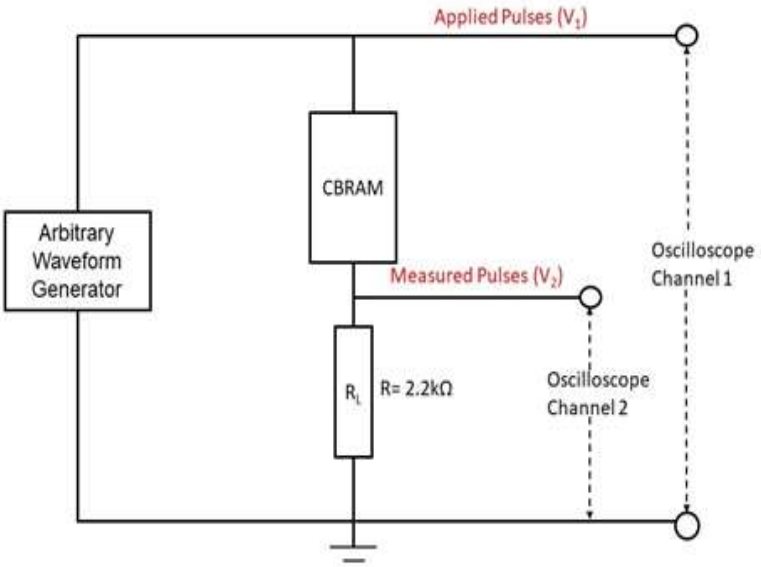


Fig. 59. Pulse test setup for radiation testing.

Table IVI
Dose Step and Exposure Time

DOSE STEPS (krad)	EXPOSURE TIME (hrs.)
45	3.30
100	4.04
300	14.68
400	7.35
600	14.68
800	14.68
1000	14.69

Experimental Results

Fig. 60 shows the potentiation and depression characteristics of the control and pre-irradiated samples. The figure shows that prior to irradiation the devices exhibited good linearity over a suitable conductance range. For potentiation, the pre- irradiated cell had an average G_{max} = $3.2x10^{-3}\Omega^{-1}$, G_{min} = $0.5x10^{-3}\Omega^{-1}$, and a non-linearity factor α_p = 0.58. For depression, an average G_{max} = $3x10^{-3}\Omega^{-1}$, G_{min} = $0.5x10^{-3}\Omega^{-1}$, and the non-linearity factor α_d = 0.74.

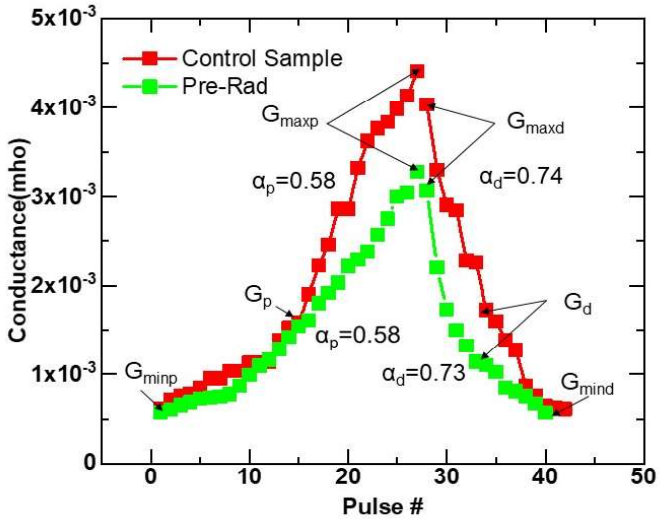


Fig. 60. Potentiation and depression characteristics of Ag-$Ge_{30}Se_{70}$ CBRAM.

The average neuromorphic response of the devices at various dose steps is depicted in Figs. 61 - 67.

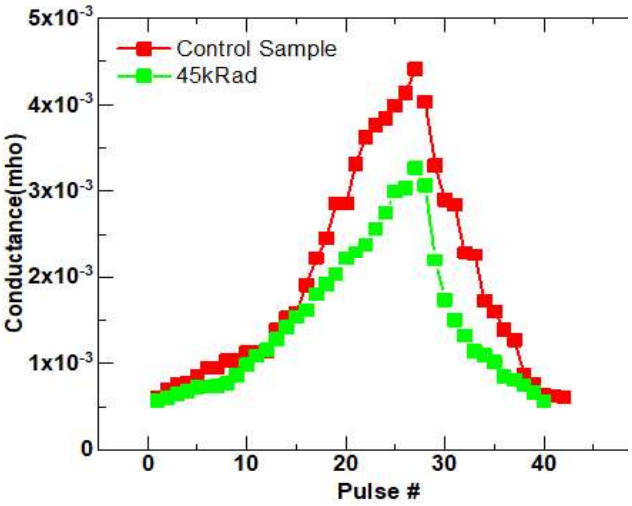


Fig 61. Potentiation and depression characteristics of the device at 45krad.

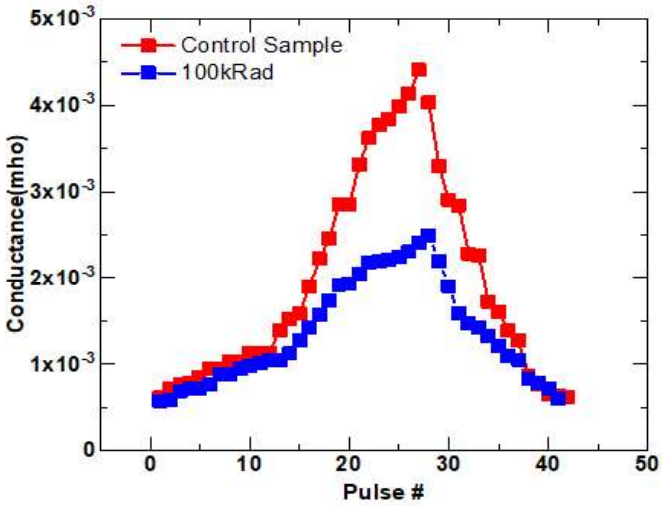


Fig 62. Potentiation and depression characteristics of the device at 100krad.

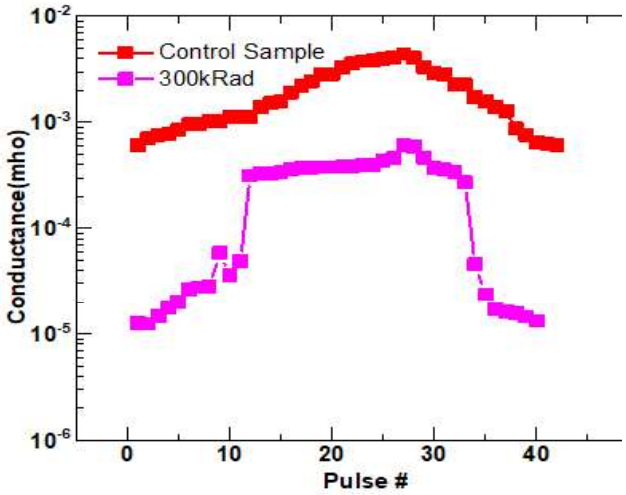


Fig 63. Potentiation and depression characteristics of the device at 300krad.

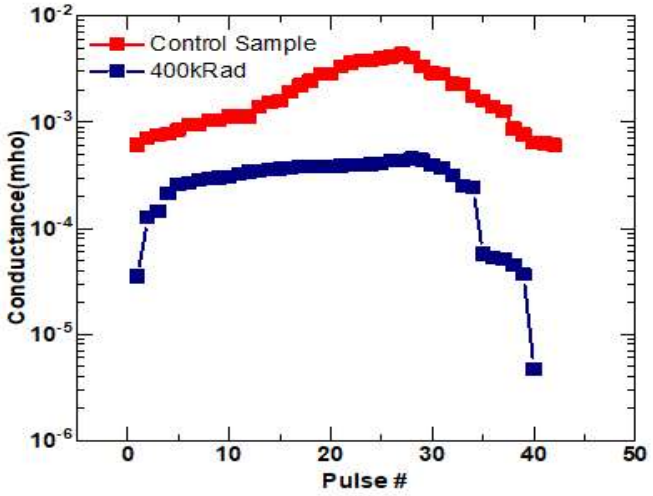


Fig 64. Potentiation and depression characteristics of the device at 400krad.

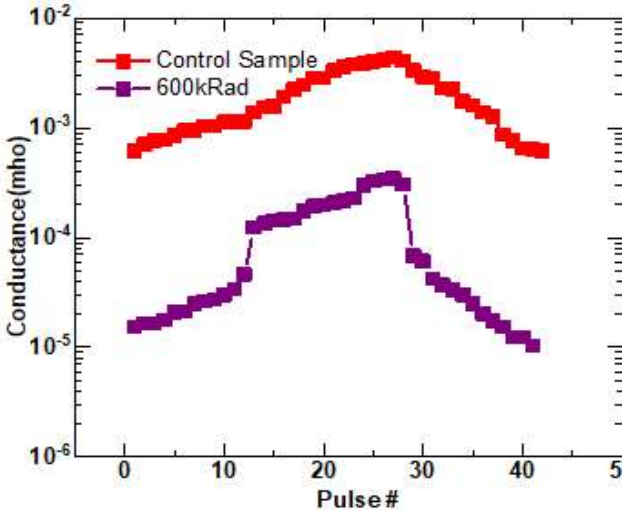


Fig. 65. Potentiation and depression characteristics of the device at 600krad.

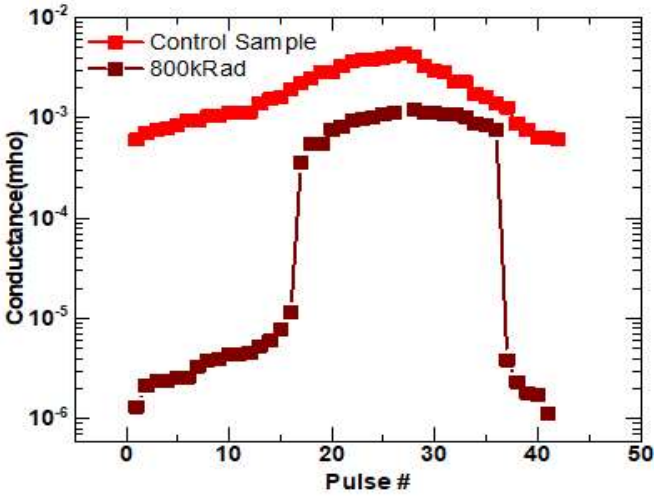


Fig. 66. Potentiation and depression characteristics of the device at 800krad.

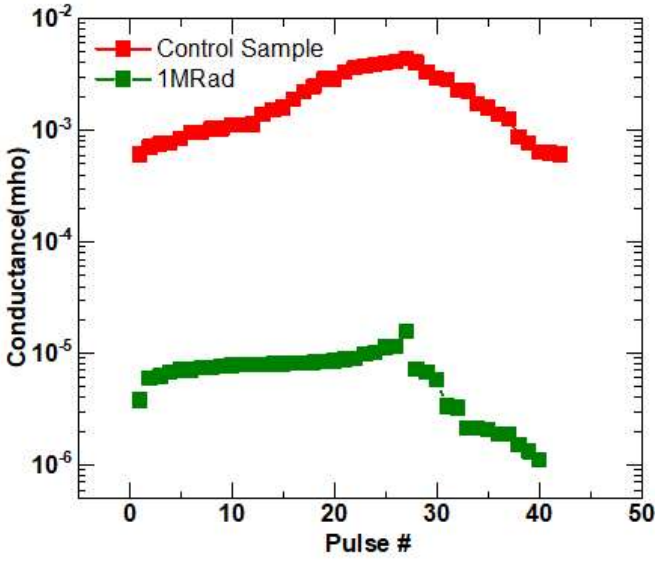


Fig. 67. Potentiation and depression characteristics of the device at 1Mrad.

Considering the neuromorphic responses of the irradiated devices, it is evident that the filament modulation becomes more erratic and uncontrollable with increase in TID levels. The maximum and minimum conductance state i.e. G_{max} and G_{min} are severely affected at high TID. The effects of irradiation on parameters G_{min} and G_{max} for potentiation and depression are shown in Fig. 68 and Fig 69, respectively. It is evident from this data that the neuromorphic programming response of the device is significantly impacted by increasing TID levels. Up to a TID level of 100 krad (SiO_2), the minimum and maximum conductance show a small, but measurable reduction. However, above 100 krad (SiO_2) we observe a catastrophic decrease in the G_{min} and G_{max} parameters.

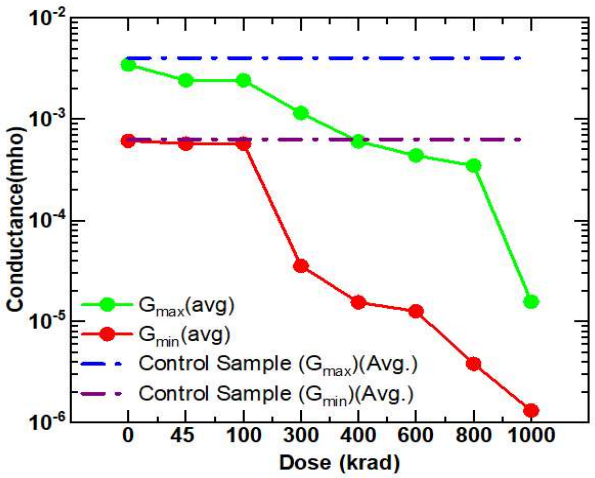


Fig. 68. Minimum and maximum conductance change with increasing TID levels for potentiation.

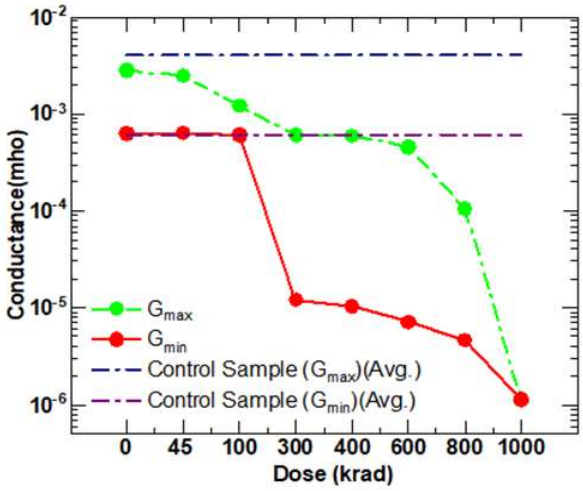


Fig. 69. Minimum and maximum conductance change with increasing TID levels for depression.

Fig. 70 plots the change in the non-linearity factors, α_p, and α_d. Both factors show large and non-monotonic changes above 100krad. It should be noted here that the non-linearity is of concern only when the conductance change is within ±10% of the pre-irradiated conductance. When the conductance reduces drastically, the non-linearity factor becomes obsolete.

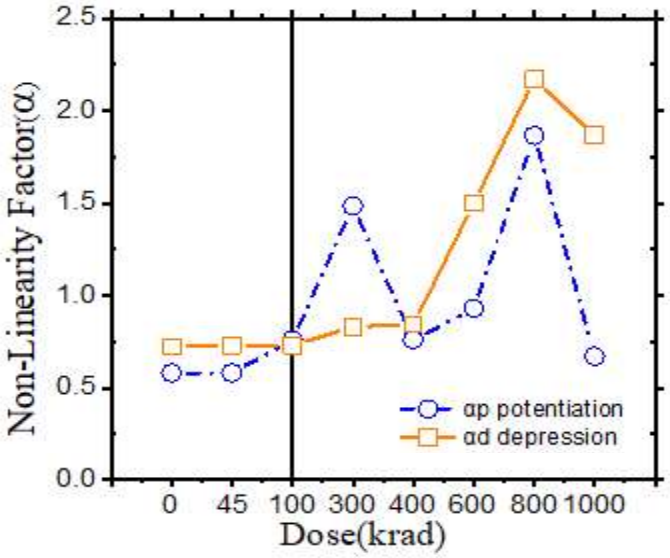


Fig. 70. Effect of irradiation on non-linearity factors, αp, and αd.

To investigate the drop in conductance and the effect of radiation environment on the radiated biased devices, Field Emission Scanning Microscopy was performed as shown in Fig. 71.

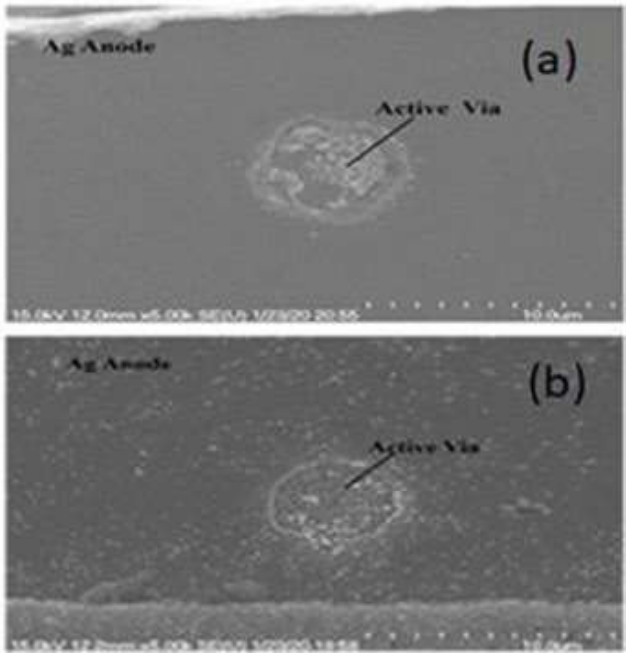


Fig. 71. FESEM images (a)biased pre-radiated and (b) biased irradiated device.

Discussion

Previous XRD studies have demonstrated the presence of several phases such as Ag_2Se and Ag-$GeSe_6$ for an Ag-$Ge_{30}Se_{70}$ material stack upon exposure to radiation. Formation of such crystallites on top of the electrode reduces the availability of active Ag, which is necessary for making contact to and for broadening the existing filament. Hence maximum conductance is not reached for the same train of applied pulses. In addition to reducing the available active metal, another impact that the new material phases have is that they increase the resistivity of the top electrode, which leads to a large voltage drop at the contact. Therefore, the pulse amplitude is not large enough to bring about the required change in the conductance. Another potential cause of reduced conductivity is the oxidation of Ge to GeO_2 in a radiation environment as previously studied in [4]. Due to oxidation, the conductivity of the switching layer is reduced. Hence the switching layer becomes more insulative and is not able to transport Ag^+ ions to the filament. Under these post-irradiated conditions, the neuromorphic response of the device is severely impacted due to the formation of different Ag phases as well as the reduction in the conductivity of the switching layer.

To better understand the effect of radiation on linearity and accuracy of a neural network, Sandia's Cross-Sim program was used. MNIST small image data set, which is an 8X8 version of MNSIT handwritten digit images, was used for the training purpose. A two-layer neural network with 64 inputs, 36 hidden nodes and 10 output nodes were used for image classification. A normalized plot of classification accuracy versus the neural network iteration (epoch) for the test device at various TID levels is depicted in Fig 72. The accuracy was seen to drop non-monotonically from 94% for Pre-rad condition to 87% at TID level of 1Mrad.

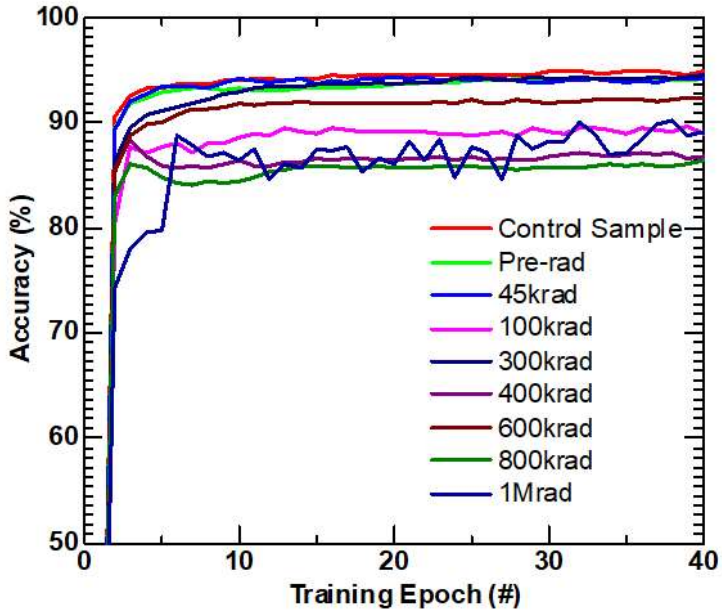


Fig. 72. Accuracy after training exposed devices at different TID levels.

TID Effects on Pulsed Programmed RRAM Synapse

The pulse responses of the control and the test devices prior to irradiation are shown in Figs. 73 and 74, respectively. The pulse amplitude and the pulse width were chosen to obtain a smooth monotonic change in conductance for both potentiation and depression. The exposed CBRAM devices were tested at the six TID levels: 50 krad, 100krad, 250krad, 500krad, 700krad, and 1Mrad($Ge_{30}Se_{70}$). The programming responses at each TID level except 1Mrad ($Ge_{30}Se_{70}$) are shown in Figures. 75-79. The maximum conductance level for the control sample obtained at the same time as the step stress measurements on exposed devices is indicated by the red line on Figures. 74-79. For clear observability of the conductance change profile vs. pulse number, only 5 of the 20 potentiation/depression cycles are shown. Exposed responses after 1Mrad ($Ge_{30}Se_{70}$) are not provided since those devices failed to switch at that dose level.

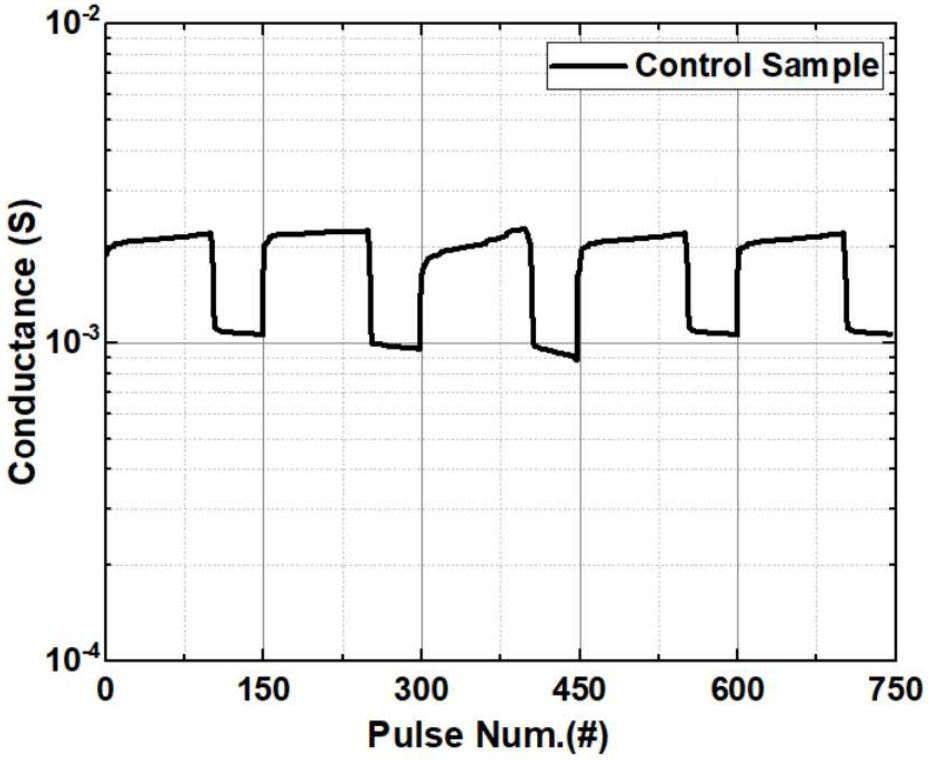


Fig 73. Potentiation and depression characteristics of the device of the control sample.

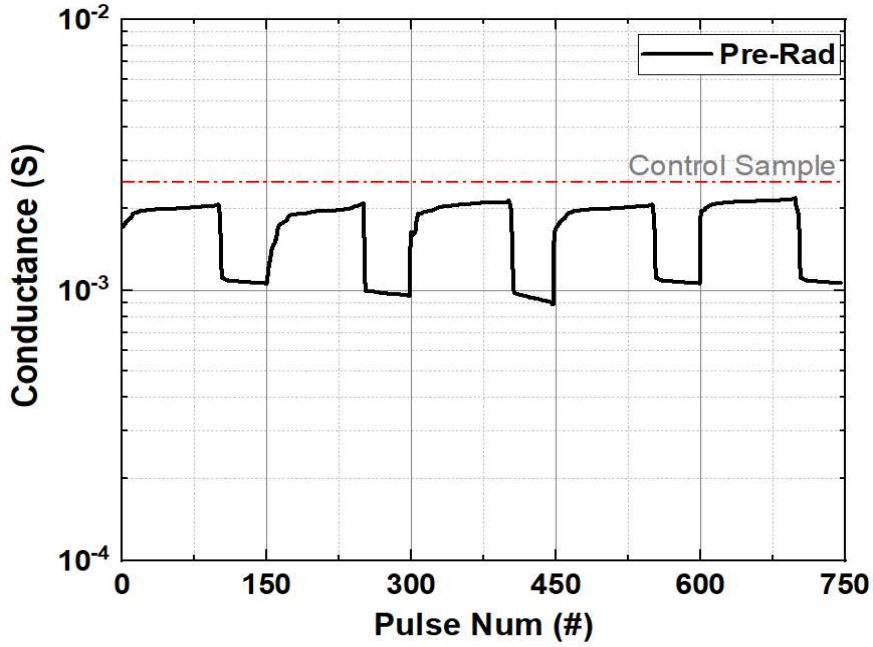


Fig. 74. Potentiation and depression characteristics of the pre-rad device.

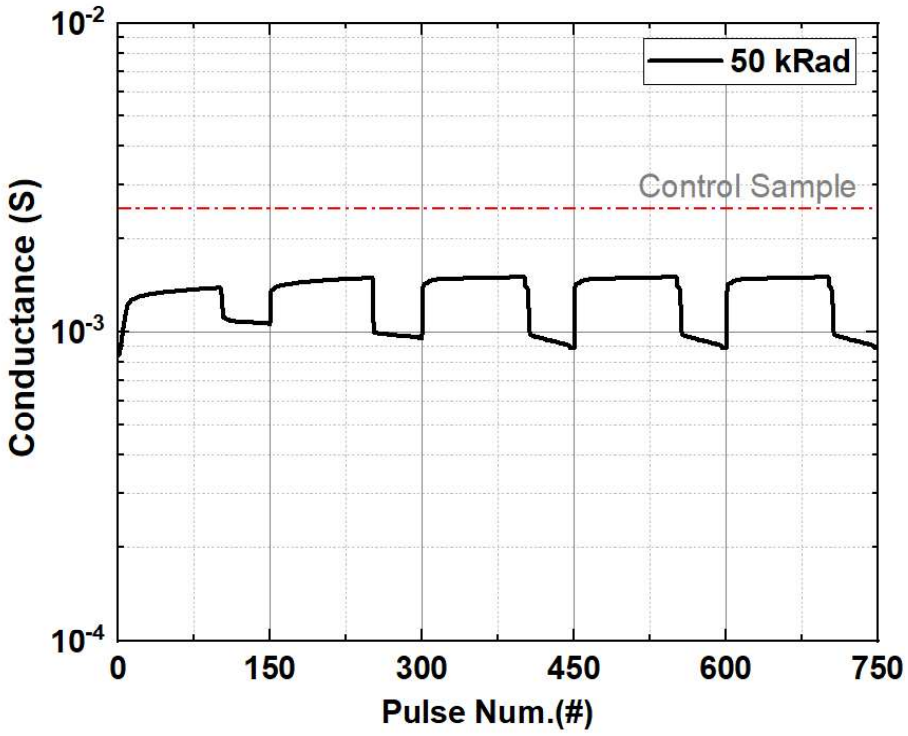


Fig. 75. Potentiation and depression characteristics of the device at 50kRad.

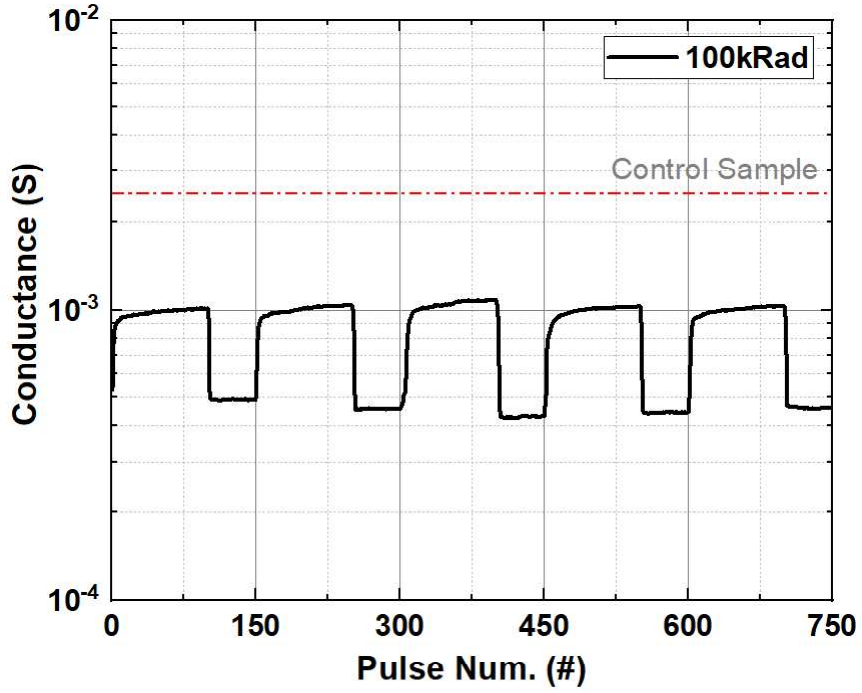


Fig. 76. Potentiation and depression characteristics of the device at 100kRad.

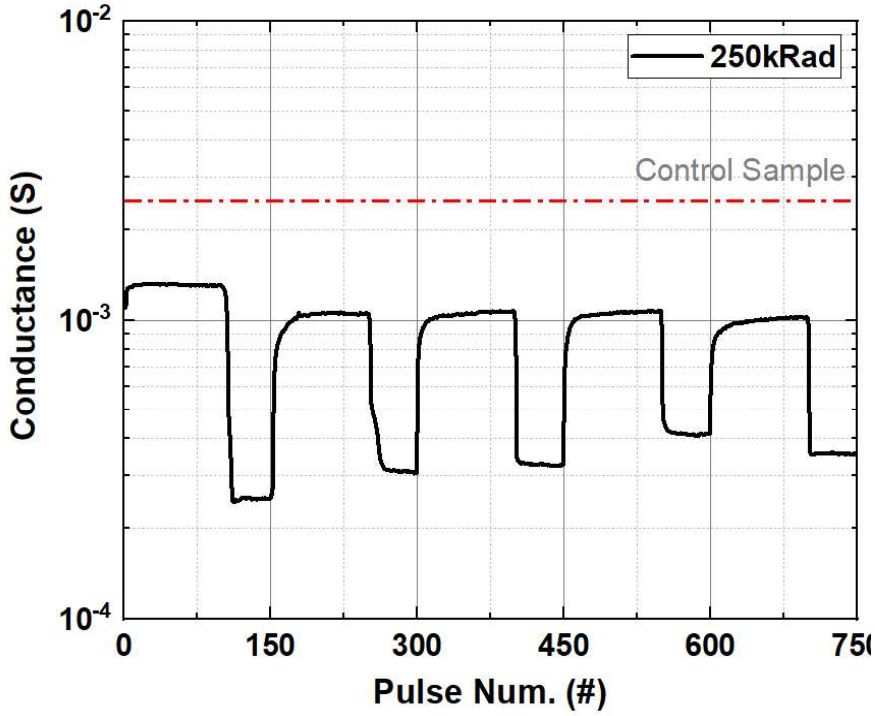


Fig. 77. Potentiation and depression characteristics of the device at 250kRad.

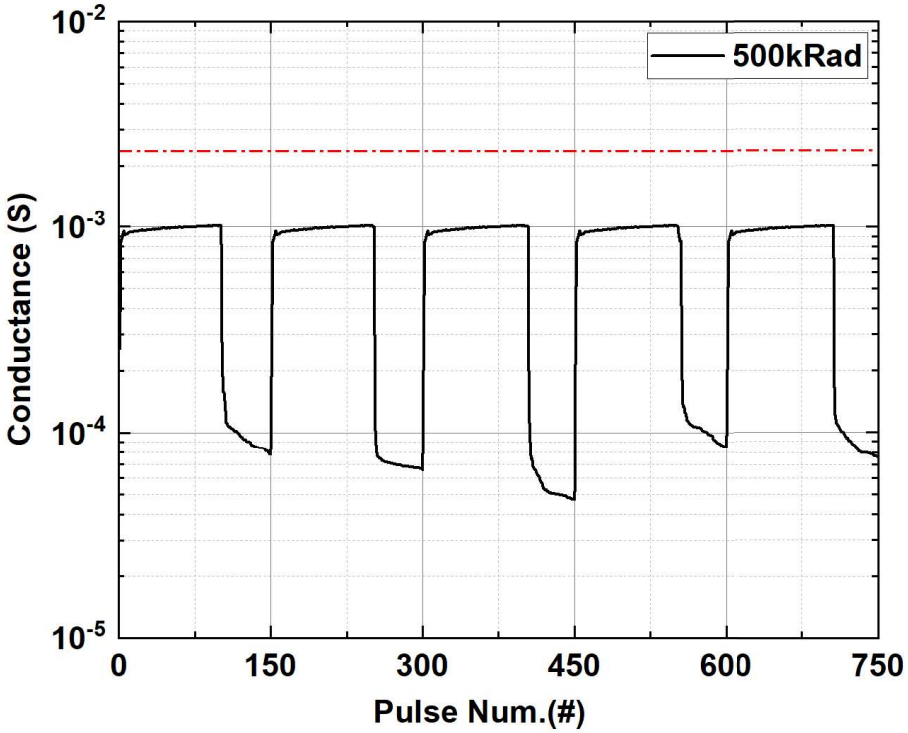


Fig. 78. Potentiation and depression characteristics of the device at 500kRad.

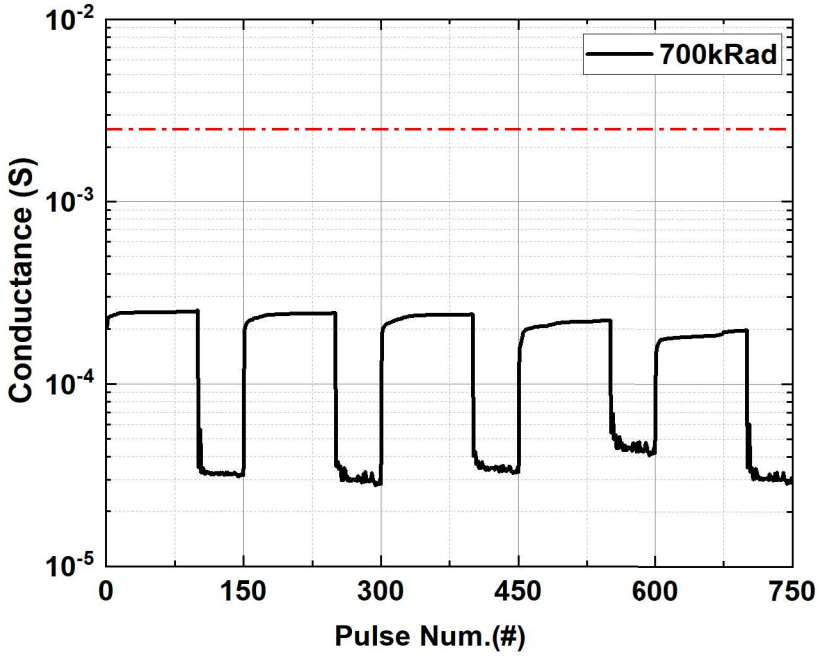


Fig. 79. Potentiation and depression characteristics of the device at 700kRad.

As is typical for many RRAM technologies, even prior to irradiation, the pulsed conductance response is highly non-linear. As the figures show, the conductance change per pulse is large at the beginning of each pulse cycle, then the rate of change slows substantially until the conductance converges to a maximum or minimum value. While linear potentiation/depression characteristics are the most ideal for analog synapses, and indeed, some devices can show this (see Figure 27 in previous chapter), recent RRAM-based accelerator studies have presented algorithms for achieving good accuracy even with non-linear characteristics. What is critical to training accuracy is that the range and cycle-to-cycle variation of the conductance behavior does not change significantly over time and stress. It is evident from the post-irradiation data that the range of conductance change is altered by TID. For potentiation (thickening of the conductive filament) the maximum conductance achieved decreases monotonically with increasing TID. This effect becomes pronounced for TID levels of 100 krad ($Ge_{30}Se_{70}$) and above. The depression programming (thinning of the conductive filament) is also affected with increasing TID levels. We observe that the minimum conductance is monotonically decreasing with increasing TID as well.

Modelling Radiation Effects on CBRAM-based Neural Accelerator

The TID impact on Ag-$Ge_{30}Se_{70}$ CBRAM devices is further studied by modelling these devices in an analog neural training accelerator. Here we use the open-source CrossSim simulator developed at Sandia National Laboratories to perform the training. An artificial two-layer neural network, shown in Fig. 80, was simulated using the pre- and post-rad data shown in Figures.

48-53. The neural network was trained using a backpropagation algorithm. This algorithm uses three important kernels: vector matrix multiply (VMM), matrix vector multiply (MVM) and outer product update to train the neural network. For this study, we use a small version (8 X 8 pixels) of handwritten digits from the 'Optical Recognition Handwritten Digits' dataset. This two-layer neural network consisting of 64 input, 36 hidden and 10 output nodes for small image dataset was used for training.

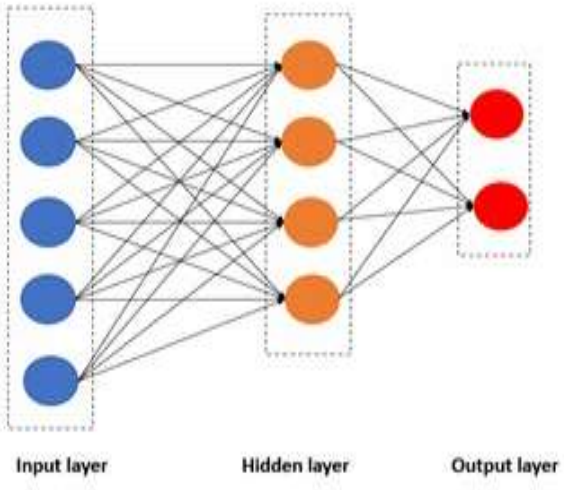


Fig. 80. Simulated 2-layer neural network.

To simulate the response of the Ag-$Ge_{30}Se_{70}$ CBRAM devices in the neural network, a look up table is created containing $\Delta G/G$ response of the devices to multiple potentiation and depression cycles at a given TID level. The training accuracy of the neural network was seen to degrade monotonically from 86 % prior to irradiation to 15 % at the 700 krad ($Ge_{30}Se_{70}$) dose level. A plot of classification accuracy versus the neural network iteration (epoch) for the test device at various TID levels is shown in Fig. 81.

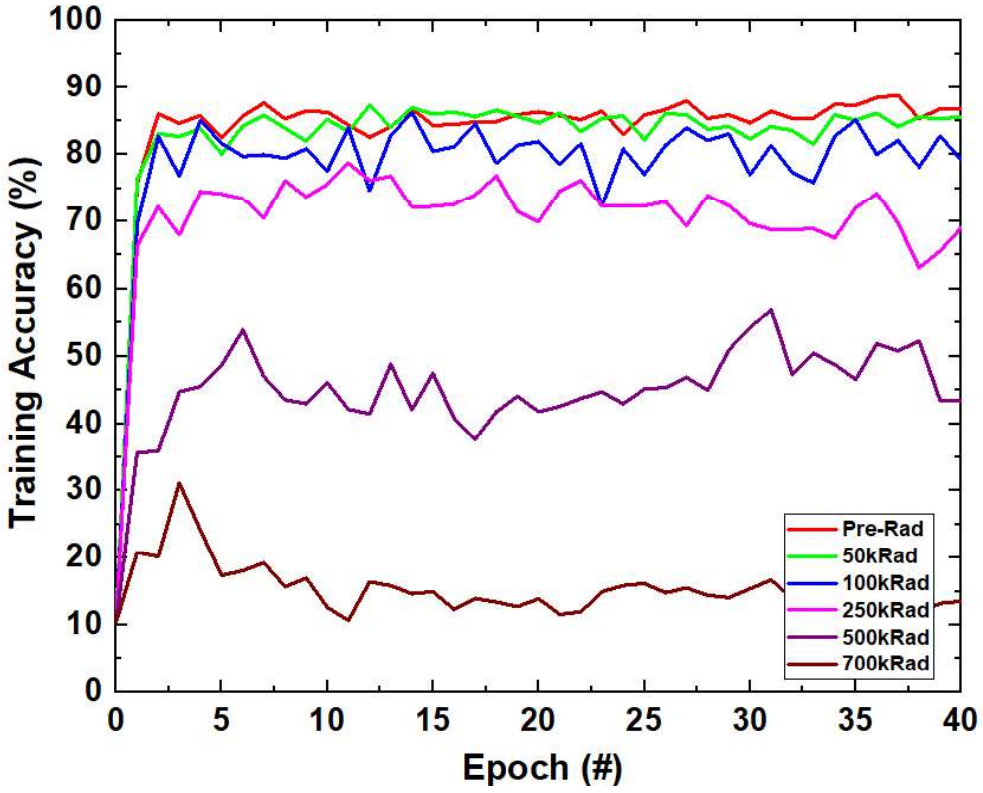


Fig. 81. Training accuracy of Ag-Ge30Se70 CBRAM devices at different TID levels.

Task 2. Array-level Testing/Simulation of Radiation Effects on Neuro Crossbar

Task 2.1 Pre-rad testing of neuro-crossbar RRAM array

Design and fabrication 256×256 1T1R RRAM Array

One-Transistor-One-Resistor (1T1R) RRAM Cell Operation

One-transistor-one-resistor (1T1R) array is one of the common RRAM array architectures. The selected cell can isolate from unselected cell to minimize the crosstalk issue. The RRAM top electrode connects to the bit-line (BL) while its bottom electrode connects to the transistor drain. The transistor gate connects to the word-line (WL), therefore, tuning the WL voltage can control the compliance current delivered to the RRAM. The transistor source connects to the source-line. 1T1R RRAM cell has four operating states including forming, read, reset, and set, as shown in Fig. 82. In forming (Fig. 82a), the bit-line (BL) is pulsed, the transistor gate (Vg) has a DC voltage loaded, and the source-line (SL) connects to ground, which initializes the cell for resistive switching and sets the OxRAM device into a resistor (R) by applying large positive pulse to form conductive oxygen vacancy filaments (CF) in the HfO_x layer. Followed by a read operation that can confirm CF formation. During read (Fig. 82b), the cell resistance is calculated by the bit-line voltage divided by its current. The bit-line current results from a positive pulse on BL, a fully turn-on DC voltage on Vg, and 0V on SL. Within the reset (Fig. 82c), the BL is biased to ground, a positive pulse and fully turn-on voltage are applied on SL and Vg, respectively, resulting in a negative bias on the top electrode of OxRAM device. The negative bias ruptures the conductive filaments in OxRAM device, forcing R into a High Resistance State (HRS). During the set operation (Fig. 82d), the RRAM is set into non-HRS with a positive pulse applied to bit-line, a turn-on voltage applied on the transistor gate, and grounding the source-line. The value of non-HRS is dominated by the transistor gate voltage. The read is necessary after programming operations including forming, set, and reset to confirm the expected states of RRAM cell.

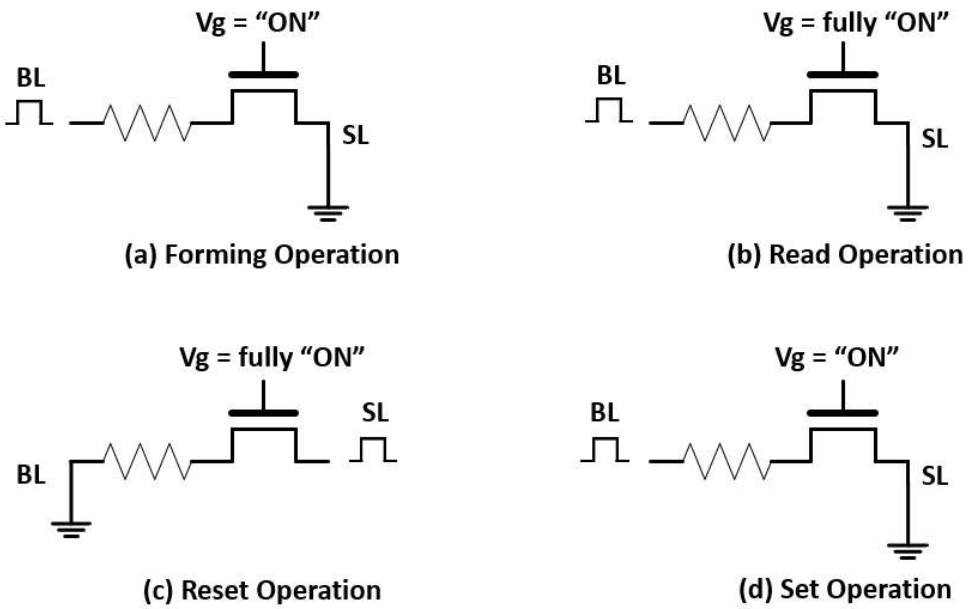


Fig. 82. The operation schematics of 1T1R cell: (a) forming, (b) read, (c) reset, and (d) set.

The RRAM Prototype Chip Design

Based on the operation of 1T1R configuration, a prototype test chip was designed and fabricated (Fig. 83) to characterize the synaptic performance and total ionizing dose response of OxRAM arrays. In this case, instead of reading/writing/storing as an analog or binary weight, the RRAM resistance (conductance) can be set and held in one of various multi-level states.

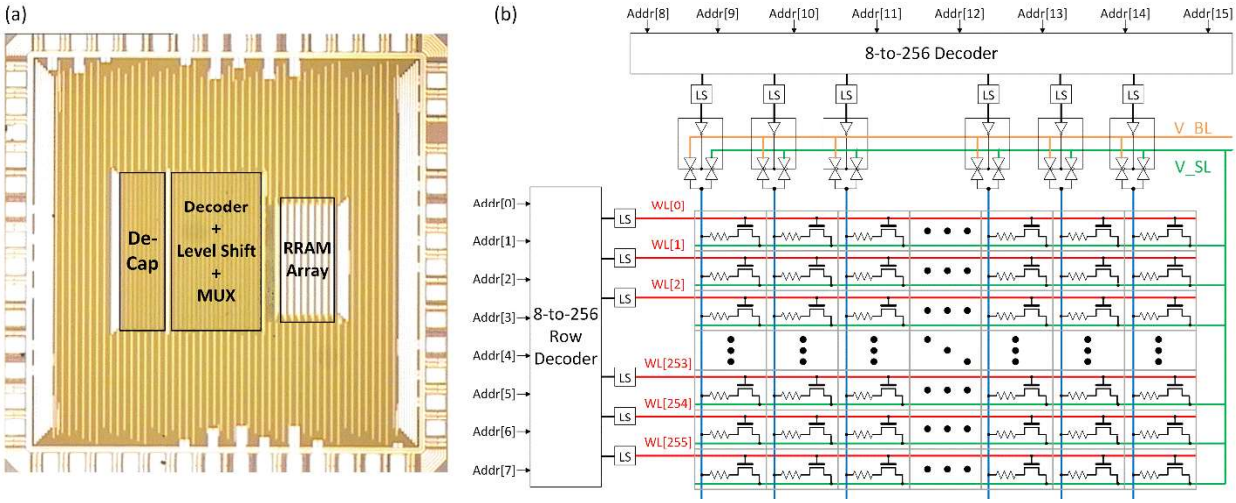


Fig. 83. (a) Die micrograph of our 64kb RRAM test chip. (b) Circuit schematic of RRAM prototype chip including: 256×256 1T1R RRAM array, column and row decoders, level shifter, and mux based on transmission gates.

The chip has CMOS peripheral circuits to control read/write into a 64Kb (256 × 256) 1T1R array. The test chip was fabricated in a 90nm technology node by Winbond Electronics Corporation

(Taiwan). Each RRAM cell consists of an NMOS transistor (high-voltage I/O) in series with a switching OxRAM resistor, i.e., the 1T1R configuration. The switching layer of this OxRAM device is HfO_x locates between the first and second layer metals, M1 and M2, respectively. Fig. 83b shows a representative schematic of the test chip. It diagrams the array, digital decoder logic, voltage level shifters (LS), and transmission gate switches for multiplexing. When a 16-bit address is fed into row and column decoders, only one 1T1R RRAM cell is selected for programming or read. For this selected cell, the digital Word-Line (WL) voltage applied to the NMOS transistor gate is a level-shifted positive voltage high and either the analog Bit-Line (V_BL) or Source-Line (V_SL) is pulsed to a positive voltage depending on the selected operation (forming, set/reset programing, or read). If V_BL is pulsed, V_SL is fixed to 0V and vice-versa. All the cells in the same column as the selected cell share the same V_BL, but transistor gate potentials are low. All cells in the same row as the selected cell share the same high WL voltage on transistor gates, but the internal bit lines are shorted to V_SL. All the cells in different rows and columns have 0V applied to transistor gates and internal bit line nodes are shorted to V_SL. The read voltage in this work is chosen to be 0.2V to prevent read disturb issues.

Electrical testing of 256×256 1T1R RRAM Array

Testing System and Programming Setup

A National Instruments (NI) PXI system is adapted to build the testing hardware setup (Fig. 84a), which contains several subsystems, including source measurement units (SMUs) to measure the cell resistance, signal multiplexer (MUX) to multiplex analog bit-line/source-line signals, arbitrary waveform generator (AWG) to generate programming pulse signals, digital waveform instrument (DIO) to generate/acquire digital I/O signals, and analog output modules (AO) to generate gate voltages. For the software, an automated testbench was created in LabVIEW. Fig. 84b shows the front panel of the main testing virtual instrument (VI) file. The control button, named without “Array”, indicates the run operation on the selected RRAM cell whose address is located in the “Address (16 bits)”. The 64Kb (256 × 256) 1T1R array was divided into 16 subarrays for programming. The control button, named with “Array”, indicates run operation on the selected subarray whose number is presented in “Subarray_Program”. The controllable parameters, including pulse width (PW), pulse voltage (V), gate voltage (Vg), read voltage, upper and lower bound of resistance, enable the write-verify based programming scheme. During forming/set operation, the system conFig.s the multiplexer (MUX) to select an arbitrary waveform resulting in a applied voltage pulse on the Bit-Line (V_BL) and biasing the Source-Line (V_SL) to ground. For reset operation, it conFig.s MUX to select an arbitrary waveform resulting in a voltage pulse applied on V_SL while V_BL is fixed to ground. During read operation, it conFig.s MUX to select SMU to obtain the RRAM cell resistance by measuring the column current of the active cell.

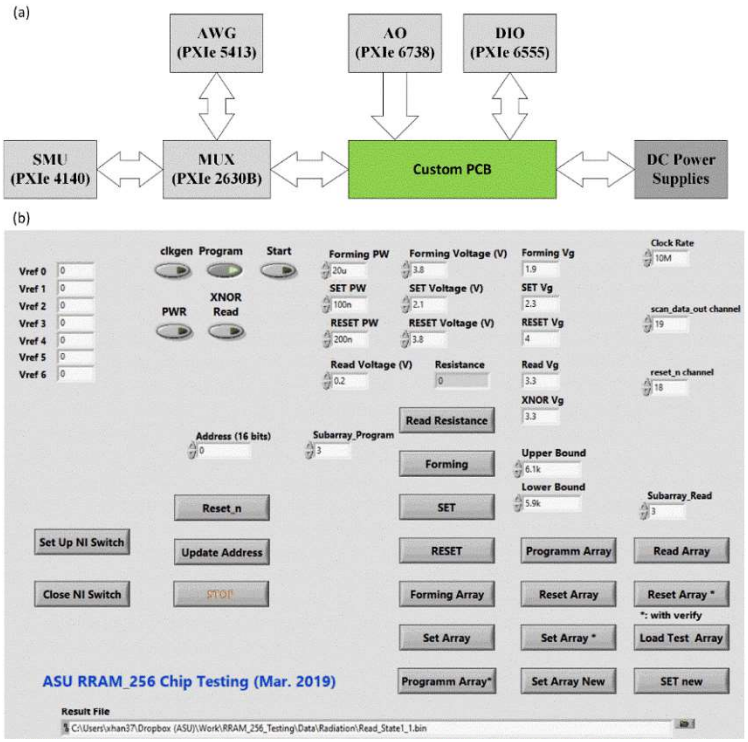


Fig. 84. (a) The diagram of the testing system setup. The chip is connected to the custom PCB via a test socket. (b) The main view of the LabVIEW automated testbench

Multi-State RRAM Verification

For our test chip, a simple, yet effective verification algorithm has been developed to enable multi-state resistances to be programmed in the array. Fig. 85 diagrams the algorithm for programming and verifying a multi-state resistance in an RRAM cell. The process goes through cell by cell. Before the verification, there are several steps, which are running the forming-read operation, reset-read operation, and set-read operation to check the functionality of each RRAM cell. The verification process has two components. HRS verification is performed through the reset operation in the reset loop (i). HRS verification fixes a distribution of HRS across the array by increasing the pulse voltage applied to the V_SL over multiple cycles of reset operations. The maximum number of HRS verification loops (M) is limited to 15 to avoid high resistance locking. Non-HRS verification is performed by the set operation in the write-verify loop (j), which is an outer-inner loop (j+k) depends on the variability of programming each RRAM 1T1R cell. For example, some cells would be programmed into the target range after 2 or 3 inner loops (k) in the current verification cycle. Some cells would need more inner loops or cannot be programmed into the target range under in the current bias conditions, thus, it is necessary to break the inner loop and start a new verification cycle. The maximum number of the inner loop is set to 10 to avoid endless loops and resistance locking. It should be noted that all non-HRS states will fit into resistance distributions lower than the range of targeted HRS values.

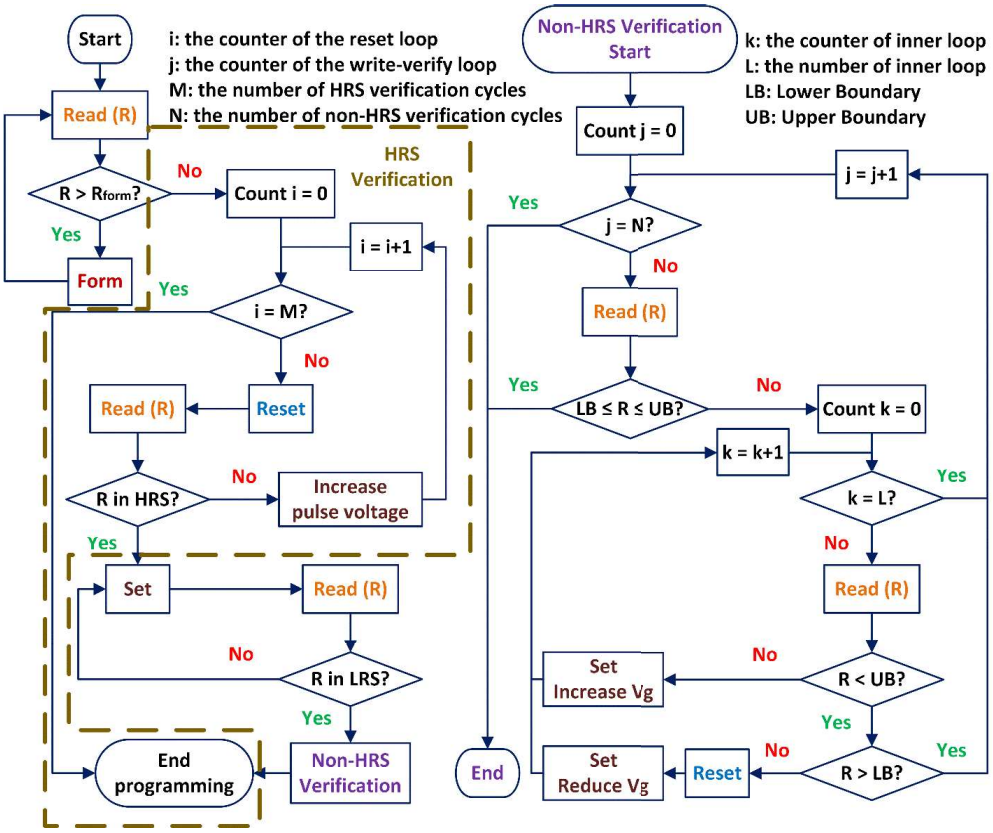
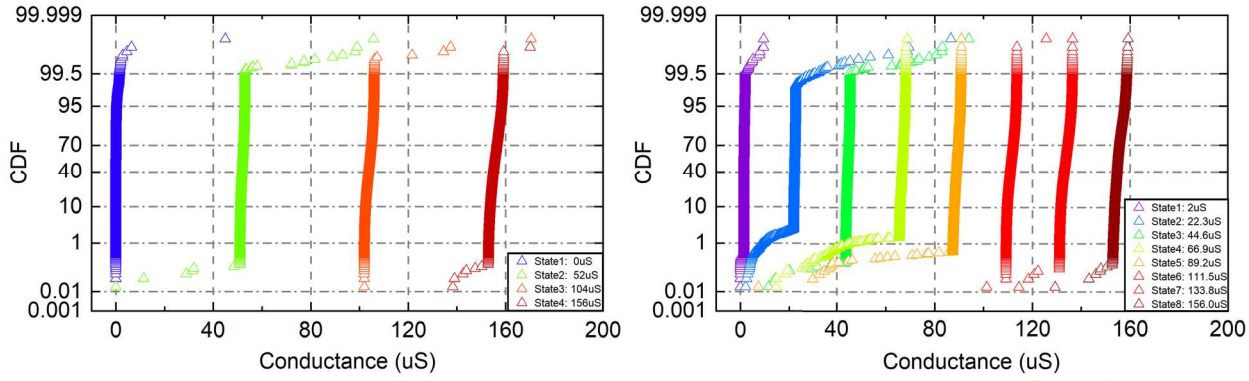


Fig. 85. Flow chart of programming and verifying multi-state resistance algorithm. The algorithm parameters are the number of verification cycles (M&N), the number of inner loop (L), upper boundary resistance (UB), lower Boundary resistance (LB), and counter i, j &k.

It is known that applying different WL voltages (i.e., Vg) can generate different non-HRS resistances in RRAM cells. Upper and lower boundary resistances for the non-HRS distribution are determined by the WL voltage (Vg) applied during the set operation. We used up to 30 loops for each state verification. The non-HRS verification process is as follows. First, read the cell resistance R. If R is within the targeted non-HRS range, the verification process ends. If R is above the upper boundary value, an operation with increased Vg is performed. This step is repetitive until the maximum loop number (L) is achieved or R is smaller than the upper boundary value. If R is smaller than the lower boundary value, a reset operation is performed to drive the cell back to HRS, followed by running set operations with decreased Vg. This step is repeated until the maximum loop number (L) is achieved or R is greater than the lower boundary value.

Based on the programming algorithm, we achieved 4-state conductance and 8-state conductance, corresponding to 2-bit and 3-bit weight precisions, in two RRAM prototype chips. Fig. 86 diagrams the cumulative distribution function (CDF) of the multi-state conductance for each chip. Each state has 4096 1T1R cells. Depending on the value of HRS and LRS, we determined the mean conductance value of first and last states, respectively. The targeted

mean conductance of intermediate states is achieved by adding equal interval conductance between HRS and LRS. All HRS verification cycles are limited to 15 cycles and non-HRS verification cycles are limited to 30 cycles. There are some stubborn cells that cannot be programming into targeted range after running all required verification cycles. Those stubborn cells slowed down the programming speed. Actually, programming 8-state conductance took couples months! To reduce the total test time, only four subarrays are programmed for radiation test, which means only selected 16 kb 1T1R cells in the 64 kb RRAM test chip were programmed and verified. The color map of the programmed 64 kb RRAM test chip identifying the selected 16 kb cells is shown in Fig 87. There are four subarrays which were programmed into 4 different conductance levels in the 256×256 1T1R array, and each state is programmed into 4096 cells.



(a) 4 states conductance of RRAM array

(b) 8 states conductance of RRAM array

Fig. 86 The cumulative distribution function (CDF) of (a) 4-state conductance (2-bit weight precision) and (b) 8-state conductance (3-bit weight precision)

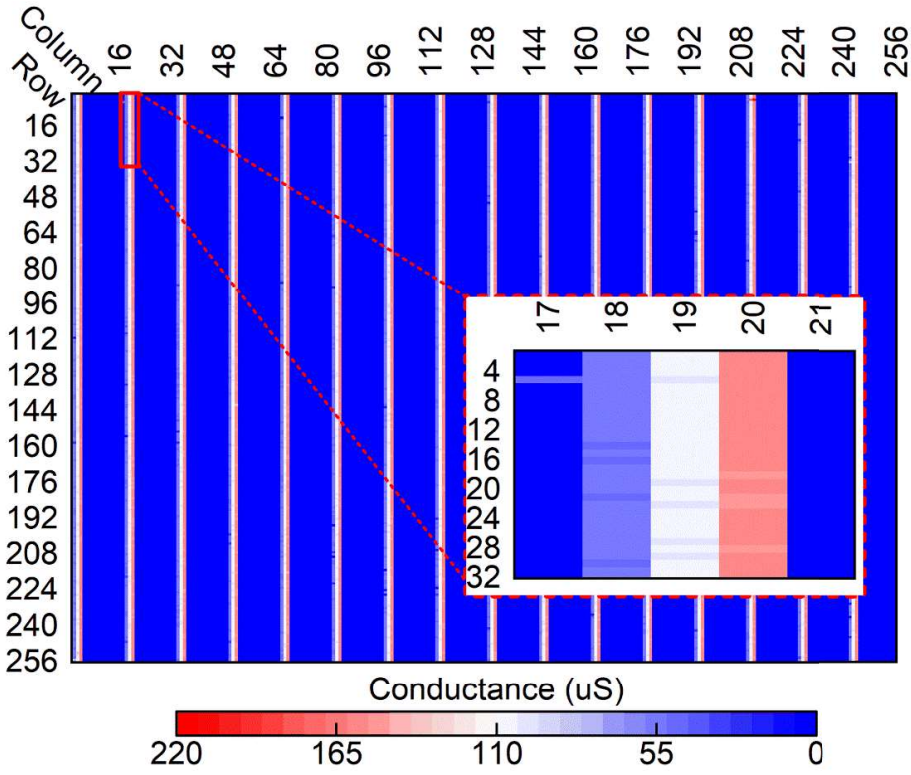


Fig. 87. Color Map of Programmed 4-State Conductance in a 256×256 1T1R Array. Each State has 4096 1T1R Cells

Task 2.2 Modeling and validation of radiation effects in core neuro-processing elements

Design full un-hardened RRAM-based integrated neuromorphic system.

Overview

The core function for our prototype neuromorphic system is a vector matrix multiply, or dot product operation. The approach we use to realize this function is to sum currents created from input voltages impressed across a set of resistors sharing a common lead. If the common is held constant, the sum is the dot product of the input row voltages and the resistor's conductances, i.e.,

$$\sum_{j=1}^{N} \frac{v_i}{R_{ij}} = \sum_{j=1}^{N} v_j \cdot G_j \tag{17}$$

The design shown in Fig. 87 illustrates the basic architecture of the analog VMM perceptron that we have designed. It integrates a front-end-of-line CMOS (foundry fabricated) with a back-end RRAM (ASU process). The CMOS portion of the chip includes row and column drivers, "pulsers", RRAM isolation switching, transimpedance amplifiers, digital-to-analog converters,

comparators, and 10b shift registers. The matrix resistors R_{ij} on Fig. 88 represent the RRAM deposited on top level metal at ASU.

The VMM chip implements a controlled pulsed write into and digitized readout of a nonvolatile RRAM cross-bar array. The array is integrated onto a CMOS chip through a specialized top metal layer planar device deposition process. While the matrix multiply operation is analog, the dot product sums are converted to a 10-bit digital word using an on-chip analog-to-digital converter. Each row or column in the array can be individually addressed and isolated to either be included or excluded for any given read or write operation. Also, on-chip is a variable width pulse generator used to control the width of the write pulses that alter and tune the array's RRAM conductances. Configuration of the control bits for array addressing, transimpedance gain, and test modes are all done through a simple shift-register chain.

The analog VMM operation is a scaled dot product of input row voltages and RRAM conductances. These dot product sums (currents) are converted to a 10-bit digital output using an on-chip analog-to-digital converter (ADC) on each column during Read operation. Each row or column in the RRAM array can be individually isolated for a given Read or Write operation, which is critical to mitigating cross-bar "sneak path" errors.

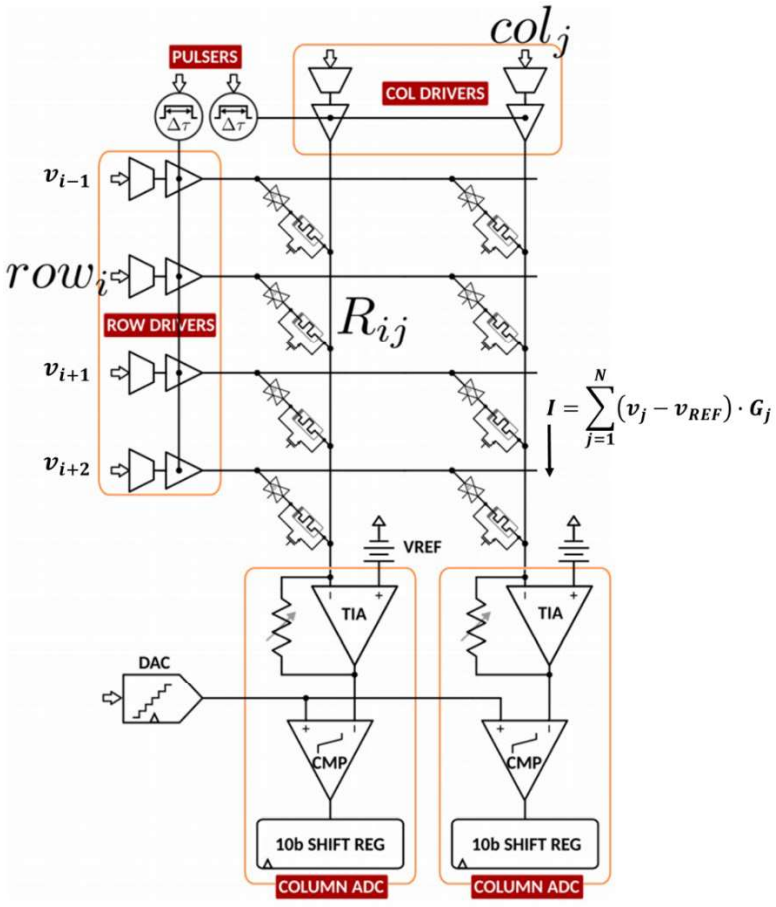


Fig. 88. Diagram of neuromorphic system architecture comprised of row drivers, column drivers, pulser, column ADC and resistive RAM (RRAM) crossbar array.

VMM Architecture

Array - The embedded RRAM array is divided into 8 columns. The VMM operation is 8 dot product operations performed in parallel, once down each column with the 8 row input voltages shared across the columns. Every row or column in the array has 2 mode bits that control how the array is connected to row and column drivers. This allows for subsets of the array to be selected for a set of operations while floating or “idling” the other ones. The eighth column is populated with fixed poly resistors instead of RRAM devices and serves as test column to evaluate the rest of the readout circuitry.

Readout - The dot product sum is realized by passing in external voltages to the array’s rows and summing the currents through a programmable gain transimpedance amplifier (Fig. 88). The transimpedance stage sums the 8 row device currents while simultaneously converting it into a voltage for the ADC (Analog-to-Digital Converter).

TIA (Transimpedance Amplifier) - the TIA has programmable gain making it simpler to scale the dot product or changes in it into a range that spans more ADC LSBs. The 10 transimpedance resistors are roughly logarithmically spaced over a range from 1 to 100 *(base) *k* Each of the 10 impedance resistors can be placed in parallel to create an effective transimpedance gain that is interpolated between any of the individual resistor gains. The TIA also ensures the summing node is held to a fixed reference voltage (due to feedback) which is necessary for dot product sum to compute correctly.

Analog-to-Digital Converter - The ADC area and power can dominate the energy efficiency and circuit density. For this design, the ADC is reduced to a single comparator and counter per column by use of a shared ramping DAC (Digital to-Analog Converter). This columnar style ADC amortizes the area and power cost of a traditional ADC over all the columns. It works by ramping a DAC incrementally through all its values while the counters simultaneously increment from zero; when the column comparator trips, that the ramp DAC has reached the same voltage as that output by transimpedance amplifier, which is the scaled dot product. At this point, the counter is stopped, and the value held for readout. The count is proportional to dot product value.

$$\frac{D}{2^{10}} \cdot V_{REF} = Z \cdot \sum_{1}^{8} v_i \cdot G_{ij} \tag{18}$$

The ADC implementation in the VMM has 10-bits of resolution and is based on a R2R ladder DAC with an amplifier buffer capable of driving the capacitive load of all the column comparators. A schematic of the adjustable TIA, ADC ramp/clock control sub-block is shown in Fig. 89 below

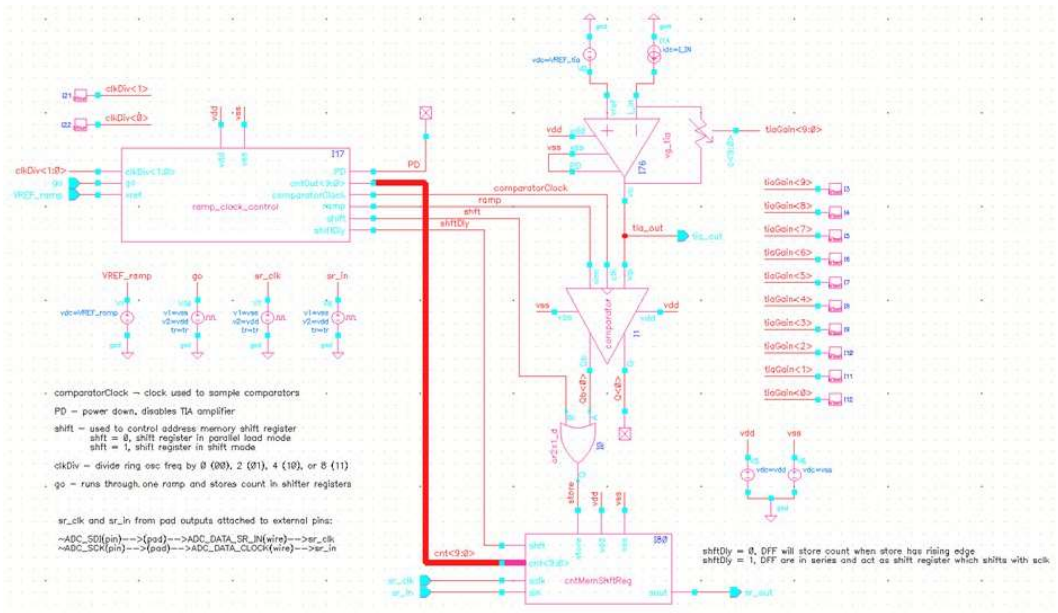
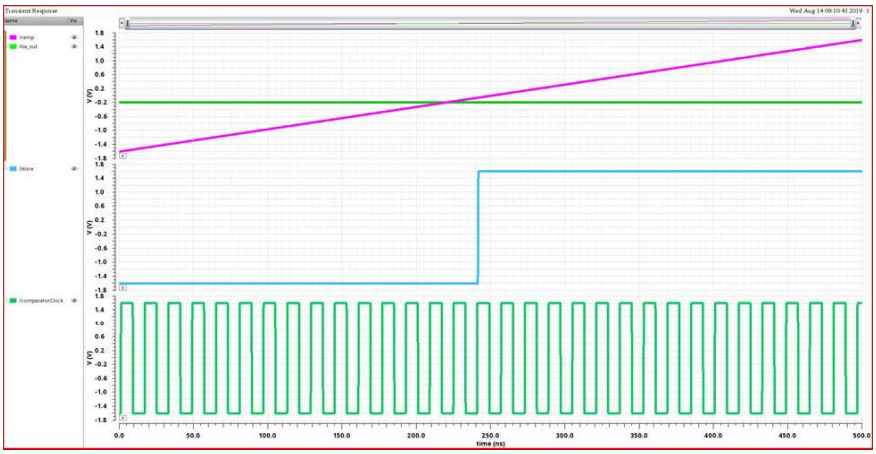


Fig. 89. Adjustable TIA, ADC, and ramp/clock control sub-block.

The comparator output is plotted for two current inputs in Fig. 90a and Fig 90b below. TIA input DC current level found in the table above. The comparator used in the design is a clocked comparator. As such, when the ramp input (green curve) reaches the TIA output voltage, the comparator output triggers on the next clock cycle. When the comparator output shifts high, the number of clock periods registered by the counter is recorded as a digital representation of the VMM calculation for each column. Fig. 91 is the schematic of the clock generator circuit. Fig. 92 is the schematic of the ramp generator circuit. And Fig. 93 is the schematic of the ramp/clock control sub-block.

(a)



(b)

Fig. 90. Comparator input and output with associated clock signal for input current of (a) 800 uA and (b) 400 uA.

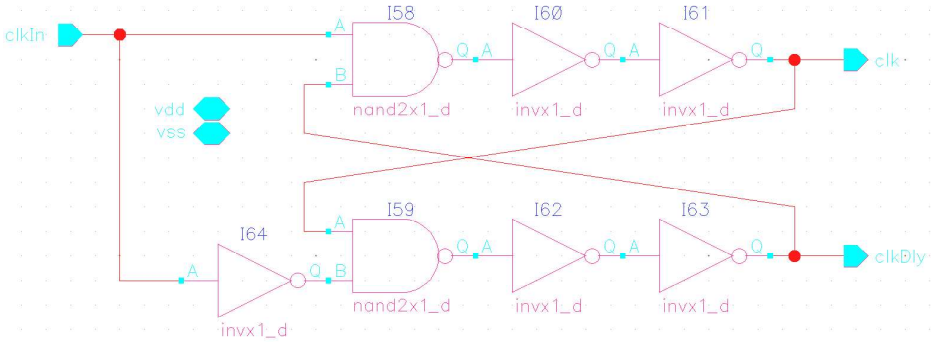


Fig. 91. Clock generator circuit.

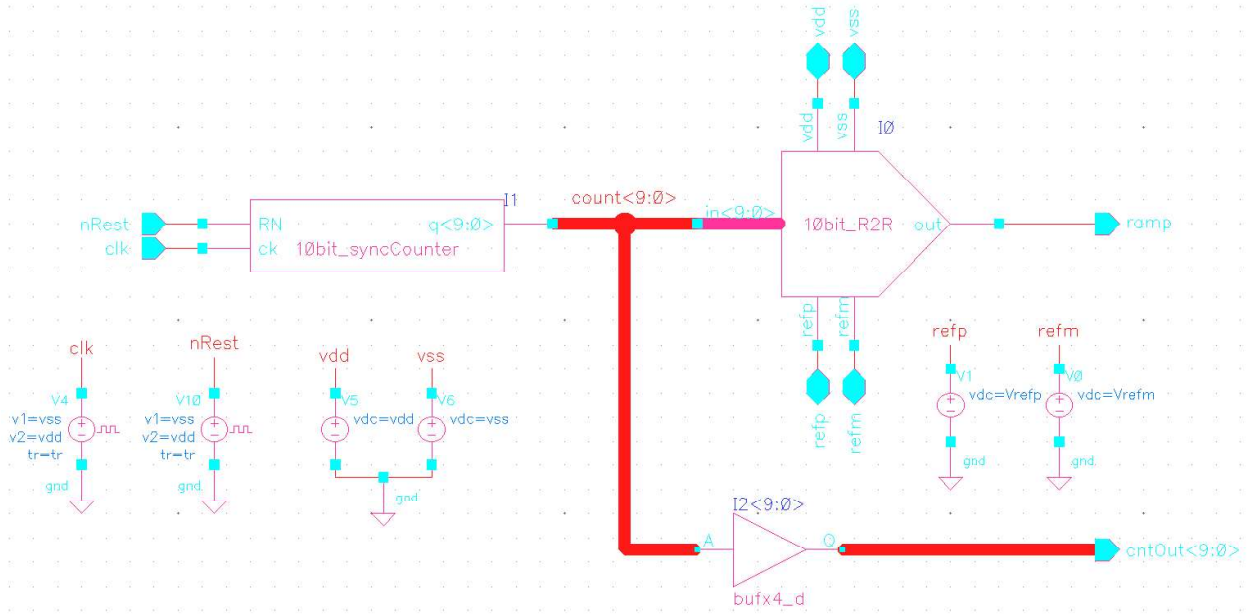


Fig. 92. Ramp generator circuit.

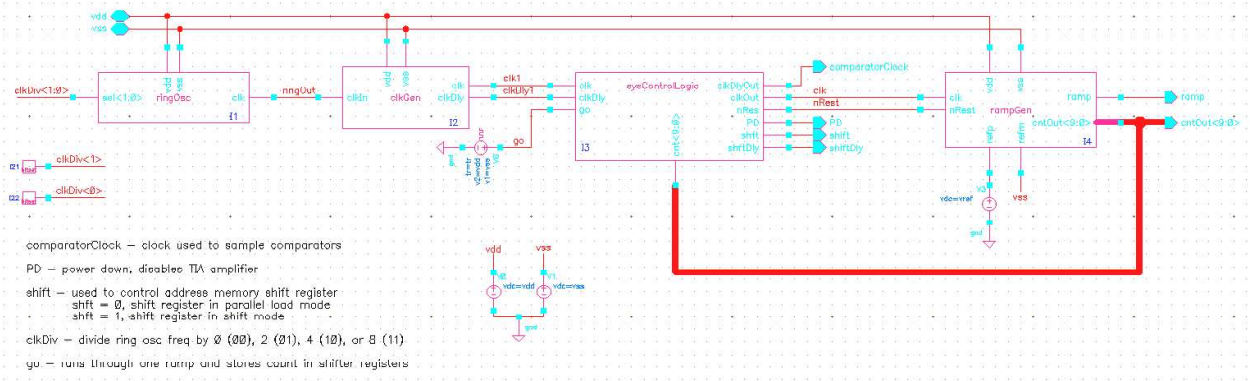


Fig. 93. Ramp/Clock Control sub-block

Write Mode - setting the RRAM resistance levels is done by setting modes for the array column and row drivers, setting a pulse width, and then finally triggering the pulser to turn of the drivers for duration of the pulse. The row and column drivers form an h bridge for devices at each crosspoint in the array. The drive voltages are provided from external pins and there is no active buffering or regulation of these voltages. To reduce any RC effects after a pulse, the shared row and column bitlines are forced to the same VMID reference after each pulse. This also ensures those lines start from the same value for each programming cycle reducing any previous state dependence in the writes.

Pulse Generator (Pulser) - There is one pulse generator whose pulse output is buffered to both the row and column drivers. The schematic of the pulse generator sub-block is shown in Fig. 94.

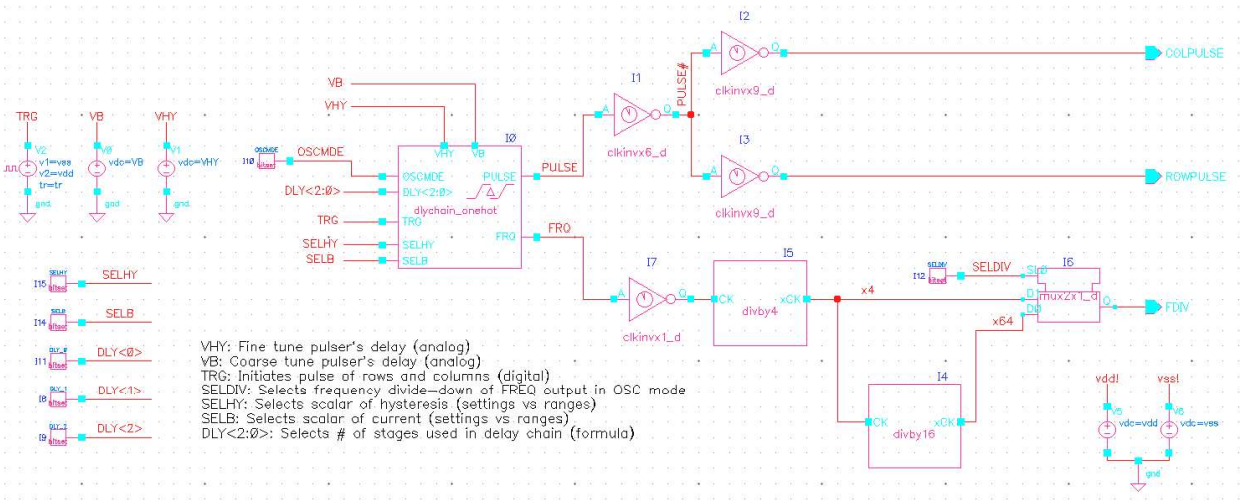


Fig. 94. Pulse generator sub-block

The pulsers have wide range (< 10ns to > 100 us) and require 2 analog voltages (VB and VHY). VB is a coarse grain adjustment while VHY is a fine-grained adjustment. Internally, the pulser is a current-starved delay line with current-adjusted hysteresis. Mapping control voltages to pulse widths is done by placing the pulser in a special test mode where it acts like an oscillator. Measuring its frequency output in this mode gives a direct measurement of the pulse width generated.

Fig 95 plots the Cadence simulation results for parametric steps in VB. As shown in Table V, VB is stepped between -1V and 1 V. Fig 96 plots the Cadence simulation results for parametric steps in VHY. As shown in Table VI, VB is stepped between -1V and 1 V.

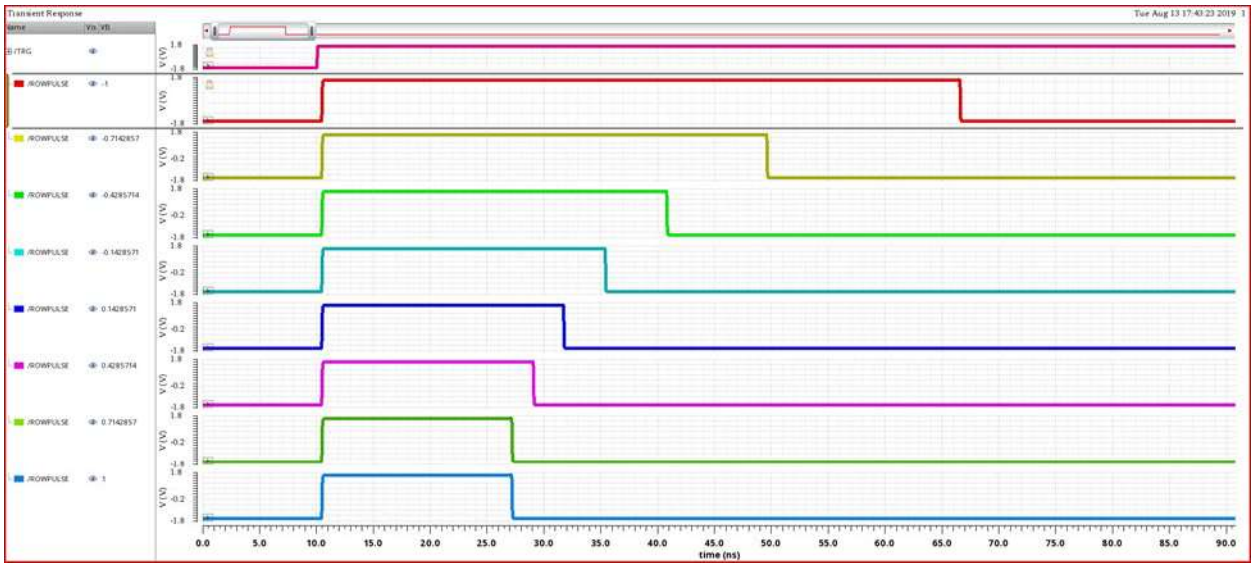


Fig. 95. Pulse Circuit simulation: VB parametric steps (VHY = 0V)

Table III
Steps in Pulse Coarse Grain, (VB).

VB(V)	ROWPULSE Width (ns)
-1.00	56.158
-0.714	39.160
-0.429	30.354
-0.143	24.941
0.143	21.275
0.429	18.619
0.714	16.728
1.00	16.722

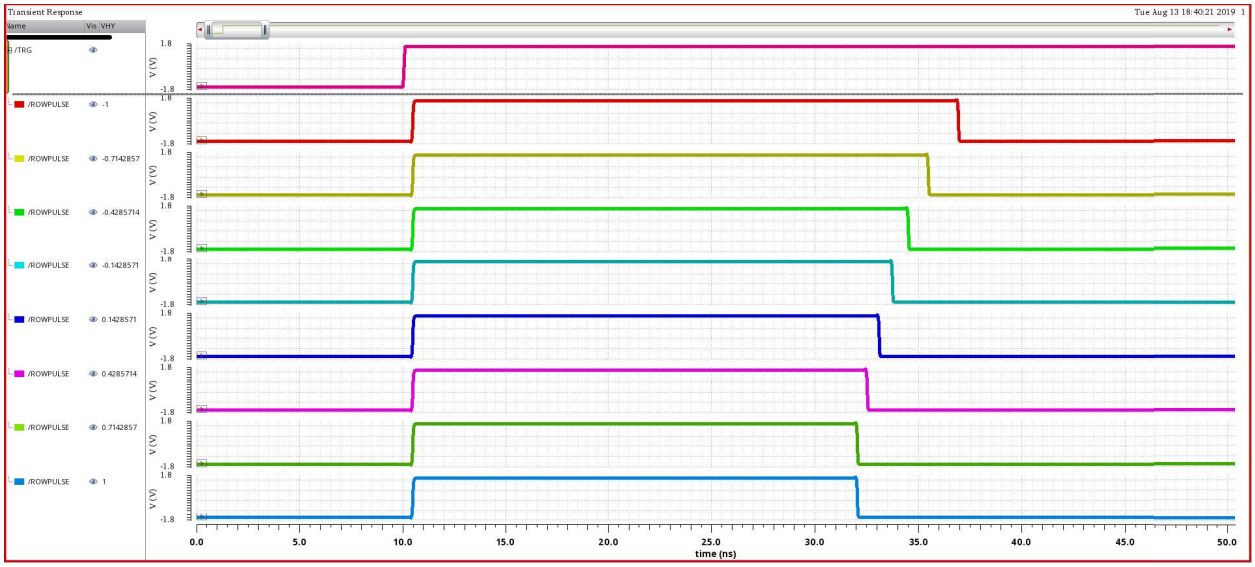


Fig. 96. Pulse Circuit simulation: VHY parametric steps (VB = 0V)

Table IV
Steps in Pulse Fine Grain, (VHY).

VHY(V)	ROWPULSE Width (ns)
-1.00	26.485
-0.714	25.010
-0.429	24.057
-0.143	23.278
0.143	22.637
0.429	22.060
0.714	21.579
1.00	21.579

Operation Modes

There are 8 row drivers and 8 column drivers which can be put into different modes. The main modes are reading, writing, a high-impedance (Z) state, or a half-select state. The high Z modes are essential for mitigating the impact of sneak paths in the RRAM cross-bar array. The rows and columns states are controlled by mode bits set in digital control registers. The 01 register controls the read/write state of the rows and columns, and the 10 register controls which mode the row or column driver is in.

Read mode - When a row is set to read, each *Vrowi* is driven onto the row and the dot product of the conductances and *Vrowis* results in a scaled voltage at the output of the TIA in each column. A schematic of the read mode configuration is shown in Fig. 97.

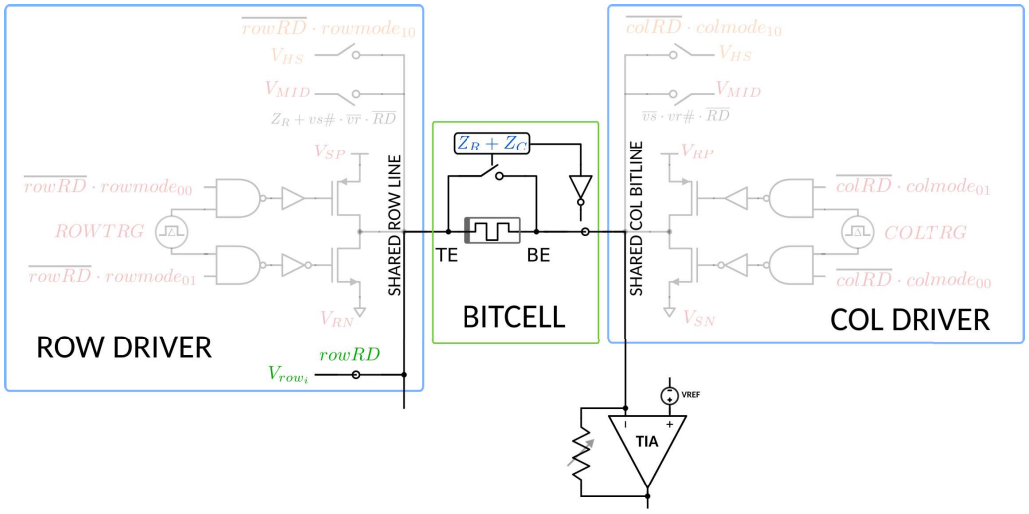


Fig. 97. Read mode configuration

Write mode - the write operation on an RRAM consists of 2 distinct configurations: SET and RESET. In the SET mode the top (anode) electrode of the RRAM device is driven high by the row trigger pulser while the bottom (cathode) electrode is driven low by the column trigger pulser A schematic of the write mode SET configuration is shown in Fig. 98.

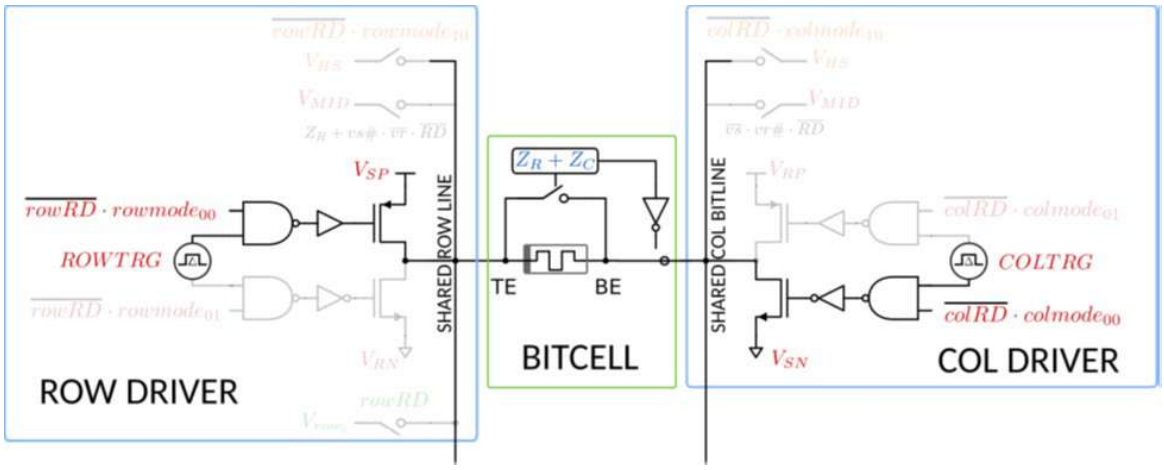


Fig. 98. Write mode SET configuration

In the RESET mode the top (anode) electrode of the RRAM device is driven low by the row trigger pulser while the bottom (cathode) electrode is driven high by the column trigger pulser A schematic of the write mode RESET configuration is shown in Fig. 99.

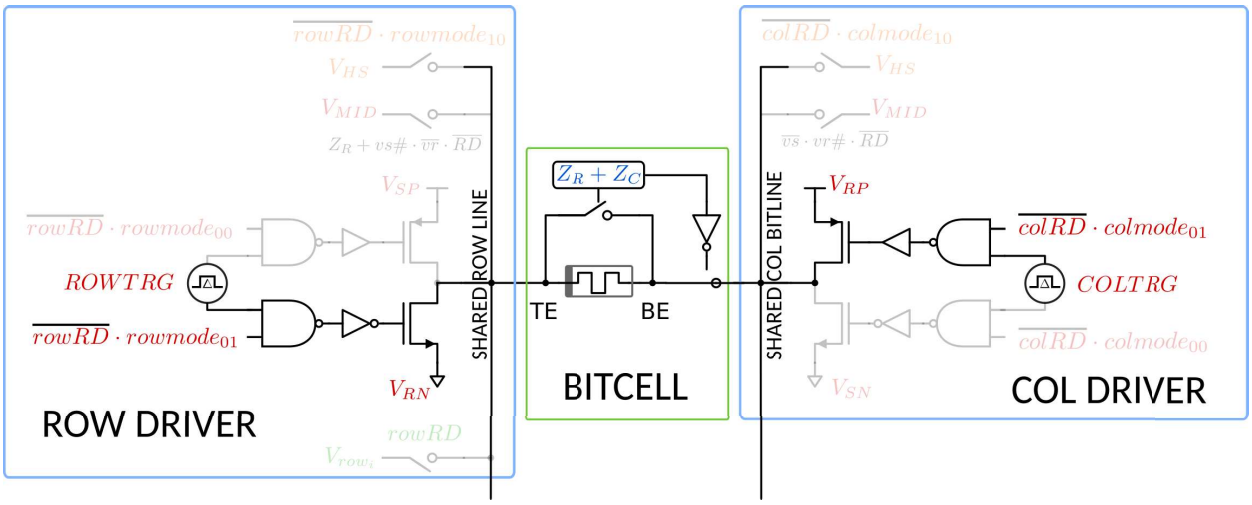


Fig. 99. Write mode RESET configuration

High Impedance Modes - Every row and column can be put into a “disconnected” mode (or high-impedance mode) where its access transistor is open so that it does not participate in any of the read or write actions impressed across its shared bitline. Simultaneously, a shunt transistor is placed in parallel with any RRAM devices placed in high-z mode. This ensures that any transient (dis)charging has a low impedance to go thru instead of the RRM device if the rowline is being driven. A schematic of the High-Z configuration is shown in Fig. 100.

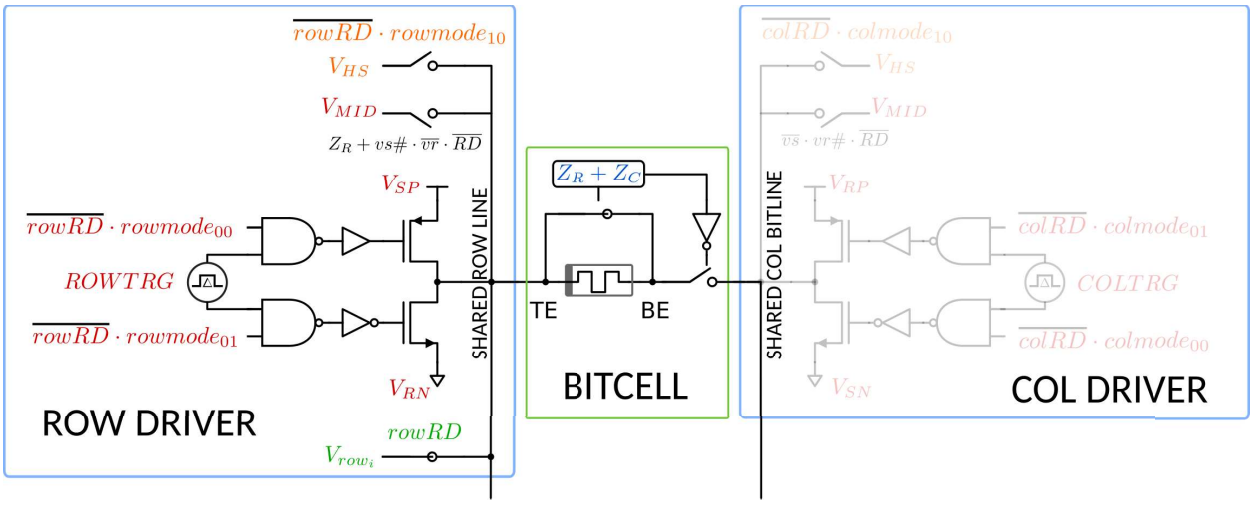


Fig. 100. High-Z configuration

Fig. 101 is a picture of the VMM chip layout. In addition of the 8x8 resistor array, , the chip also included transistors and MOS chain structures to support TID characterization of the CMOS. The full chip and CMOS TID characterization is currently being performed at ASU.

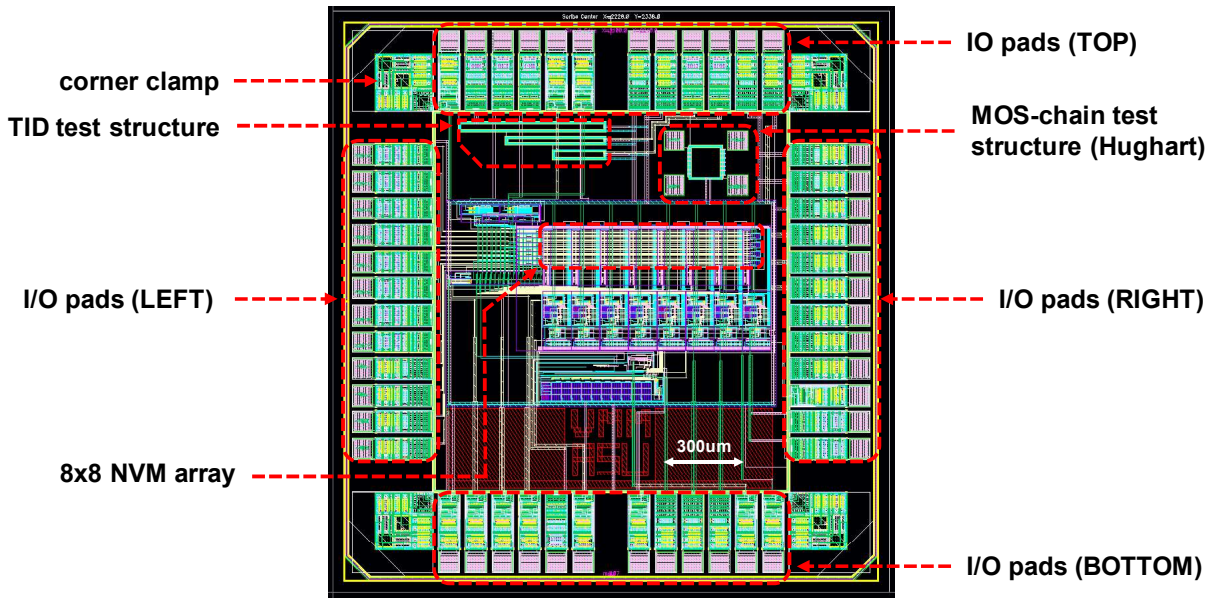


Fig. 101. VMM_8x8 Layout

Eight wafers of the CMOS chips were fabricated by ON-semiconductor in their 0.35 mm process in Pocatello ID. Fig. 102 is a photomicrograph of the final die.

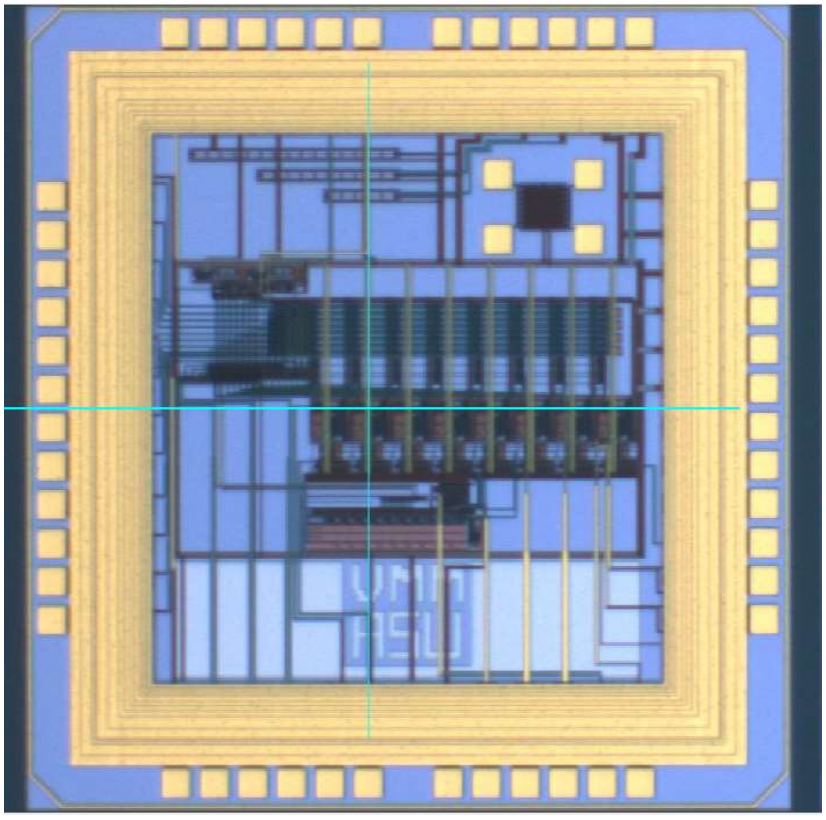


Fig. 102. Die photomicrograph of fabricated CMOS-RRAM hybrid Vector-Matrix Multiply (VMM) Chip

Task 2.3 Simulation of radiation effects in SRAM array with peripheral neuron circuits

Simulate TID effects in non-spiking XNOR architecture based on SRAM arrays.

See Task 3.3.

Task 3. Architecture-level Modeling Radiation Effects in Neuromorphic Systems

Task 3.1 Implement neuro-inspired algorithms on neuromorphic hardware platforms

Develop the software codes and test bench for evaluating deep feed-forward multi-layer perceptron network.

VMM ASIC Test Bench

We use combination of an FPGA and a PCB to connect and communicate with the ASIC. The system is illustrated in Fig. 103 as a block diagram and in Fig. 104 is the physical set up. The PCB is populated with DACs which will allow us to supply analog voltages to the ASIC despite the digital nature of the FPGA. The PCB also includes a series of OpAmps which will be used to shift the voltage level to be compatible with the ASIC. The ASIC is equipped with an ADC that allows us to read back the analog data from the VMM operation onto the FPGA.

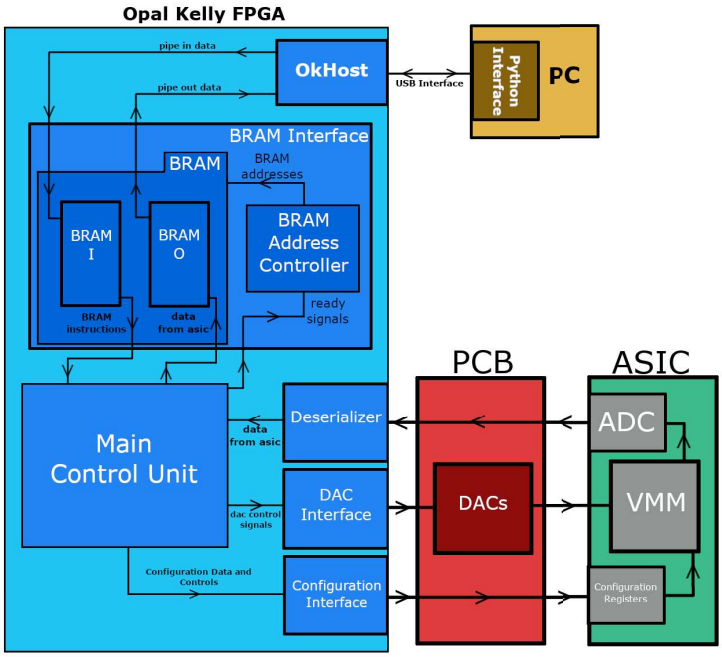


Fig. 103. Block diagram of the connections between the user (at the PC) and the ASIC

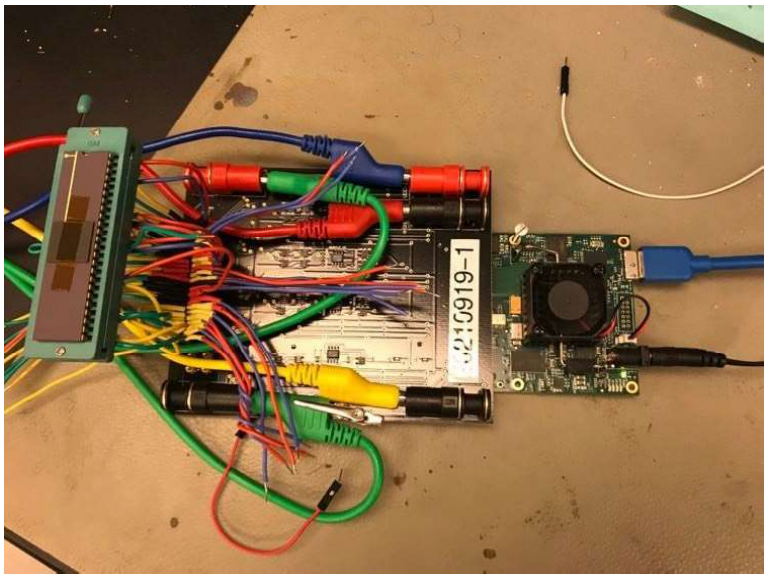


Fig. 104. Physical set up. From left to right: the ASIC on the turquoise ZIF connector with colorful wires leading to and from the PCB, the black PCB, and the smaller, green PCB that houses the Opal Kelly FPGA. The blue cable out of the FPGA connects via USB to a PC.

To evaluate the VMM ASIC we created a series of Verilog modules to program onto the FPGA. These modules give us control of signals going into the ASIC. Data can be sent into the FPGA via a Python interface. We created several Python functions that encode instructions for the FPGA to generate a way for a user to interact with it. These Python functions allow us to perform tasks such as write to the DACs, change settings on the configuration registers, and other FPGA actions easily. These functions have been updated and optimized throughout the scope of our project. Using these functions, we needed to run tests on the pulser to determine its characteristics across different conditions. The finished characterization of the pulser will enable us to set the RRAM devices to the values that we desire once they are populated onto the chip.

Following the creation of the Python interface, we began characterizing elements on the PCB. These characterizations helped us identify and fix an issue that could have become more significant when running tests. Additionally, validation of the TIA (used for converting the output current into a readable voltage) and ADC have been completed and we have successfully run a dot product operation. It should be noted that software-driven machine learning is inefficient because of the high computing cost of a VMM operation when done with software. Using the VMM hardware approach will significantly improve the rate at which we can achieve machine learning because it enables easy performance of VMM operations with simple functions.

Test Results

Verilog modules were created to program onto the FPGA, and a Python interface through which data can be sent into the FPGA. Included in this interface are several Python functions we created that encode instructions for the FPGA to generate a way to interact with it. These

functions have been created with the idea that they are easy to implement at a higher level. Using these functions we have been able to characterize the pulser, and because of these we have the means to be able to set the RRAM values when they are populated onto the chip. The RRAM must be set using an iterative write-verify process so we will use the pulser to iteratively write until converging on a desired value. The verification portion of the write-verify process will require us to read out the data from the column where the RRAM is located while all the other devices are cut off from the circuit. We decided it would be best to handle the entire process on the FPGA and ASIC to save time rather than going through the CPU. Piping the data back to the CPU for every readout in an iterative write process would take a significant amount of time. After deciding to implement this iterative solver within the ASIC we created FPGA logic to read the data without instruction from the PC and decide whether to initiate positive or negative pulses to the RRAM device until the target value has been achieved. It should be noted that we have not yet had the opportunity to test this iterative write-verify scheme due to the absence of the RRAM as mentioned earlier.

Additionally, from a software perspective, test benches have been written for testing the pulser, ADC, and TIA. Research has been conducted to learn about coding a perceptron algorithm to do machine learning as soon as the VMM is verified (**EM indicator Q**). Much like Bayat's article on "Advancing Memristive Analog Neuromorphic Networks: Increasing Complexity, and Coping with Imperfect Hardware Components," we use a multilayer perceptron for the various binary inputs and RRAM arrays. Ultimately, we will need to utilize the RRAM devices in a way that uses hardware neurons to further improve the power and chip area efficiencies. Furthermore, it has been confirmed that code is working properly by observing a square wave output on pin VB and VHY while performing several tests with an oscilloscope. The generation of the square wave is the indicator of the output voltage which was to be expected. For example, the initial code for some of the testbenches made it difficult for the DACs to be characterized, so the Python code has been edited for better user interaction with testing each bit setting. Examples of these modifications in the code are shown below in Figs. 105 and 106.

```
#Settings that can change
Vin = [3.27, 3.27, 3.27, 3.27, 3.27, 3.27, 3.27, 3.27]
TIA_VREF = 0
data_string = '1111111111111111'
EN_EXTIA = 3.27
################################################################################
```

Fig. 105. Setting that can be changed for the TIA with EN_EXTIA being an output voltage of 3.27. Settings are all grouped together for user ease.

```
# For Write Reg Sctrl to 11 which = 3,
instr = device.encode_config_instr(data_string, 3)
device.append_byte_array(instr)

#Activate pulse trig activates a signal into the ASIC to send a pulse
device.activate_pulse_TRG()

device.pipe_data_in()
device.start_program()

device.xem.Close()
```

Fig. 106.General settings for information to be put on a buffer, and piped into the BRAM of the FPGA. In this example, for the pulser. Improved functions to simplify top-level code.

In the third year of the program began characterizing the PCB, DACs, and the OpAmps to identify and fix issues with the VMM ASIC. We originally saw a significant difference between the PC-requested voltage output of the DACs and OpAmps, and the actual output voltage seen on the output pins. This error increased as the requested voltage increased, and is documented in Fig. 107, which shows the difference between the requested, set value and the actual output voltage seen for the DACs and OpAmps as that offset increases with the requested value from the PC. From this comparison, it is evident that the rate of change of offset is the same between the DACs and the OpAmps, but the DACs had no offset at a 0V requested value, whereas the OpAmps had an offset that was shifted down from the DAC curve. We were able to identify that this problem was born from the voltage rail being set in the code at 3.3V. When we decreased this slightly to 3.27V, such that we did not limit the use of the device, we were able to completely eliminate this issue.

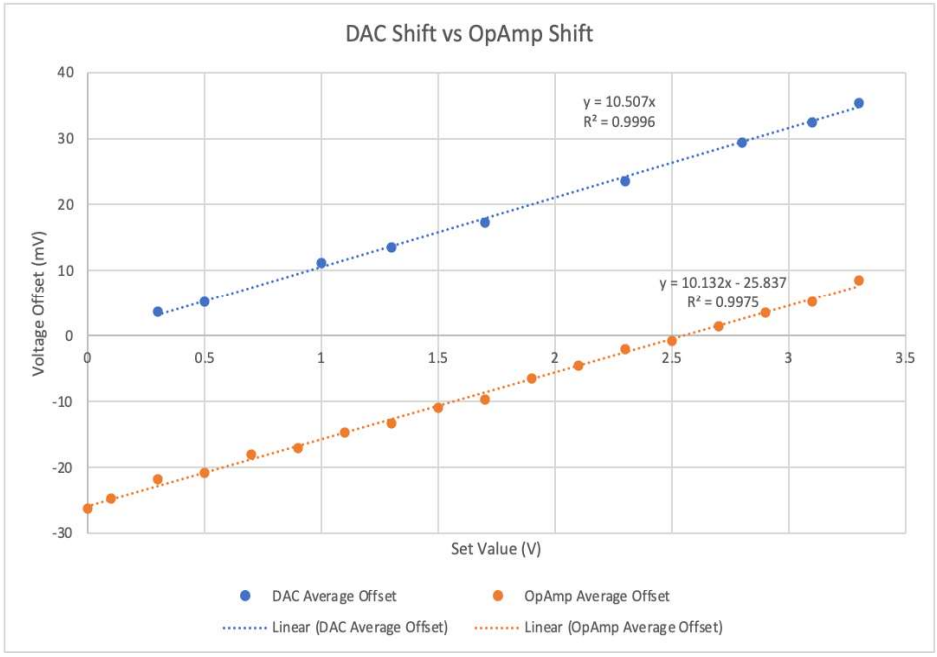


Fig. 107. The difference between the requested, set value and the actual output voltage seen for the DACs and OpAmps as that offset increases with the requested value.

We have validated the functionality of the Pulser was also validated. Data is shown in Fig. 108 to illustrate the correct and expected functionality of the pulser.

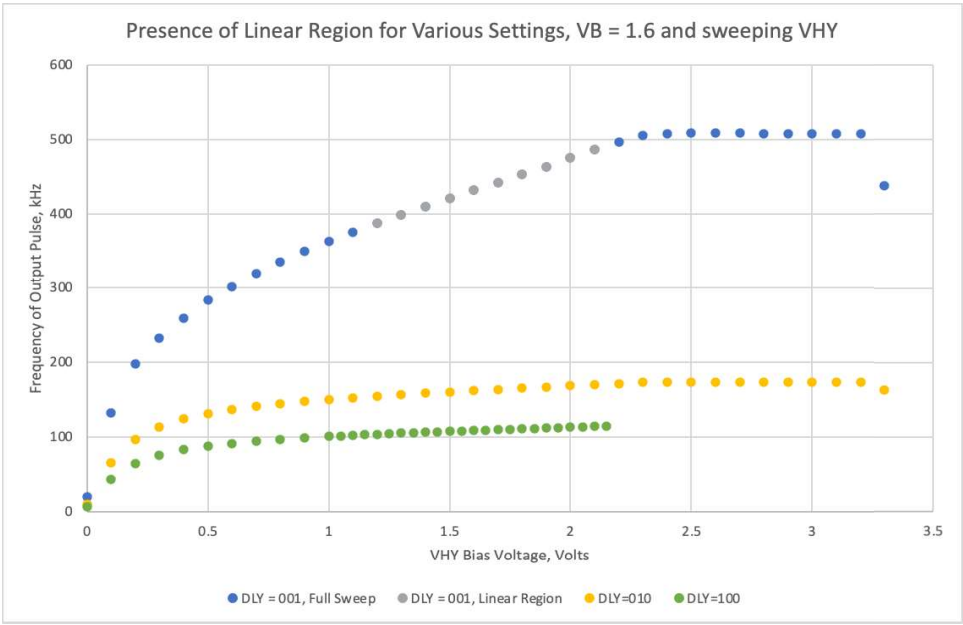


Fig. 108. Resulting frequency of output pulse at various delay settings across varying bias voltages. Used to illustrate the presence of a linearly predictable range of output frequencies that will be used to

accurately set the RRAM values. This shows the Pulser is functioning properly and is not a full characterization of the pulser.

The functionality of the TIA and ADC were verified. If the ADC is working properly, the output of the ADC will be the digital value of the TIA output voltage. The data from the ADC is deserialized by the FPGA and then piped back to the PC. The PC then decodes the data into an analog form the user can understand. We were able to verify that we can receive data from the ASIC and that we can change that data by changing input voltages and gain settings. Below (Fig. 109) is a side by side comparison of the TIA output voltage vs a sweep in input voltage and the data received from the ADC vs the same input voltage sweep.

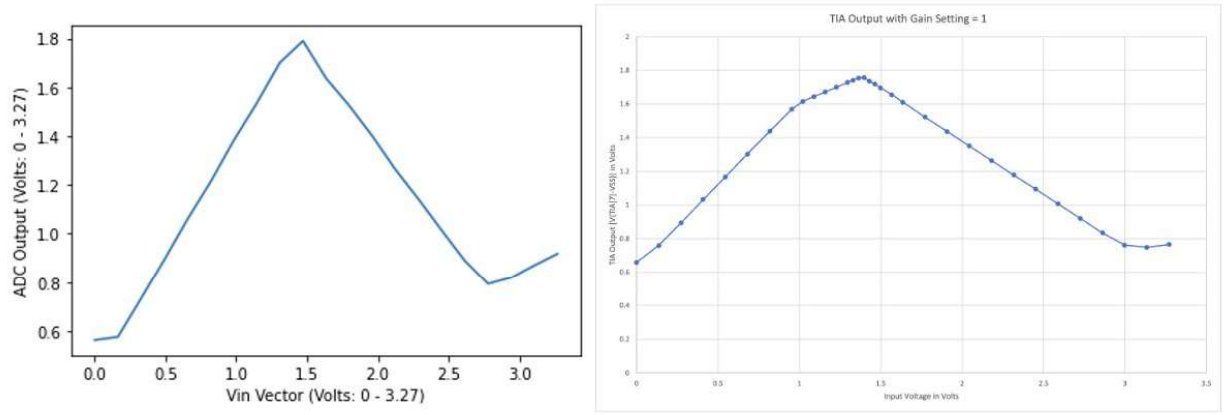


Fig. 109. Side by side comparison of the ADC output and the TIA output

The linear region with the negative slope between approximately 1.5 and 3 Volts on the input will be the region of operation for the output of the TIA. After these tests, a single test column of the TIA has also been verified which will allow us to run simple tests to verify functionality without the population of the RRAM. We have also written code that will allow us to run a dot product operation as well as a simple VMM. The dot product code has been tested and we are able to successfully perform a dot product operation using our PCB and ASIC setup. The next step for the ASIC is to run the VMM operation and apply it to a machine learning problem. This will require RRAM to be populated onto the ASIC. This was not completed after Year 3 of the program.

Modelling Effect of Linearity on Pattern Recognition Accuracy

The impact of linearity and weight update behavior on system response to different pulse schemes is studied by modelling these devices in an analog neural training accelerator. An artificial neural network was simulated using the device properties of the Ag-$Ge_{30}Se_{70}$ CBRAM. The CrossSim simulator was used to perform the supervised learning. A three-layer neural networks depicted in Fig. 110 were trained using a backpropagation algorithm. This algorithm is a computationally intensive algorithm that uses the two important kernels: vector matrix multiply (VMM) and outer product update. For this work, we have used two datasets: a small image version (8 X 8 pixels) of handwritten digits from the 'Optical Recognition Handwritten Digits'

dataset and a large image version (28 X 28 pixels) of handwritten digits from the MNIST dataset. A two-layer neural network consisting of 64 input, 36 hidden and 10 output nodes are used for small image dataset. The larger MNIST dataset was trained on a network of dimensions 784x300x10. A crossbar array is simulated using the analog synaptic properties. The crossbar a part of the neural core performs the required kernel operations. To perform the vector matrix multiplications, the conductance states of the synaptic devices are programmed by applying input voltages or input pulse lengths to each row and the corresponding output vector is read in the form of current. Parallel read, multiplication and summation operations are performed in this single step. This helps in reducing the total energy of vector-matrix multiply (VMM) and highlights the main advantage of using analog resistive memories for these operations.

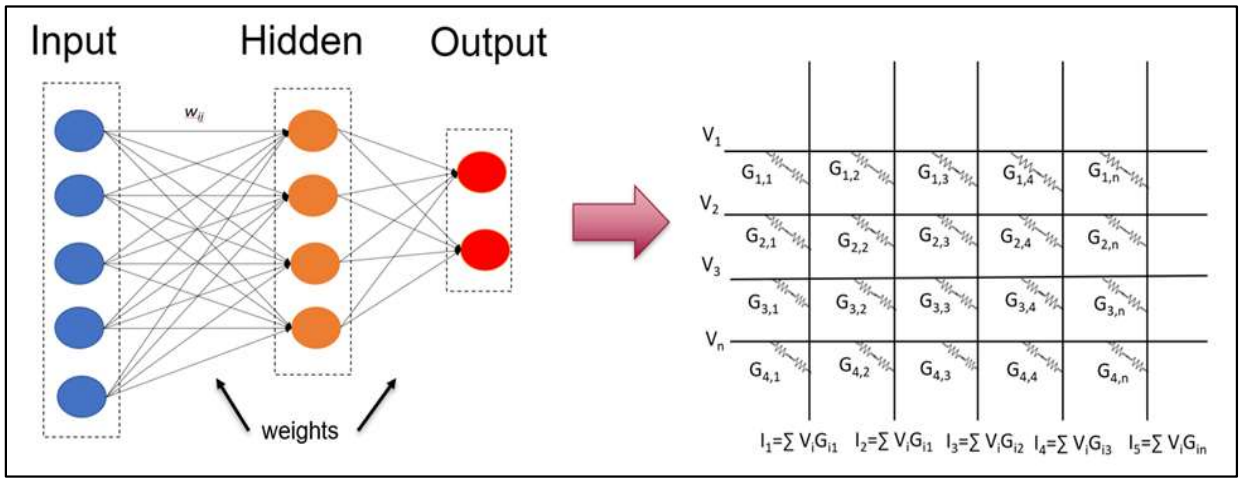


Fig. 110. Mapping the neural network layers to crossbar array.

Table V
Datasets used for System Modeling

Data set	**#Training Examples**	**#Test Examples**	**Network size**
UCI Small images	3823	1797	64X36X10
MNIST large images	60000	10000	784X300X10

To simulate the response of the Ag-$Ge_{30}Se_{70}$ CBRAM devices in the neural network, a lookup table is created containing $\Delta G/G$ response of the devices to multiple potentiation and depression cycles at a given pulse widths using data from Figs 111 and 112. For small image dataset, we observe the training accuracy of 85% for constant amplitude pulsing as depicted in Fig. 111(a) while the training accuracy for the same dataset with increasing pulse amplitude was found to be 92% as seen in Figure 111(b). For large digits the accuracy was seen to be

improved from 79% to 87% for increasing pulse amplitude as seen in Figure 112a & 112b. The ideal training accuracy using a double-precision CPU or GPU is 98%.

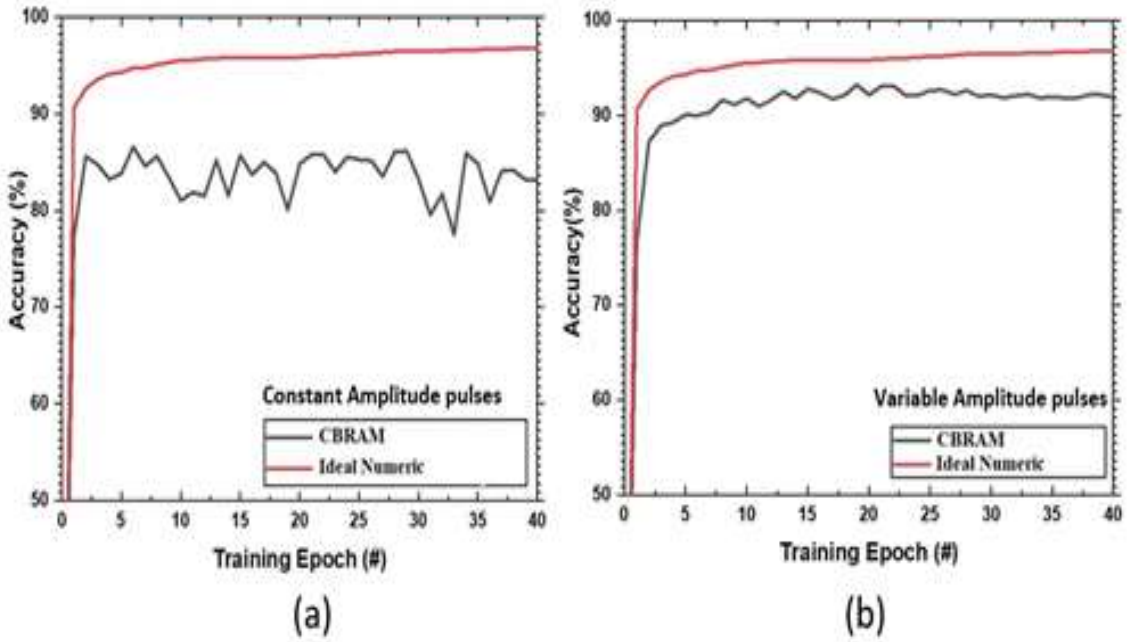


Fig. 111. Classification accuracy for small image dataset.

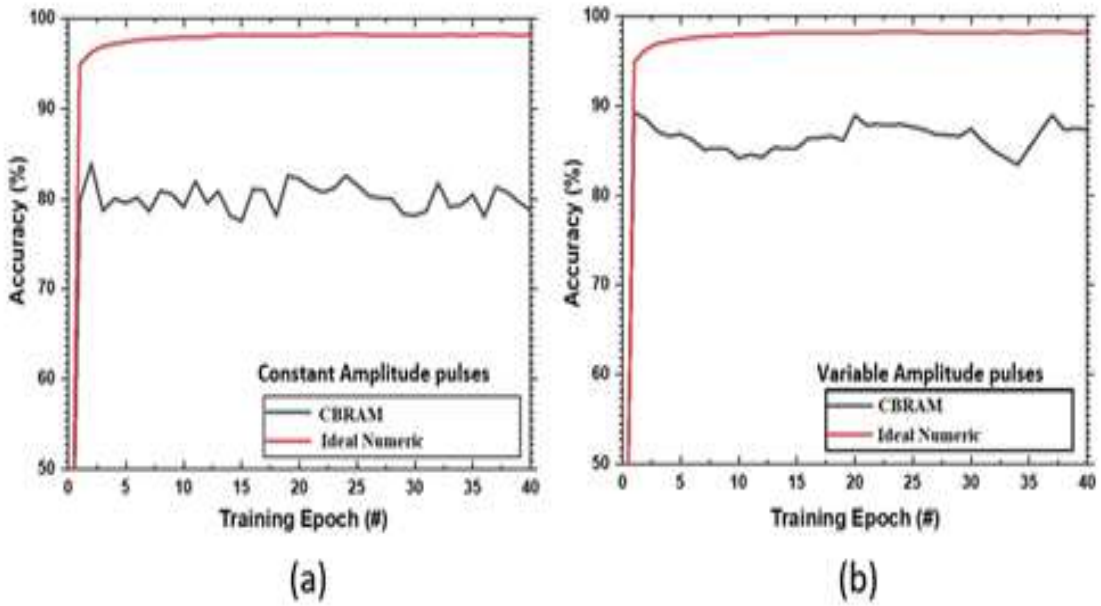


Fig. 112. Classification accuracy for large image MNIST dataset.

Circuit level performance of Ag-$Ge_{30}Se_{70}$ CBRAM in a crossbar array architecture -
To benchmark the circuit level performance of Ag-$Ge_{30}Se_{70}$ CBRAM device we use NeuroSim, a neuro-inspired architecture at circuit level, to estimate the area, latency, dynamic energy of a neuromorphic hardware accelerator. NeuroSim utilizes the circuit/device model parameters on a small but realistic neuro-inspired architecture, enabling fast early-stage design exploration. The overview of neuromorphic hardware design utilized by NeuroSim module is depicted in Fig. 113.

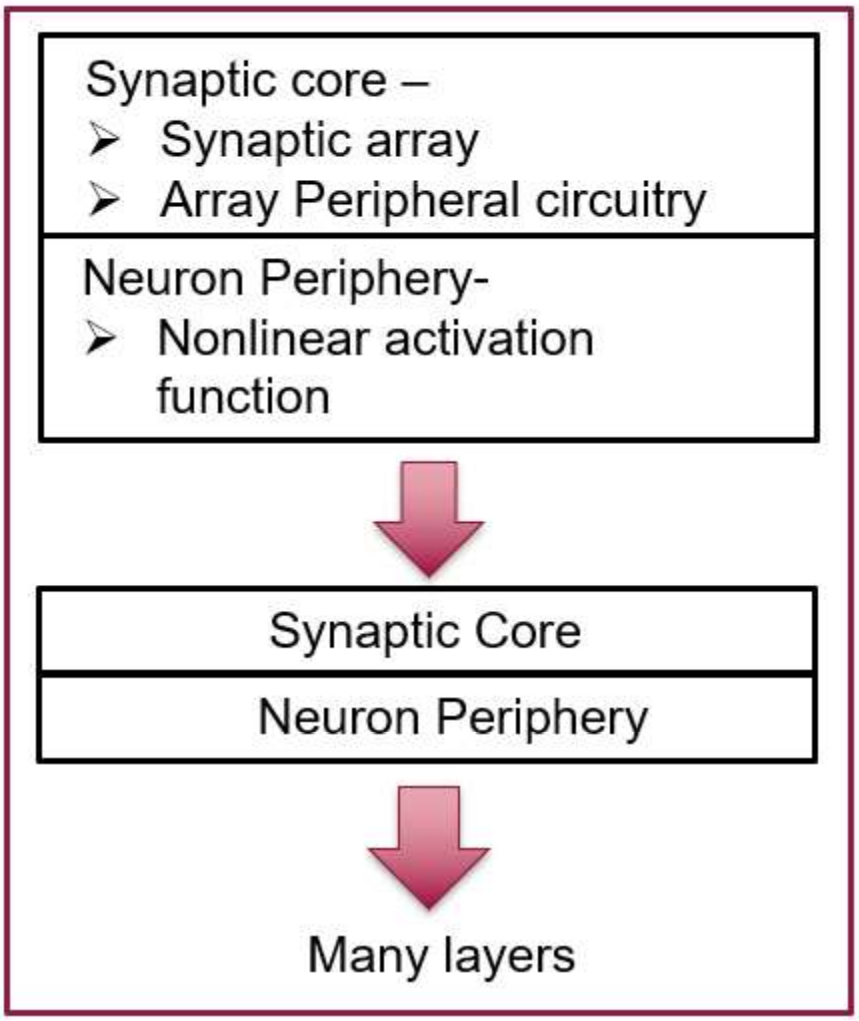


Fig. 113. An overview high level neuromorphic accelerator design used in NeuroSim to implement neural network.

To estimate the circuit level performance of Ag-$Ge_{30}Se_{70}$ CBRAM, a 2-layer MLP network with 400 input neurons, 100 hidden neurons and 10 output neurons is mapped to the analog eNVM crossbar array architecture. Each synaptic layer consists of synaptic core and neuron periphery. The synaptic core consists of a crossbar array architecture as depicted in Fig. 114.

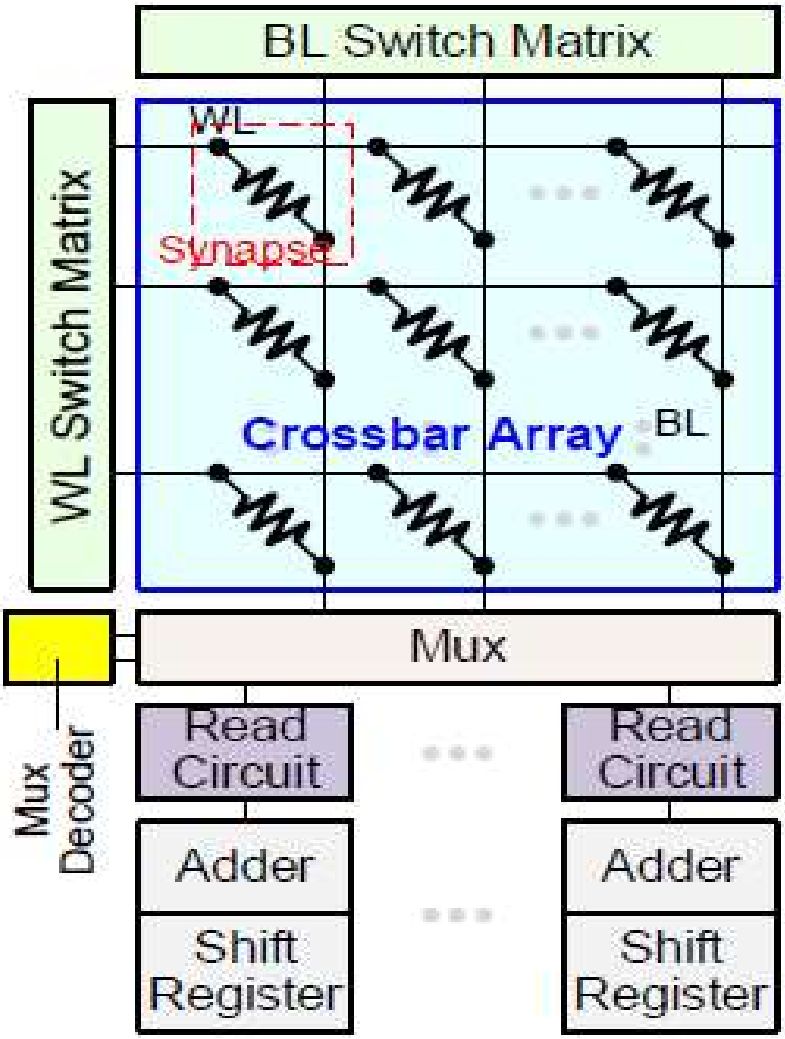


Fig. 114. Circuit block diagram of analog eNVM synaptic core. © 2017 IEEE

In this type of array structure, each CBRAM device is located at the cross point of a word line (WL) and a bit line (BL). This is the most compact and simplest form of array structure. This layout can achieve the highest integration density of $4F^2$/cell. The input vector can be coded in read voltage signals and the weighted sum operation (i.e., MVM) can be performed in a parallel fashion. A word line (WL) switch matrix can be used to switch multiple rows or columns at a time thus enabling parallel read voltage input for the weighted sum operation or selecting a rows or columns to perform weight update scheme that require different voltage biases. Adder and shift registers are used to perform the shift and the addition on all the bits cycles to obtain the final weighted results. To convert the analog weighted sum currents to digital outputs, read circuitry is used. The read circuit uses the principle of integrate-and-fire neuron model. The neuron

peripheral circuitry performs the non-linear activation function on the digital outputs and is also responsible for the communication between different synaptic cores.
A list of the simulation parameters is provided in Table VI. The parameters are extracted from the synaptic response of the Ag-$Ge_{30}Se_{70}$ CBRAM to constant amplitude pulses.

Table VI
Simulation circuit and device parameters -Constant Amplitude Pulses

Parameters	**Values**
Technology node (F)	32 nm
Clock frequency	2 GHz
Assigned height of cell	2 F
Assigned width of cell	2 F
Crossbar eNVM cell area	4 F^2 (F = tech node)
Max. Conductance	6 E-4
Min. Conductance	3.8 E-5
Write Voltage	0.35V / -0.25V
Read Voltage	30 mV
Write pulse width (P.W.)	100 ns
Read pulse width (P.W.)	100 µs
Non-linearity	6.65 / -2.5

Simulation Results
The simulated results for NeuroSim for constant amplitude pulses are depicted in Table VII. The total area is calculated as the summation of the synaptic core area and the total neuron peripheral circuit area. The latency in a crossbar array is caused by wire parasitics (*R* & *C*) and the peripheral circuitry. The power consumption is mainly due to the static energy consumption in an eNVM synaptic array. The power consumption is directly proportional to the conductance of the cell, write voltage used for conductance increase and decrease, number of applied write pulses and the pulse widths used.

Table VII
Represents the circuit level performance of Ag-Ge30Se70 CBRAM device at 32nm.

Area (μm^2)	Read Energy (J)	Write Energy (J)	Read Latency (s)	Write Latency (s)
1.4028E5	1.0176E-01	5.0520E-02	1.5796E-02	4.5121

One should note that when using variable amplitude scheme that yields highly linear weight update behavior, a read-before-write verify step is necessary to determine the current conductance state. After the conductance state is read, a correct pulse amplitude is applied. This increases the peripheral circuit complexity along with increasing the latency and energy consumption. Fig. 115 depicts the crossbar array circuitry used for non-identical pulse application.

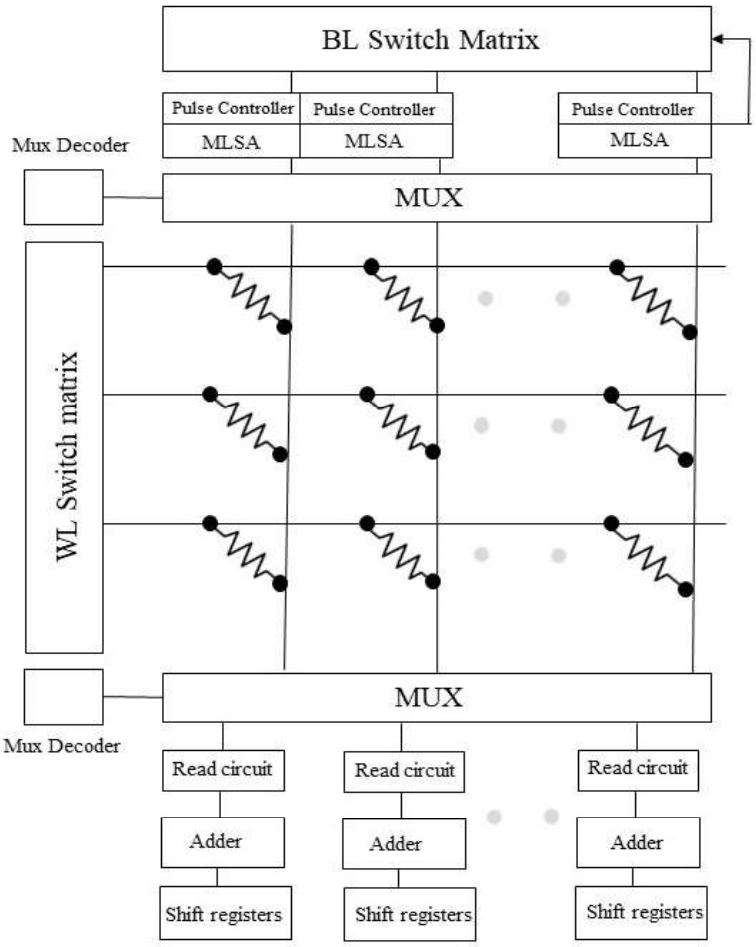
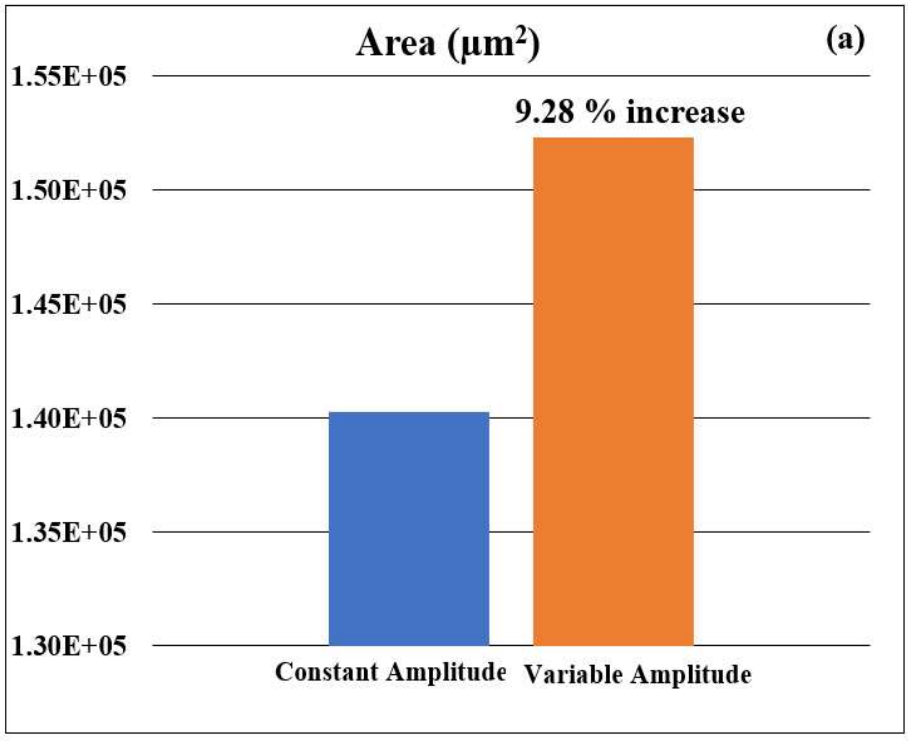


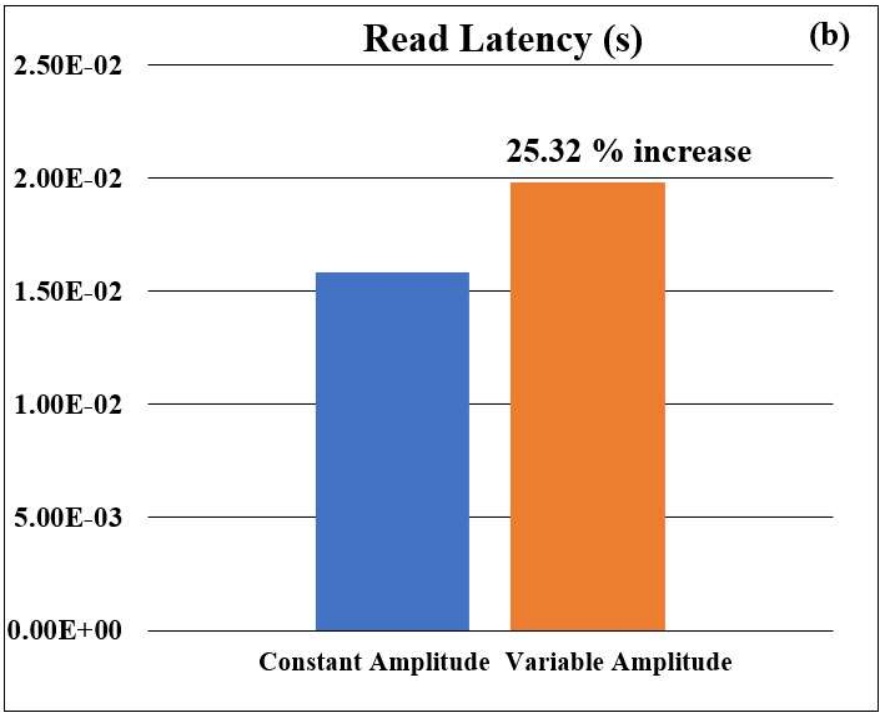
Fig. 115. Block diagram of Ag-Ge30Se70 CBRAM crossbar array with peripheral circuitry used for variable amplitude pulse application.

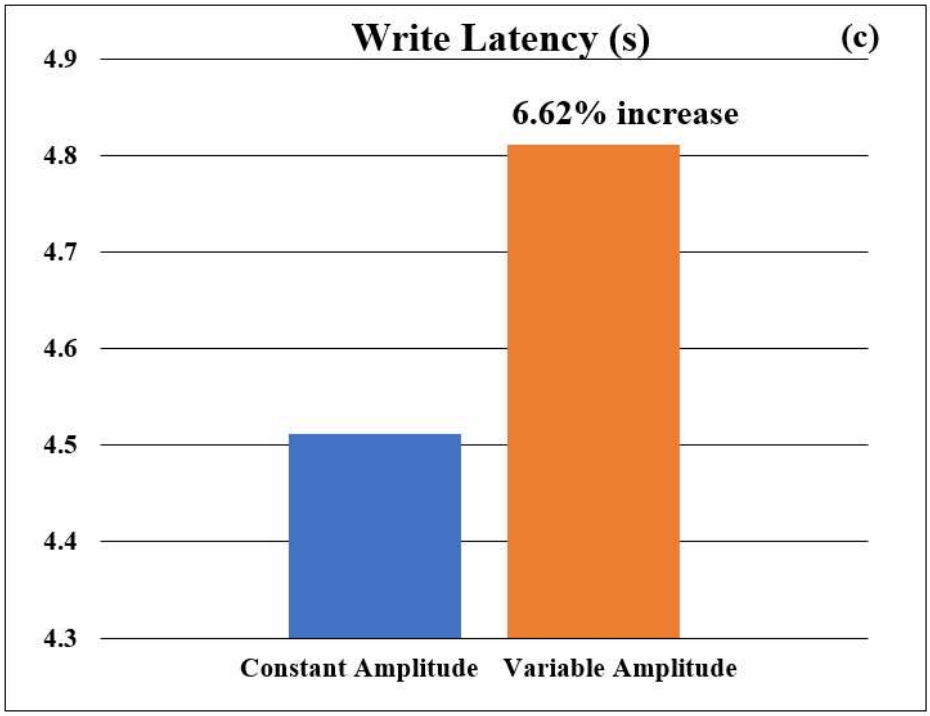
Additional circuitry like pulse controller and multilevel sense amplifier are added to the crossbar array to feed back the current state to choose appropriate pulse amplitude. Circuit simulation was performed using the variable pulse amplitude response of the Ag-$Ge_{30}Se_{70}$ CBRAM devices as depicted in Fig. 47. The simulated results are shown in Table VIII. The contrasts with a constant amplitude pulse scheme are shown in Figs. 116a-e. Comparing the performances, we observe that there is 9.28% area overhead, 6.62 % write latency, and 768% write energy, 426 % read energy penalty is introduced when using variable amplitude pulses as compared to constant amplitude pulses.

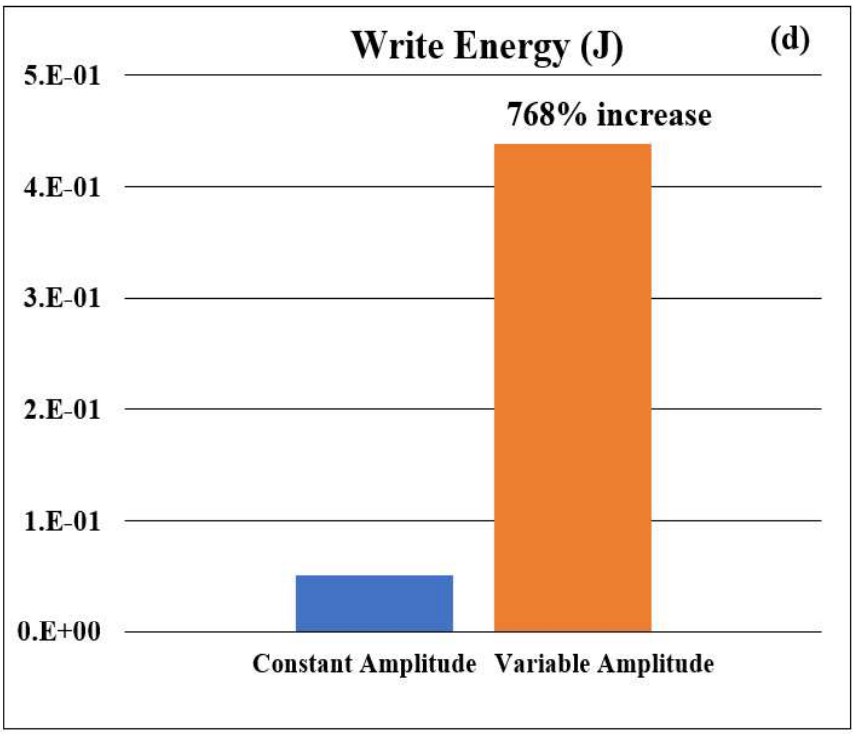
Table VIII.
Represents the circuit level performance of Ag-$Ge_{30}Se_{70}$ CBRAM device at 32nm with Variable pulse amplitude.

Area (μm^2)	Read Energy (J)	Write Energy (J)	Read Latency (s)	Write Latency (s)
1.53297E5	5.2522E-01	4.3830E-01	1.9796E-02	4.8107




Read Latency (s)
(b)
2.50E-02
2.00E-02
1.50E-02
1.00E-02
5.00E-03
0.00E+00
25.32 % increase
Constant Amplitude
Variable Amplitude


Write Latency (s)
(c)
4.9
4.8
4.7
4.6
4.5
4.4
4.3
6.62% increase
Constant Amplitude
Variable Amplitude



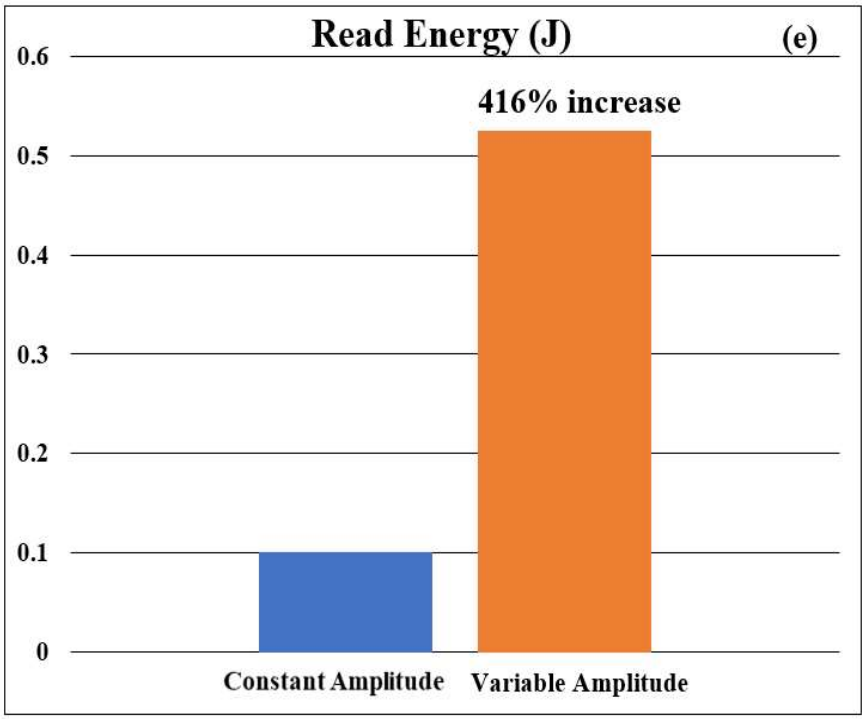


Fig 116. Comparing the circuit level performance of Constant amplitude and Variable amplitude pulses.

The overhead energy and latency problem can be resolved by using combination of resistors in series with RRAM device as suggested in some previous works. K. Moon *et.al* suggest using a voltage divider arrangement in which a fixed resistor is connected in series with the CBRAM device as shown in Fig. 117. However, for potentiation, i.e., increasing the CBRAM conductance (i.e., reducing the resistance), it will not be possible to use a fixed voltage divider to realize an increase in the pulse voltage across R_{CBRAM}. To increase the V_{CBRAM} pulse while R_{CBRAM} switches to a lower resistor a more sophisticated design is required.

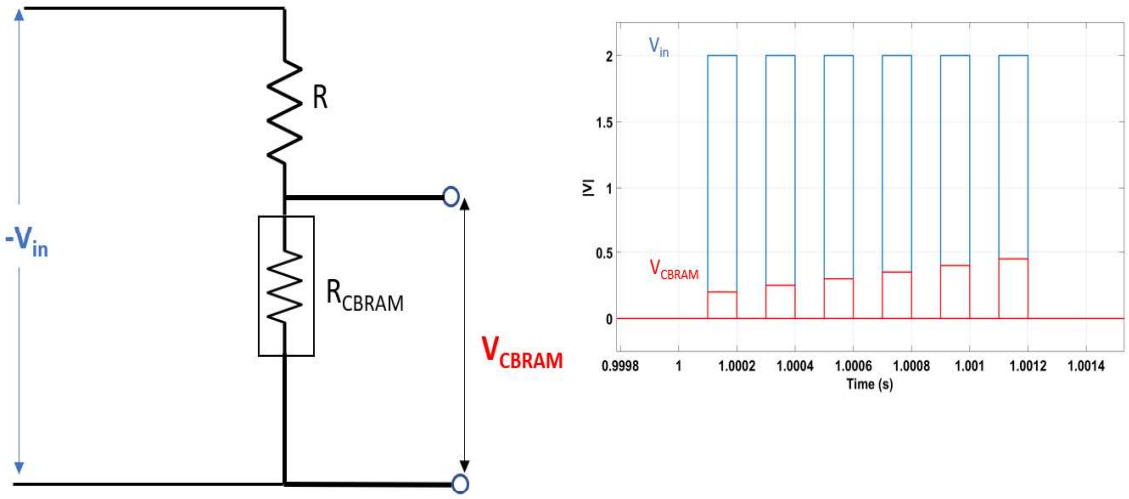


Fig. 117. Two resistor circuit to implement increasing amplitude effect.

One proposed design is shown in Fig. 118. In this design, the constant voltage pulse input, V_{IN}, toggles a custom 4-bit mixed signal counter, CNTR1, which selects eight different bias currents to CBRAM device. When the 4th bit of CNTR1 is set low, a second 4-bit counter, CNTR2, is enabled. The first 3 output bits of CNTR2 combine with the CNTR1 currents to add another four monotonically increasing currents to the sequence (the last three 3-bit outputs of CNTR 2 can be neglected as the CBRAM device will have reached a level close to its minimum value by that step). The 4th bit of CNTR2 is fed back to the enable switch of CNTR1. When the 4th bit on CNTR2 is toggled, CNTR1 is disabled, and the pulse sequence terminated. The currents are set by biasing a select combination of p-channel transistors in saturation mode. The counter outputs are toggled between VDD (off-state) and VDD-V_{ref} (saturation), where V_{ref} is a tunable voltage slightly larger than the p-channel threshold voltage, which ensures the p-channel transistors are biased in saturation. When the input voltage is low, an n-channel transistor in parallel with R_{CBRAM} is activated to shunt current to ground, which debiases the R_{CBRAM} during that half cycle of the input pulse.

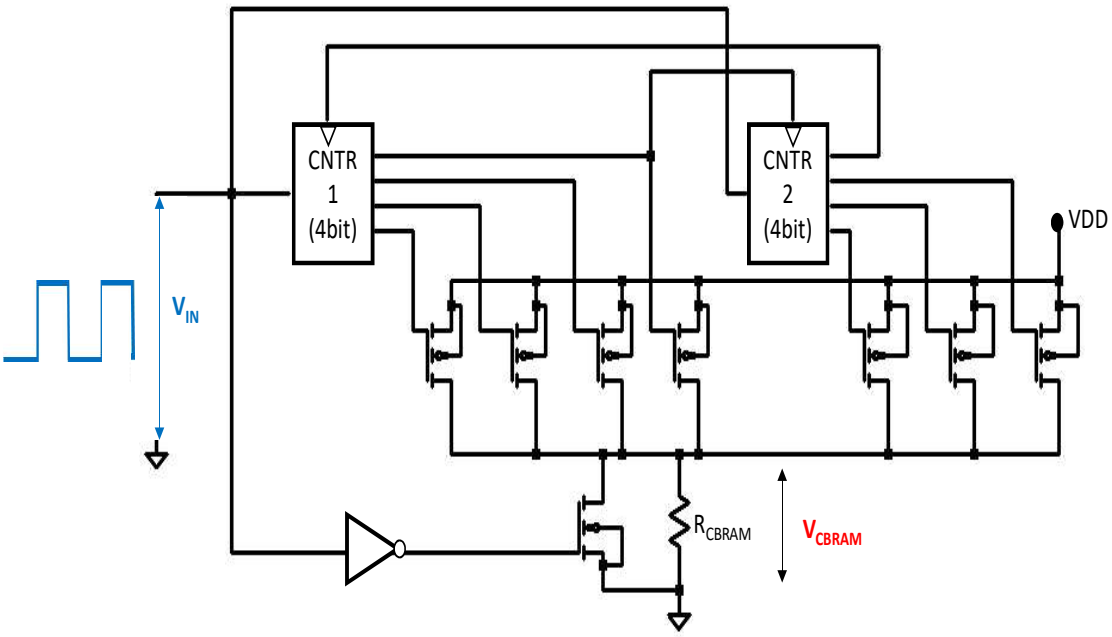


Fig. 118. Switched current design for potentiation.

The increasing current magnitude sequence selected by the counter is determined as follows. To achieve a V_{CBRAM} that increases approximately linearly as R_{CBRAM} is reduced, the following condition must be satisfied:

$$V_{CBRAM,i} = R_{CBRAM,i} I_i \approx M R_{CBRAM,i} + B. \quad (19)$$

In Eq. 19, I_i is the selected current at pulse number i. M is a negative value and B is positive value set to match the targeted (V_{CBRAM}, R_{CBRAM}) order pairs, e.g., (0.2V,2kΩ) and (1.36V, 730Ω). By rearranging Eq. 19, the targeted I_i is therefore,

$$I_i \approx M + \frac{B}{R_i} \quad (20)$$

Fig. 119 shows the simulated response of R_{CBRAM} and V_{CBRAM} with the input voltage pulse sequence for V_{IN}. It should be noted that using switch programming, the hardware of the potentiation circuit (Fig. 118) can be reconfigured for use in R_{CBRAM} depression.

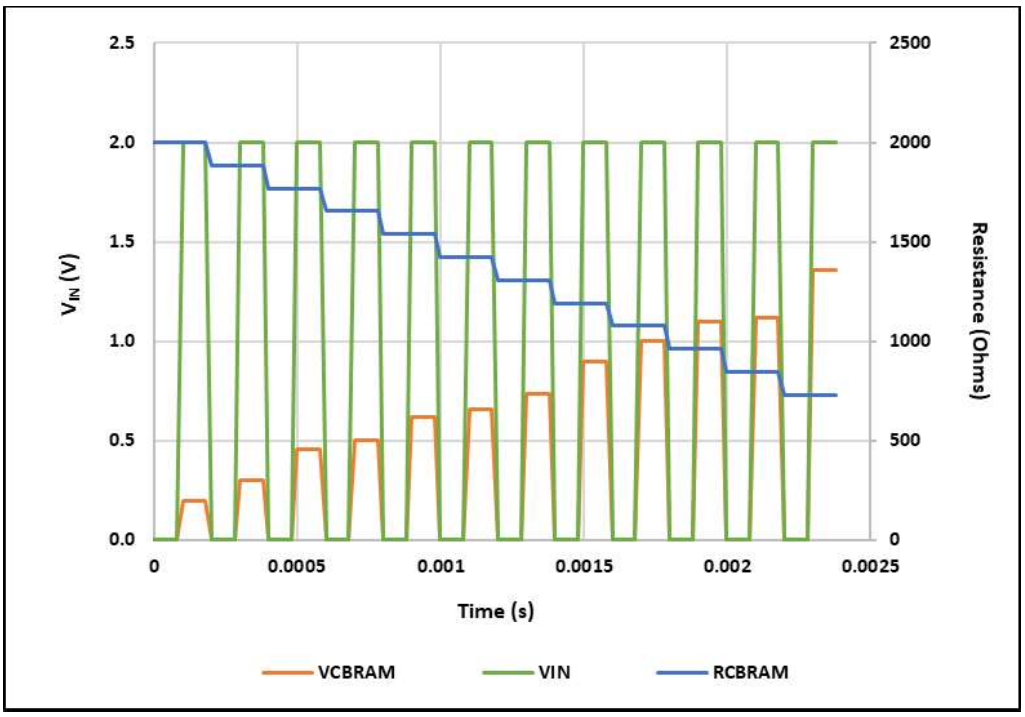


Fig. 119. Reducing Resistance and increasing VCBRAM for input voltage pulse sequence.

Task 3.2 Post-rad testing of neuro-crossbar RRAM array (with peripheral neuron circuits)

Perform TID tests on 256×256 1T1R RRAM Array

Irradiation Test Conditions

Fig. 120 plots our final 4-state conductance distribution in a control on the test chip that would be irradiated. Before the programming, the HRS and LRS values are determined to be state 1 and state 4, which is smaller than 1.25uS and centered at 156uS, respectively. The targeted mean conductance of intermediate states is acquired by adding equal interval conductance between HRS and LRS, with state 2 to be 52uS and state 3 to be 104uS. The tolerance of each state is set to be +/- 10% of its mean conductance. The active 1T1R cell in the control was forced to be programmed identically as the radiation response test chip to eliminate the effect of voltage stress on the radiation results.

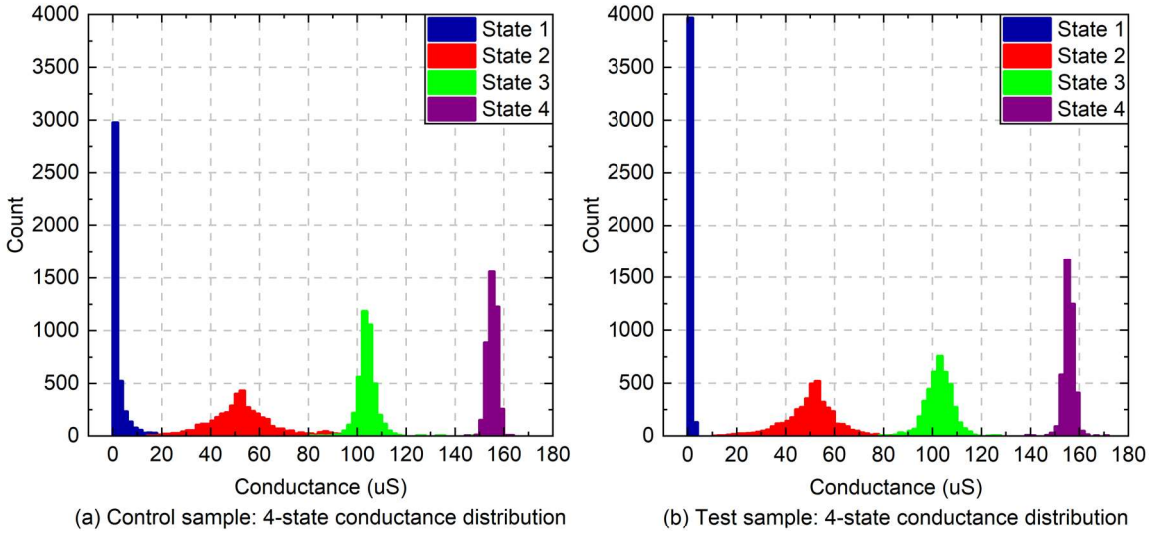


Fig. 120. 4-state conductance distribution in HfO_x-based RRAM 1T1R array after verification. Each state has 4096 1T1R cells tested. (a) control sample; (b) test chip

After programming the 4-state conductance in the RRAM array, we irradiated the entire test chip with different levels of total ionizing dose (TID) under Co-60 gamma ray: 36krad[Si], 255krad[Si], 358krad[Si], 652krad[Si], 932krad[Si], 1.9Mrad[Si], and 2.9Mrad[Si]. Both the test chip and the control sample were identically biased during irradiation in a read operation mode: addr[15]~addr[0]=0V (digital low), V_SL=0V, V_BL=0.2V. This bias condition results in a 3.3V DC voltage applied to WL[0], while all other WLs are biased to digital low. In columns, all the cells in the first column are biased to V_BL=0.2V, while other columns are shorted to V_SL=0V. In this bias configuration, the first 1T1R cell in the array has been biased for read operation, while the other cells are inactive.

Pre- and Post-irradiation Testing Results

Fig. 121 shows the comparison between the pre- and post-irradiation results of the 4-state conductance distributions in the test chip and control sample. After the time required to reach each targeted TID level, a read operation was performed for both chips to obtain the 4-state conductance distributions. States 1 and 4 correspond to the HRS and LRS and states 2 and 3 are two intermediate states. Figure 121a shows that the programmed cells in the control sample still remains a similar 4-state conductance under the same bias condition as the TID test chip during the radiation test. Notice that, the 0.2V read voltage stress alone does not affect the conductance of the active cell, which is stable in its high resistance state (HRS) in the control sample.

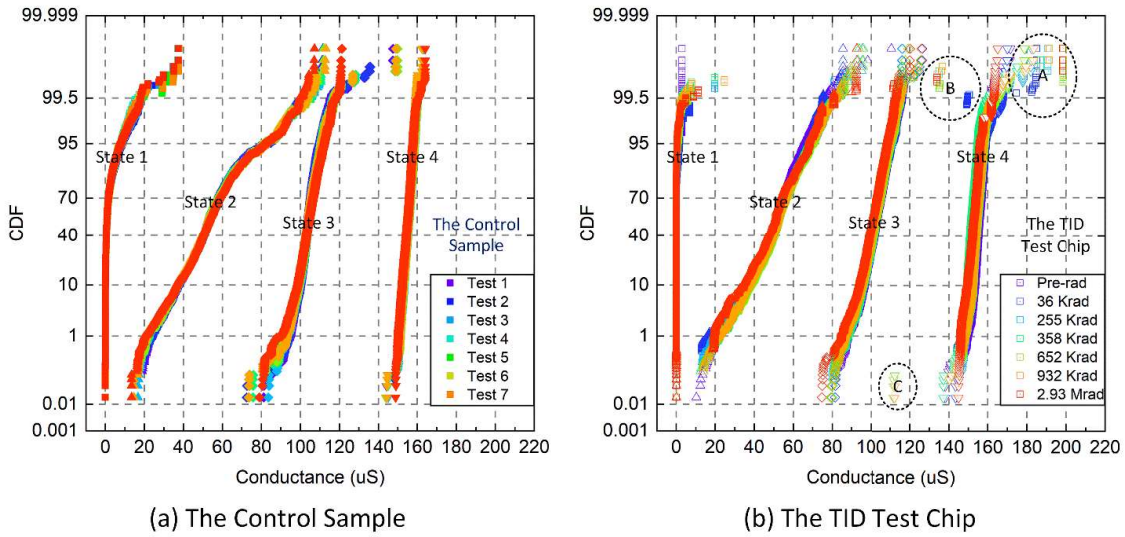


Fig. 121 Cumulative Distribution Function (CDF) of the Conductance for the 4 States in (a) Control Sample and (b) TID Test Chip.

In Fig. 121b, the cumulative distribution function (CDF) of the 4-state conductance in the TID test chip indicates a small impact on the intermediate states (state 2 & state 3) when the chip is exposed to TID levels up to 2.93 Mrad(Si). Indeed, some shifts in cells were detected during the TID exposure. First, the active cell that was biased as read operation during irradiation, shifted from HRS to LRS, and was stabilized on LRS eventually. Second, a few HRS cells in state 1 switched between HRS and LRS (Fig. 121b, dot circle A, called behavior A). Third, some cells in state 1 switched between HRS and the intermediate states without consistent degradation (Fig. 121b, dot circle B, behavior B). Moreover, some anomalous LRS cells in state 4 switched between LRS and the intermediate state without consistent degradation (Fig. 121b, dot circle C, called behavior C). Table IX shows the number and percent of the anomalous cells that have the behaviors A, B, and C following total ionizing dose (TID). Each state was programmed into 4096 cells. Behavior A and B both occurred in state 1, high resistance state (HRS). At 36 krad(Si), there were 16 cells exhibiting behavior A, 16 cells with behavior B, and no cells showing behavior C. In state 1, the percent of the cells exhibiting behavior A decreased from 0.391% (16 cells) to 0.073% (3 cells) with the increase of TID from 36 krad(Si) to 358 krad(Si). When the TID level reached 652 krad(Si), the percent of behavior A cells increased to 0.195% (8 cells), which would decrease to 0.049% at 932 krad(Si) and increase to 0.098% (4 cells) at 2.93 Mkrad(Si) TID. For behavior B, the anomalous cells decreased from 0.391% (16 cells) to 0 when TID level was 255 krad(Si); it would increase and hold a value at 0.098% when TID was larger than 652 krad(Si). This may be caused by the radiation-induced oxygen vacancies in the switching layer and subsequent decrease in the conductive filament gap changing the resistance into a lower resistance. Behavior C occurred in state 4, low resistance state (LRS). This behavior may be due to damage to the conductive filament caused by radiation-induced non-lattice oxygen ion reactions with oxygen vacancies around the filament region, rupturing the filament. This would increase the resistance of cell. Behavior C only occurred under 652 krad(Si) and 932 krad(Si) TID levels. Once TID level reached to 652 krad(Si), there are 0.098%

(4 cells) cells in LRS showing behavior C and this percent would decrease to 0.094% (2 cell) when the TID progressed to 932 krad(Si). No behavior C occurred during TID below 652 krad(Si) and above 932 krad(Si). In general, those few anomalous cells have a negligible impact on 4-state conductance in full array.

The conductance accuracy is defined as the percentage of the cells within the pre-defined +/- 10% tolerance, and it is reported in Table X. The data indicates the nearly negligible impact on the 4-state conductance accuracy under 2.93Mrad(Si) irradiation. Compared to the pre-radiation data, the TID results demonstrate the non-HRS states (state 2/3/4) are rather stable, while the HRS state 1 shifted to higher resistance on average. It should be noted that, although few anomalous cells appeared in state 1 (HRS), the average resistance in state 1 has been enhanced following the accumulation of TID level. It seems that for the most cells, the radiation-induced non-lattice oxygen ions distributed near the conductance filament region and recombined with original oxygen vacancies increases the conductance filament gap, forcing the resistance into a higher resistance, yet still within the state 1 range.

Table IX
Number and percent of Anomalous Cells Following TID

TID Level	Behavior A	Behavior B	Behavior C
36 krad(Si)	16 (0.391%)	16 (0.391%)	0 (0%)
255 krad(Si)	4 (0.098%)	0 (0%)	0 (0%)
358 krad(Si)	3 (0.073%)	2 (0.049%)	0 (0%)
652 krad(Si)	8 (0.195%)	4 (0.098%)	4 (0.098%)
932 krad(Si)	2 (0.049%)	4 (0.098%)	2 (0.049%)
2.93 Mrad(Si)	4 (0.098%)	4 (0.098%)	0

Table X
The Multi-state Conductance Accuracy

Total Ionizing Dose (TID)	State 1 <1.25uS	State 2 52uS±10%	State 3 104uS±10%	State 4 156uS±10%
0 krad(Si)	90.70%	49.46%	94.12%	99.80%
36 krad(Si)	91.63%	45.34%	92.26%	100%
255 krad(Si)	97.41%	47.00%	92.65%	99.80%
358 krad(Si)	97.36%	46.41%	93.26%	99.80%
652 krad(Si)	95.80%	46.04%	92.48%	99.80%
932 krad(Si)	97.58%	44.21%	91.60%	99.80%
2.93 Mrad(Si)	97.88%	43.51%	90.09%	100%

To characterize the impact of TID on the 1T1R RRAM cell as well as on the CMOS periphery after collecting the post-irradiation results, we reprogrammed the 4-state conductance with the previous programming scheme and bias condition in the test chip. After one-month anneal, all states tended to relax so there was the discrepancy of the data before the TID irradiation and the initial programming. An alternative way to analyze is by comparing the reprogrammed data (as post-irradiation results) with the initial programming data (as pre-irradiation results) to access the TID response. The cumulative distribution function (CDF) plot in Fig. 122 and the histogram distribution plot in Fig. 123 exhibits the comparison of pre- and post-irradiation programming results in the test chip as well as pre- and post-test programming results in the control sample. The distributions of all state appeared to be improved after reprogramming. The CMOS periphery also appears to function well, and the active cell is able to reset to HRS. Moreover, the reprogramming time for the TID test chip was much less than that for the control sample. Compared to the reprogramming data and the programming time in the control sample, we may contend that TID exposure can make the creation of conductive oxygen vacancy filaments more sensitive to bias voltages. The consistency of 4-state conductance demonstrates that our 64 kb RRAM prototype chip is mostly tolerant to TID and moreover is able to achieve a better 4-state conductance distribution after a TID exposure.

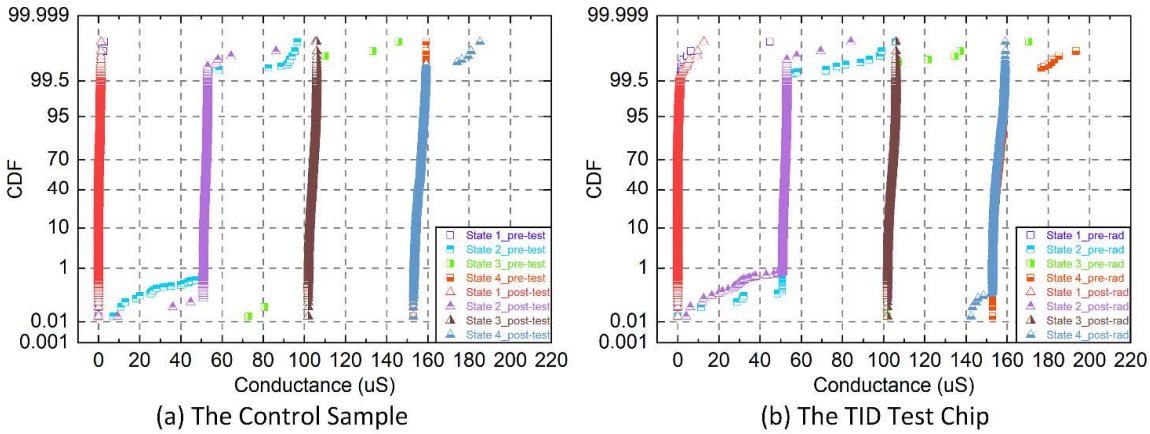


Fig. 123. Comparison for Programming Results in Control Sample and TID Test Chip Before and After Irradiation. (a) Control Sample, (b) TID Test Chip.

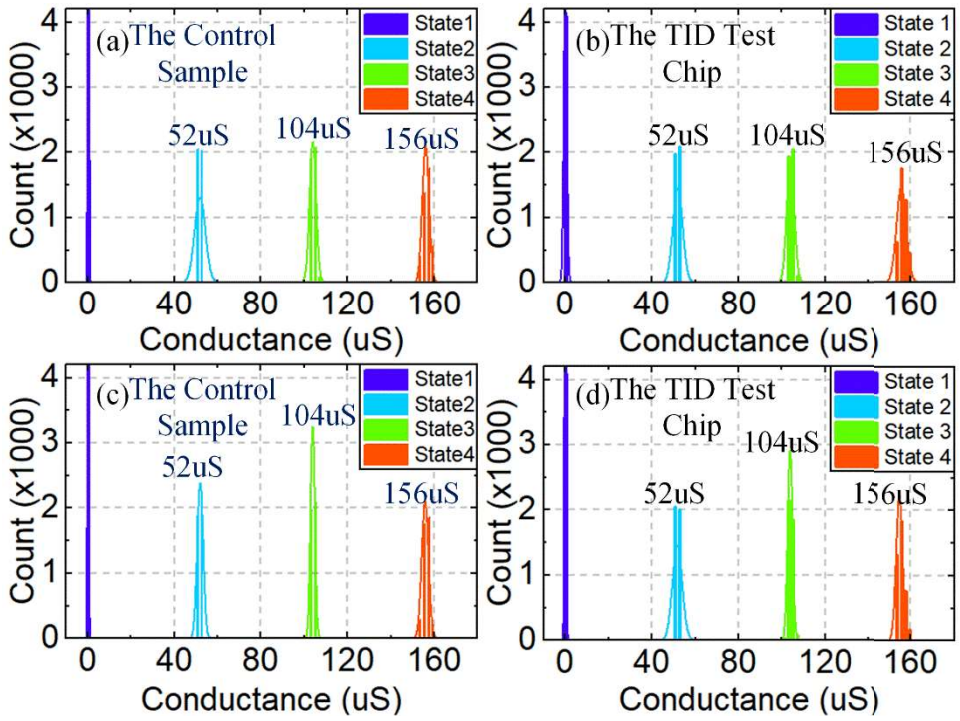


Fig. 123. Histogram Distribution of Programming Data in (a) Control Sample Before Irradiation, (b) TID Test Chip Before Irradiation, (c) Control Sample After Irradiation, and (d) TID Test Chip After Irradiation.

Task 3.3 Incorporate device and circuit-level radiation models into deep neural network

Model TID effects in SRAM-based multi-layer perceptron network.

XNOR Computation

In the SRAM neural network considered in the program, an XNOR operation is executed with weights stored in memory and input signals from the preceding layer. Recent 65nm SRAM XNOR-and-accumulate (XAC) in-memory computing prototype chips are reported in Table XI.

Table XI
Comparison Table of Recent SRAM-based XAC IMC Accelerators

	16Kb SRAM CIM [41]	XNOR-SRAM [33]
Technology	65 nm	65 nm
Bit-cell Area (μm^2)	1.8 × 4	1.45 × 2.7
Array Size	128×128	256 × 64
Operating Voltage	0.5-1.0 V	0.6-1.0 V
Supported Algorithm	MLP/CNN	DNN/CNN
Energy per Operation	1bit ADC: 1.3-16 fJ 5bit ADC: 11.5-135 fJ	2.48-7.19 fJ
EDP Reduction	7.1-18.5X (0.5V)	288-330X
MNIST Accuracy	96.62%	98.3%
CIFAR-10 Accuracy	86.51%	85.7%

XNOR-and-Accumulate Computation

The SRAM-based XNOR design has two modes of operation: 1) memory mode and 2) XNOR mode. In the memory mode, the circuit is performing the training weight update. In the XNOR mode, the circuit executes the in-memory XNOR logic and summing (accumulate) computations, i.e., classification.

Fig. 123 displays a customized design of the XNOR-SRAM bit-cell [33]. In the cell, transistors T1 to T6 form the standard 6T SRAM cell containing one write word-line (WWL), and two complimentary write bit-lines (WBL & WBLB).

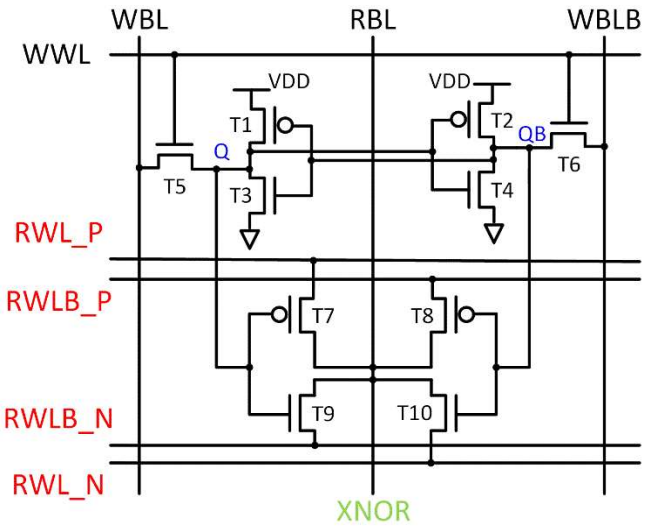


Fig. 123. Customized XNOR-SRAM Bit-Cell.

In memory mode, a write operation to the 6T SRAM corresponds to the artificial "synapse" weight update, which is required for network training. Once written, the 6T SRAM stores a weight ($w_{i,j}$) for subsequent training or classification operations. For this design, if $w_{i,j}$ is "+1", QB equals to VDD while Q is 0V. If the stored weight is "-1", Q equals to VDD while QB is 0V. Transistors T7 to T10 form the pull-up/pull-down circuits for both the XNOR classification and memory mode read operations. If the input is "+1", the PMOS read word-line (RWL_P) and NMOS read word-line (RWL_N) are at VDD while RWLB_P and RWLB_N are at 0V. If the input is "-1", RWL_P and RWL_N would be pulled down to 0V and RWLB_P and RWLB_N pulled up to VDD. The input and weight XNOR logic operation will occur through transistors T7 to T10. For a single cell, the XNOR result sets the read bit-line (RBL) voltage to either 0V or VDD. A truth table for single cell XNOR logic is given in Table XII.

Table XII
Control Table of XNOR Bit-cell

Weight \ Input		+1 *RWL_P/N = VDD* *RWLB_P/N = 0*	-1 *RWL_P/N = 0* *RWLB_P/N = VDD*
+1	Q = 0 QB = VDD	+1 *RBL = VDD*	-1 *RBL = 0V*
-1	Q = VDD QB = 0	-1 *RBL = 0V*	+1 *RBL = VDD*

Fig. 124 provides schematics describing the four XNOR logic modes. In a selected column, T7 to T10 in each bit-cell perform XNOR on the inputs (RWL_P/N and RWLB_P/N) and the stored weight. The inputs are the source voltages for T7 to T10. When the input is “+1”, the source voltage of T7 and T10 is at VDD while the source voltage of T8 and T9 are 0V, vice versa for the “-1” input. The stored weight ($w_{i,j}$) from the 6T SRAM, held by Q and QB, sets the gate voltages for T7 to T10. When the Q and QB equal to 0V and VDD respectively (weight “+1”), T7 and T10 are turned on while T8 and T9 are off. When the Q and QB equal to VDD and 0V respectively (weight “-1”), T8 and T9 are turned on while T7 and T10 are off. In the XNOR logic computation, for a stored weight “+1”, if the input is “+1”, the XNOR output will be “+1”, that is, RBL will be fixed to VDD (+1), as shown in Fig. 124a. For a stored weight “+1”, if the input is “-1”, RBL will be fixed to 0V (-1), as shown in Fig. 124b. For the stored weight “-1”, if the input is “+1”, RBL will also be at 0V (Fig. 124c). And lastly, for a stored weight “-1”, if the input is “-1”, the XNOR output will be VDD (Fig. 124d). In general, if input and weight are the same, the XNOR outputs a digital high, through the combination of a strong pull-up PFET and a weak pull-up NFET. Otherwise, the XNOR output is digital low which equals to “-1”, resulting in a strong pull-down NFET and a weak pull-down PFET. To parallelize the XNOR-and-accumulate operation in the array, multiple read word lines are activated simultaneously and the XAC is output as the column RBL voltage. Each bit-cell in the column will perform an individual XNOR logic to pull up or pull down the column RBL voltage. The combination of XNOR outputs from all bit-cells in a column form a voltage divider, where RBL is the output.

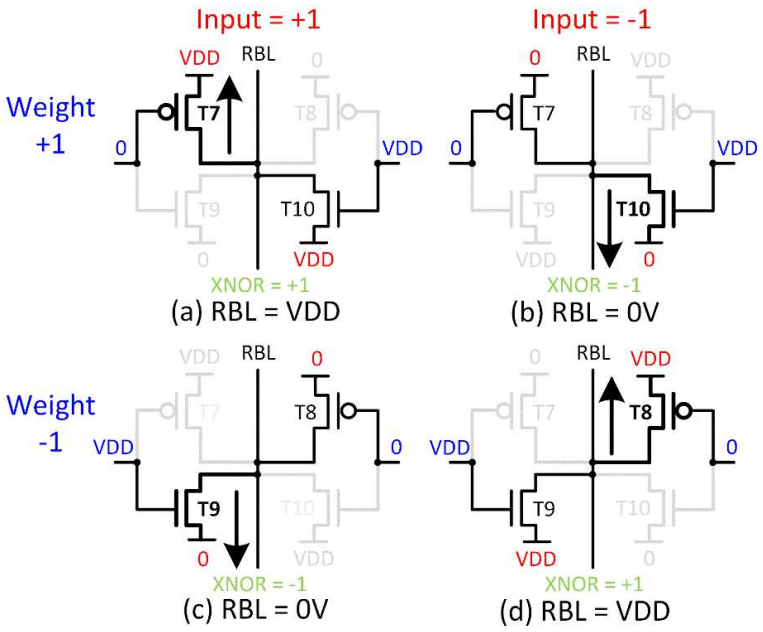


Fig. 124 The Illustration of XNOR Operation in Pull-Up/Pull-Down Circuits

XAC Array Modeling

Based on the in-memory XNOR architecture, we designed an 8×8 single layer network (array) with GlobalFoundries 22nm SLVT enhanced gate transistors. A schematic of the array is shown in Fig. 125. There are 8 parallel row inputs, 8 parallel XNOR column outputs (RBL), and 64 independent stored weights. In XNOR mode, all row inputs turn on simultaneously. In each column, every bit-cell processes its own XNOR logic. The resulting output for each column is an accumulation of the XNOR logic in each bit-cell in that column. Since the 8 parallel inputs would have 9 conditions, shown in Table II, i.e., 8 inputs = “-1”; 1 input = “+1”, 7 inputs = “-1”; 2 inputs = “+1”, 6 inputs = “-1”; 3 inputs = “+1”, 5 inputs = “-1”; 4 inputs = “+1”, 4 inputs = “-1”; 5 inputs = “+1”, 3 inputs = “-1”; 6 inputs = “+1”, 2 inputs = “-1”; 7 inputs = “+1”, 1 input = “-1”; 8 inputs = “+1”, each RBL column would have 9 corresponding voltage levels from 0V to VDD (1.8V).

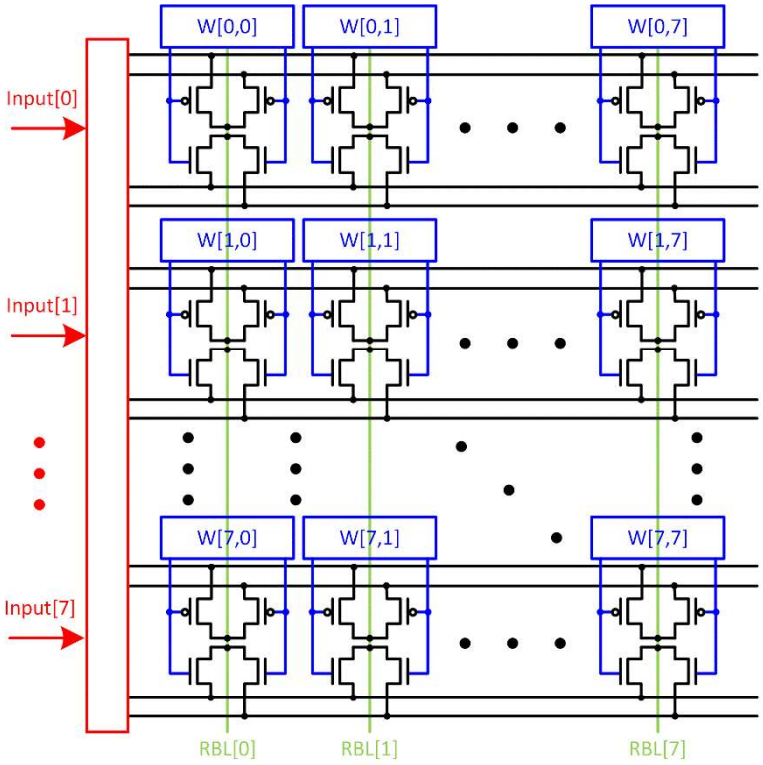


Fig. 125. Schematic of 8×8 XNOR Array

Table XIII
The 9 Corresponding RBL Voltages with All Weights "+1"

CONDITION	INPUT +1	INPUT -1	WEIGHT +1	DESIGNED RBL VOLTAGE	ACTUAL RBL VOLTAGE
1	0	8	8	0	Level 1 = 0V
2	1	7	8	1/8 VDD	Level 2 = 102.6 mV
3	2	6	8	2/8 VDD	Level 3 = 240 mV
4	3	5	8	3/8 VDD	Level 4 = 455.6 mV
5	4	4	8	4/8 VDD	Level 5 = 750.4 mV
6	5	3	8	5/8 VDD	Level 6 = 1.03 V
7	6	2	8	6/8 VDD	Level 7 = 1.285 V
8	7	1	8	7/8 VDD	Level 8 = 1.564 V
9	8	0	8	VDD	Level 9 = 1.8 V

In order to simulate all possible combinations, eight inputs are pulsed with different pulse widths. Fig. 126 shows a transient simulation of one column RBL voltage in our 8×8 XNOR array. Each column has 8 independent stored weights. In that column, if all 8 weights are set to "+1" and all 8 inputs are "+1", each XNOR logic result of those 8 parallel bit-cells would be "+1", that is, all 8 bit-cells would simultaneously pull up the column RBL voltage to VDD. If all weights are set to "+1" and all 8 input are "-1", each XNOR logic results in the 8 parallel bit-cells set to "-1", resulting in a column RBL voltage set to 0V. If all weights are "+1" and if 7 inputs are "+1" while one input is "-1", in that column, 7 bit-cells would pull-up while 1 bit-cell would pull-down. Here the resulting RBL voltage would be ~ $\frac{7}{8} \times \mathrm{VDD}$. With the same weights condition, if 7 inputs are "-1" while the rest of the inputs are "+1", in that column, 7 bit-cells would pull-down while only 1 bit-cell would pull-up, the resulting RBL voltage would be ~ $\frac{1}{8} \times \mathrm{VDD}$. Depending on the XNOR logic results of those 8 parallel bit-cells, the column RBL voltage has 9 levels. For our simulation, the 9 output RBLs for one column (bottom curve of Fig. 126) are 0V, 102.6mV, 240mV, 455.6mV, 750.4mV, 1.03V, 1.285V, 1.564V, 1.8V. These results demonstrate the ideal (pre-rad) operation of XAC function.

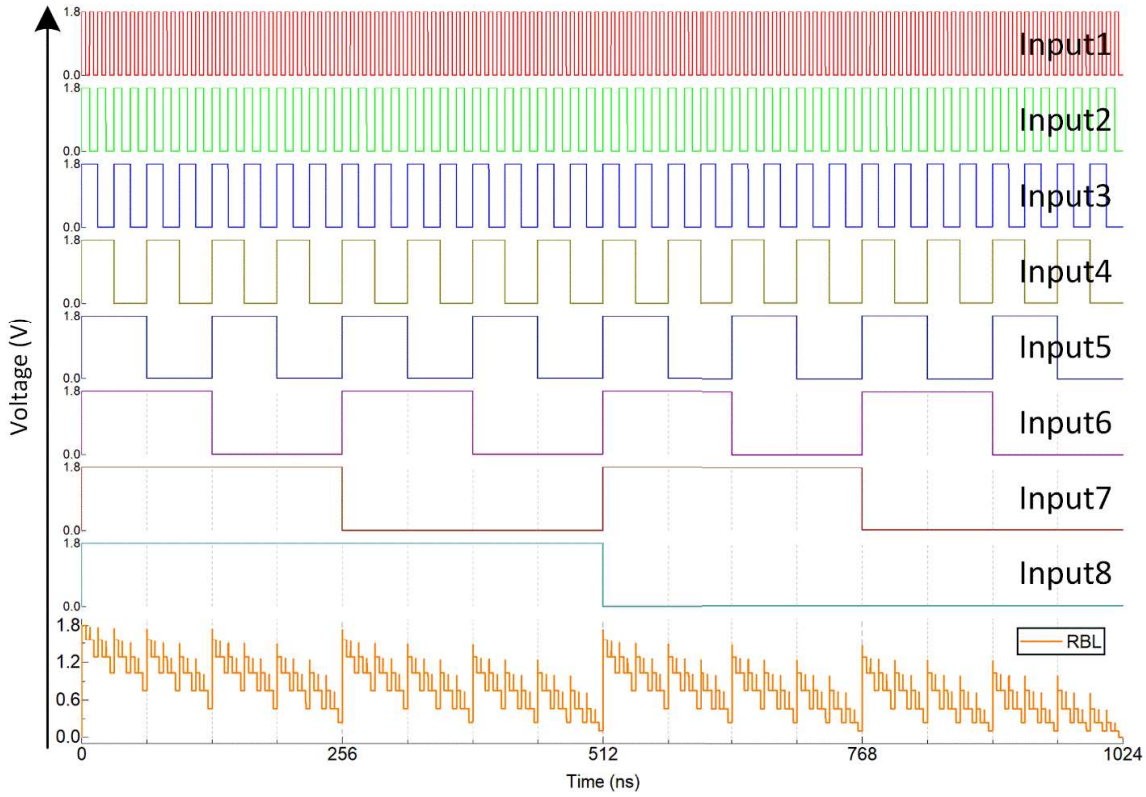


Fig. 126. Simulation Results of RBL Voltage for One Column When All Weights are “+1”.

TID Effects on SRAM-based XAC Array

The effects of total ionizing dose on SRAM memory have been the subject of extensive study. However, to date, little research has been devoted to the study of ionizing radiation effects in neuromorphic (non-von Neumann) digital systems. The circuits are designed with flipped-well enhanced-gate super-low threshold voltage (SLVT) MOSFETs manufactured in the Global Foundries 22nm FDSOI process. TID experimental data on n- and p-channel FDSOI transistors were used to generate the radiation-aware models as discussed in Task 1.1 . Circuit-level simulations that incorporate these models show how exposure to ionizing radiation can impact the operation of a SRAM-based XNOR IMC neuromorphic system.

TID Effects on XAC Computation in XNOR Array

To simulate TID effects on XAC computation, we replaced all SLVT MOSFETs in the 8×8 XNOR array with our radiation-aware models (i.e., added defect potential for specific dose levels) and ran the previous simulations to evaluate the change in corresponding XNOR-and-accumulate output (RBL) voltage levels. The simulations reveal that our 8×8 XNOR array is capable of maintaining the 9-level RBL voltages after irradiation with a voltage shifts within ±9.5% of the pre-radiation voltages. The TID-induced RBL voltage shifts for each level are presented in Table V. Only the intermediate RBL voltage outputs exhibit significant shifts, although, the voltage levels are still within an acceptable range. After 200krad(SiO_2), the maximum voltage shift for Levels 6 – 8 is less than 2%. For the same dose, the RBL voltages for Levels 2, 3, 4 and 5 have

shifted 2.622%, 7.333%, 9.592%, and 3.398%, respectively. This demonstrates that lower voltage levels are more sensitive and that higher voltage levels have better immunity to TID effects. These general patterns are also observed at the lower doses.

Table IXX
The TID-induced RBL Voltage Shift

RBL VOLTAGE	TOTAL IONIZING DOSE LEVEL			
	0 krad	50 krad	100 krad	200 krad
Level 1 = 0V	0	0	0	0
Level 2 = 102.6 mV	0	-2.339 %	-2.866 %	-2.622 %
Level 3 = 240 mV	0	-3.125 %	-4.500 %	-7.333 %
Level 4 = 455.6 mV	0	-3.995 %	-5.838 %	-9.592 %
Level 5 = 750.4 mV	0	-0.693 %	-1.226 %	-3.398 %
Level 6 = 1.03 V	0	+0.680 %	+1.068 %	+1.068 %
Level 7 = 1.285 V	0	+0.778 %	+1.323 %	+1.790 %
Level 8 = 1.564 V	0	-0.895 %	-1.087 %	-1.087 %
Level 9 = 1.8 V	0	0	0	0

TID Effects on Power Consumption in XNOR Array

Fig. 127 shows how TID affects the power consumption of the SRAM-based XNOR array. For the array, there are 64 stored weights. When all 64 stored weights are “+1”, the power consumption increases gradually with the increasing TID levels. When all stored weights are “-1”, the 8×8 XNOR array has the lowest power consumption, which, interestingly, decreases after 50krad(SiO_2) before rising for the higher doses. When half of stored weights are “+1” and half of them are “-1”, the 8×8 XNOR array has the maximum power consumption, but stays is relatively constant under various ionizing radiation exposure levels ranging from 50krad(SiO_2) to 200 krad(SiO_2).

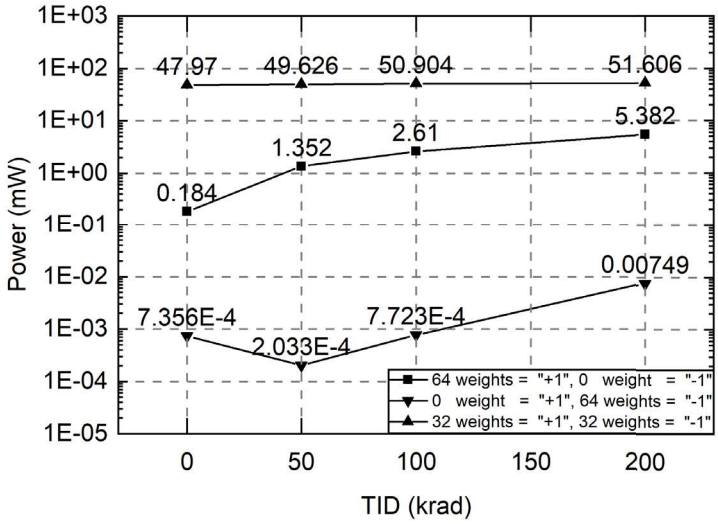


Fig. 127. Influence of TID-Induced Voltage Shifts on Power Consumption of 8×8 XNOR Array

TID Combined with Temperature Effects on XAC Computation

For MOSFETs devices, when temperature rises, the threshold voltage decreases while the subthreshold leakage current increases exponentially. Therefore, it would be interesting to evaluate TID combined with temperature effects on the XAC array.

For temperature effects alone, our 8×8 XNOR array also can maintain 9-level RBL voltage but with a larger voltage shift. Typically, RBL voltage Level 1 (0V) and Level 9 (1.8V) are stable under various temperatures. Fig. 128 shows how temperature affects the intermediate RBL voltage levels. Each level RBL voltage at room temperature (27°C) is set to its reference for corresponding level, that means no voltage shift occurs at room temperature. The test temperature range is set from -55°C to 125°C. The simulation results indicate that the RBL voltage for Level 2, 3, 4 and 5 increase gradually with the increasing temperatures from -55°C to 125°C. The RBL voltage for Level 6 and 7 are rather stable with increasing temperature. The RBL voltage for Level 8 would decrease with the increasing temperature. The various temperature simulation results reveal that the lower level (Level 2, 3, 4) RBL voltage, which shift above ±10% of nominal, are more sensitive to the temperature effects.

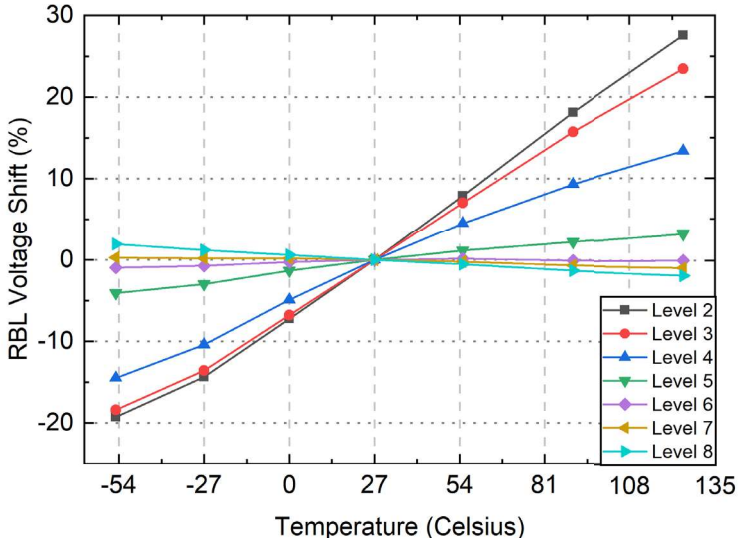


Fig. 128. Temperature Effects on RBL Voltage Shifts in 8×8 XNOR Array.

The TID combined with temperature effects on intermediate RBL voltage levels is shown in Fig. 129. For each RBL level at each temperature, its pre-rad result is set as a reference for calculating the TID-induced RBL voltage shift. In Fig. 129a, under the same TID level, the voltage shift for RBL Level 2 decrease with increasing temperature from -55°C to 125°C. At room temperature (27°C), the voltage shift for Level 2 increased when TID reached 100krad while decreased at 200krad(SiO2). Below room temperature (27°C), the voltage shift for Level 2 increased with rising TID until reached 100krad (SiO2). After 200krad(SiO2) TID, the voltage shift for Level 2 does not degrade consistently. The voltage shifts for Level 3, 4, and 5 have similar trends under the TID and temperature effects (Fig. 129b, c, d). For the same TID level, the voltage shift for Level 3, 4, and 5 gradually increased with the rising temperature from -55°C to 125°C. For the same temperature, the voltage shift for Level 3, 4, and 5 increased with increasing TID levels from 50krad(SiO2) to 200krad(SiO2). The higher level RBL voltages (Level 6, 7, and 8) have some interesting responses to TID combined with temperature effects. For Level 6 RBL voltage, under 50krad(SiO2) TID, the voltage shift was rather stable when temperature changed from -55°C to 55°C. As temperature raised from 55°C to 90°C, the voltage shift for Level 6 increased from 0.679% to 0.777% and decreased to 0.68% at 125°C. As TID reached 100krad(SiO2), the temperature rises from -55°C to 0°C and 27°C to 125°C caused rarely changes on voltage shift for Level 6, but as temperature raised from 0°C to 27°C, the voltage shift for Level 6 increased from 0.974% to 1.068%. When TID reached 200krad(SiO2), the increasing TID does not affect the Level 6 voltage shift at temperature -55°C, 0°C, 27°C and 55°C, but the voltage shift decreased from 0.98% to 0.88% when temperature changed from -55°C to -27°C, and then the voltage shift increased with rising temperature until 125°C. For Level 7 RBL voltage, at same temperature, voltage shift increased with increasing TID. Under same dose, temperature has little effects on voltage shift for Level 7. The voltage shift for Level 8 has similar response to TID combined with temperature effects as that for Level 2.

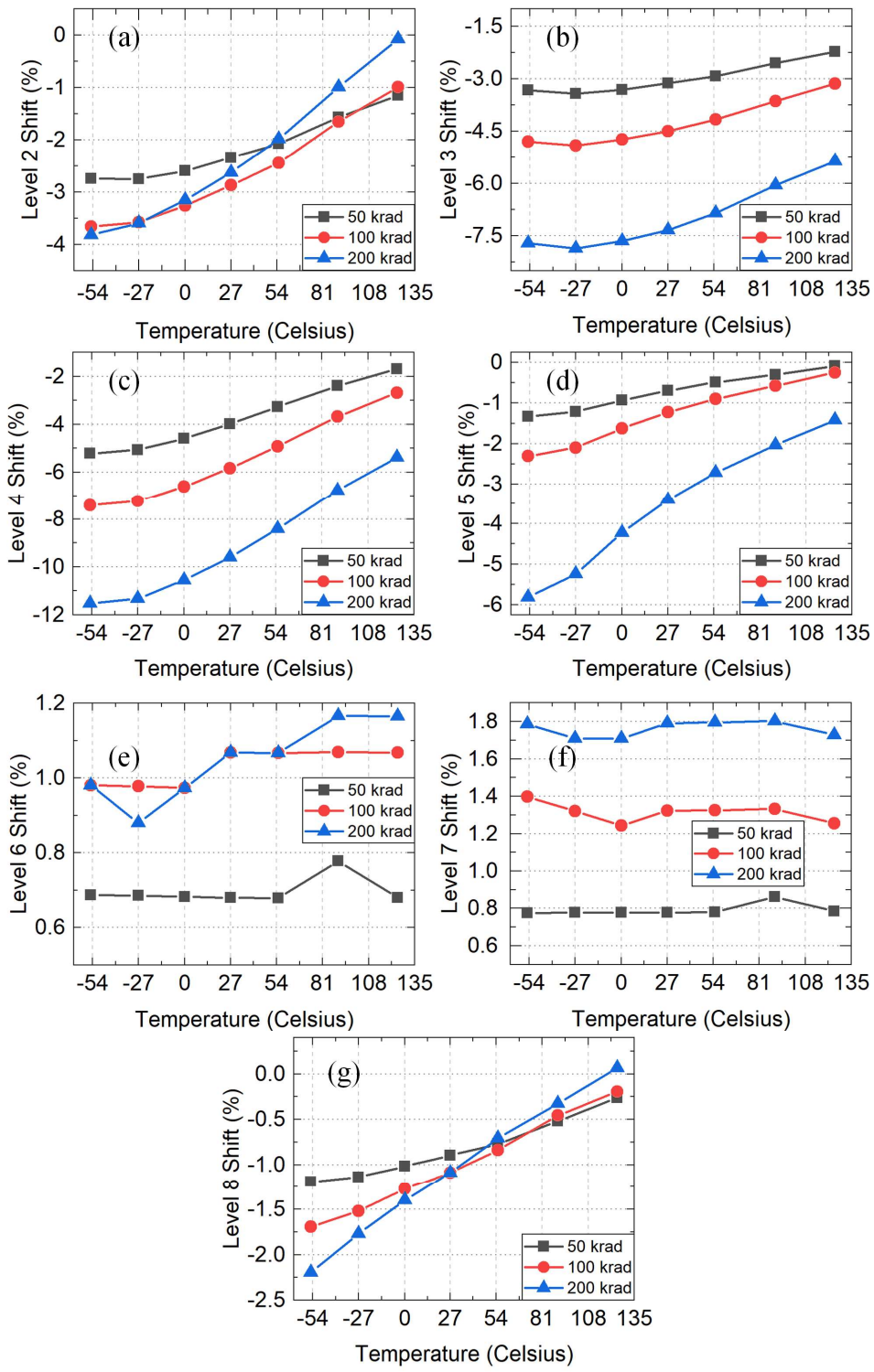


Fig. 129. TID Combined With Temperature Effects on the Intermediate RBL Voltage in 8×8 XNOR Array. (a) RBL Level 2 = 102.6mV, (b) RBL Level 3 = 240mV, (c) RBL Level 4 = 455.6mV, (d) RBL Level 5 = 750.4mV, (e) RBL Level 6 = 1.03V, (f) RBL Level 7 = 1.285V, (g) RBL Level 8 = 1.564V.

Task 4. Radiation Effects on Full-scale Neuromorphic System

Task 4.1 Fabrication, and test unhardened CMOS neuromorphic system

Fabricate multi-layer perceptron (vector matric multiply – VMM) neuromorphic system without radiation hardening techniques.

RRAM Process integration

The primary neuromorphic system architecture in this program is a CMOS-based perceptron that performs a VMM operation by implementing a controlled pulsed write into and digital readout of an 8x7 integrated nonvolatile RRAM array. The design of the system was presented in Task 2.2. The VMM column read operations sum the currents created from input voltages impressed across a set of RRAM cells sharing a common lead as shown in Fig. 130. The dot product sums are converted to a 10-bit digital word using an on-chip analog-to-digital converter (ADC). Each row or column in the array can be individually addressed and isolated to either be included or excluded for any given read or write operation. Also, on-chip is a variable width pulse generator used to control the width pulses that alter the array's RRAM conductance during write operations. Configuration of the control bits for array addressing, transimpedance gain, and test modes are all done through a simple shift-register chain which is part of the ADC circuitry. The embedded RRAM array is divided into 8 columns. A VMM operation is just 8 dot product operations performed in parallel. Every row or column in the array has 2 mode bits that control how it is hooked up to the shared row and column drivers. This flexibility allows for subsets of the array to be targeted for a set of operations while floating or "idling" the other ones. The 8th column is populated with fixed poly resistors instead of RRAM devices and serves as a test column to evaluate the rest of the readout circuitry. This means there are only 8 rows and 7 columns of RRAM devices.

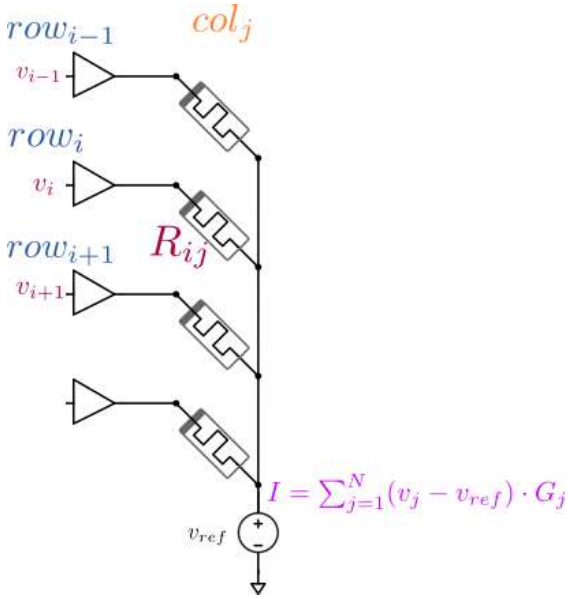


Fig. 130. Single 8x1 RRAM array for carrying out VMM operation.

The RRAM in the 8x7 device array is a vertical PMC or CBRAM. Its structure consists of 3 layers (Ag/Ag-$GeSe_3$/Ti)- a silver (Ag) anode, a germanium selenide chalcogenide doped with silver (Ag-$GeSe_3$) as electrolyte and a titanium (Ti) cathode. The electrodes are to be deposited in an 8x7 array of BEOL metal stacks. And each unit in this 8x7 electrode stack consists of top-level metal electrodes: - a top electrode (TE) acting as the anode contact and the bottom electrode (BE) acting as the cathode contact for the RRAM cell. In ascending order, the top-level metal stacks are composed of a 15 nm thick titanium layer, a 25 nm thick titanium nitride (TiN) layer, a 675 nm thick aluminum-copper alloy which is capped off by a final 25 nm thick layer of TiN acting as an anti-reflective coating (ARC) as shown in Fig. 64. Thus, the electrodes are Ti/TiN/Al-5Cu/TiN ARC stacks where the bottommost Ti layer is an inert electrode material, Al-5Cu is an alloy of aluminum (Al) with 5% copper (Cu) by weight and the TiN ARC prevents reflection of light from the underlying highly reflective metal layers during photolithography. Each stack is 740 nm thick and has the same composition of materials that is Ti/TiN/Al-5Cu/TiN ARC. The TE and the BE have different shapes as seen in Fig. 131. The TE and BE stacks are 2 μm apart from each other and the BE of one array column is 23.75 μm from the nearby metal routing which in turn is 68.6 μm away from the TE of the subsequent array column of electrode stacks.

ASU Process

In Fig. 132, the blue-colored Ag layer forms the active anode and the pink-colored Ag doped $GeSe_3$ layer forms the electrolyte. Note that the BE never touches the Ag anode layer to prevent short-circuit. There is a need to planarize the stacks to avoid the deposited CBRAM layers from falling off from the edges of the TE and BE stacks. This is achieved using a combination of spin on glass (SOG) named NDG 7000R by Desert Silicon, Inc. and LOR 3A by Kayaku Advanced Materials, Inc. Ideally just the SOG would have sufficed as a planarizer but due to two limitations the LOR 3A had to be used in conjunction with the SOG. Firstly, the SOG needs to be spun at a spin speed <2000RPM to avoid severe non-uniformities in the top layer of SOG thus limiting the maximum possible height per SOG layer. Secondly, no more than 2 layers of SOG can be stacked on top of each other as the baking process for the third layer cracks the SOG layers underneath. Consequently, according to the film thickness versus spin speed curve in and taking into account film shrinkage after curing, the total thickness of the dual layer SOG spun at 2000 RPM is limited to about 1100 nm. This is nowhere near enough to planarize 740 nm of topography. Thus, the LOR 3A resist, being an excellent planarizer can be stacked up to several thousands of nm in thickness without cracking. So, it is used as a buffer between the SOG and the wafer to allow for a taller planarization layer.

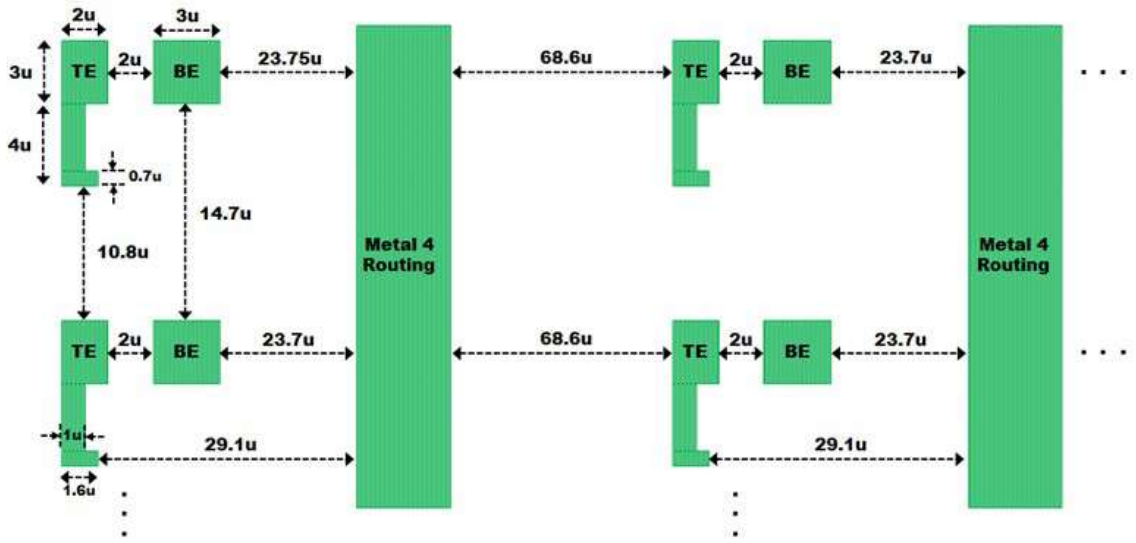


Fig. 131. Top view of a 2x2 subset of the 8x7 CBRAM array outlining the TE, BE and the metal routing.

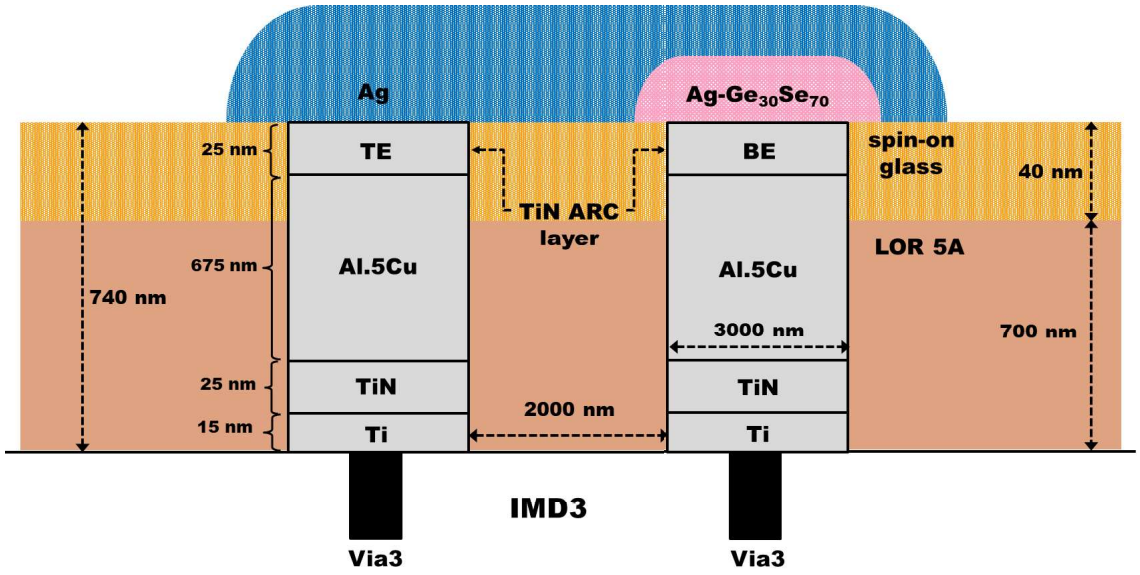


Fig. 132. Cross-section of one CBRAM and its electrodes.

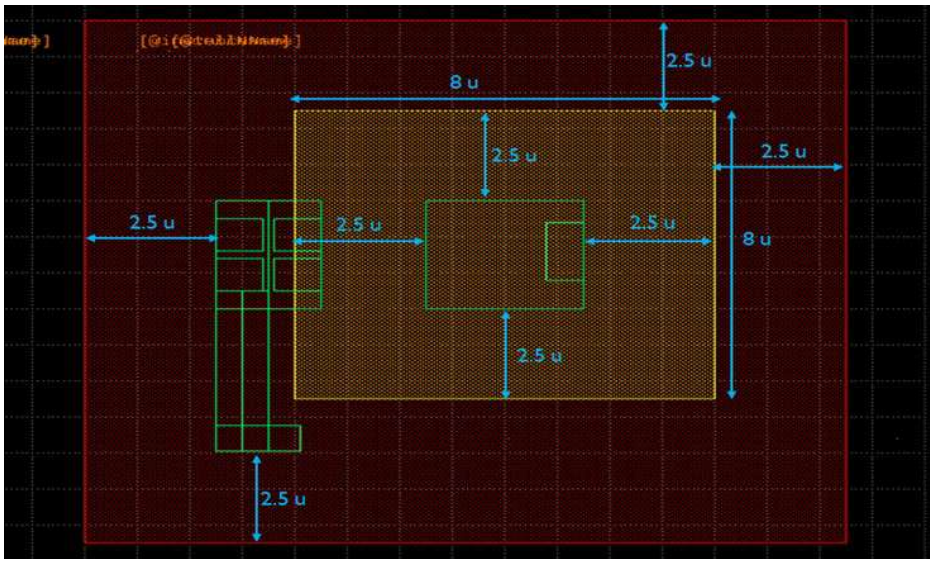
CBRAM Layout in the Neuromorphic CMOS Chips

This work requires the deposition of only two CBRAM layers- *Ag* anode and *Ag-GeSe*$_3$ electrolyte. Hence, only two masks are needed for patterning as shown in Fig. 133(a). In the same figure, the green outlines represent the TE and BE electrode outlines. The first mask outlined by a yellow square in Fig. 133(a) patterns the *GeSe*$_3$ chalcogenide layer and a thin *Ag* photo dopant layer on top of the chalcogenide. The second mask outlined by a red square in

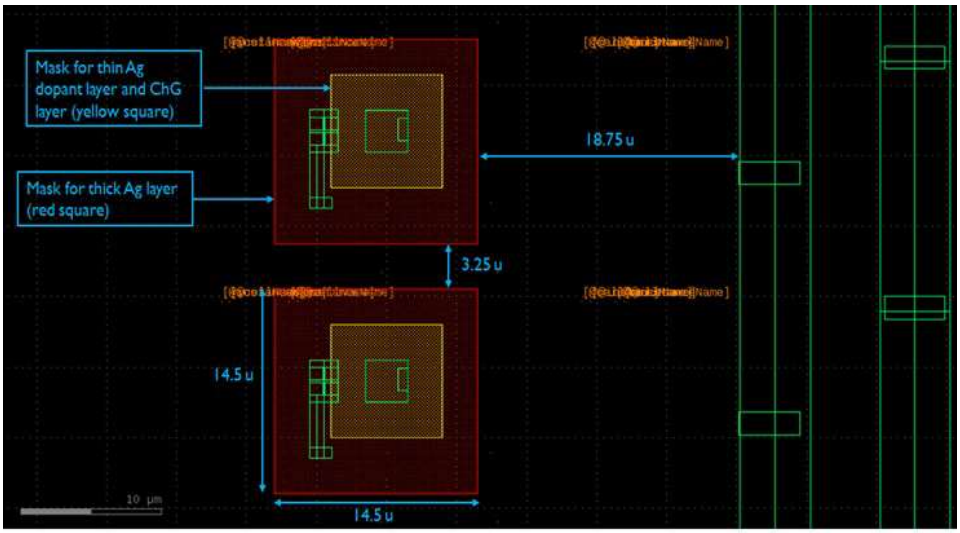
Fig. 133(a) patterns the *Ag* anode layer. After the completion of deposition, the latter diffuses into the chalcogenide through a process called photo-dissolution. Photo-dissolution involves the use of light energy greater than the optical gap of the glass to create charged defects at the interface of the thin Ag photo doping layer and the $GeSe_3$ chalcogenide. The electrons so created diffuse into the $GeSe_3$ layer while the holes are trapped by Ag. The resulting built-in electric field is enough to force the Ag^+ ions across the interface energy barrier into the chalcogenide. Thus, an Ag^+ rich interfacial monolayer is created that supplies metal ions deep into the bulk of $GeSe_3$ through drift and diffusion.

The masks were designed to fulfill three conditions: - the *Ag-$GeSe_3$* mask must completely cover the BE so that the Ag anode layer does not come in contact with it and short circuit the device, the *Ag* anode mask must completely cover the TE and the *Ag-$GeSe_3$* layer and since the aligner to be used for patterning the PR in the device deposition process flow (a lift-off process is used for depositing the CBRAM layers) is manually operated there is a predicted margin for error of ±2.5 µm. This means that the exposure of the PR can be off by ±2.5 µm. This means that the final layers can be misaligned by a maximum of ±2.5 µm after lift-off. Considering all the three conditions, the *Ag-$GeSe_3$* mask dimensions are made 2.5 µm larger than the BE outline and the *Ag* anode mask dimensions are 2.5 µm away from the TE and *Ag-$GeSe_3$* mask outline, i.e., the left and bottom dimensions of Ag anode mask are 2.5 µm away from TE outline and the right and top anode mask dimensions of the same are 2.5 µm away from the Ag-$GeSe_3$ mask outline as shown in Fig. 66 (a) and (b). Under these design rules, the *Ag-$GeSe_3$* mask is 8x8 μm^2 and the *Ag* anode mask is 14.5x14.5 μm^2 in area. Fig. 66 (c) shows the 10x10 sets of die on the coupon as designed in the mask with each die area being 2138x2248 μm^2 and the inter die distance being 90 µm.

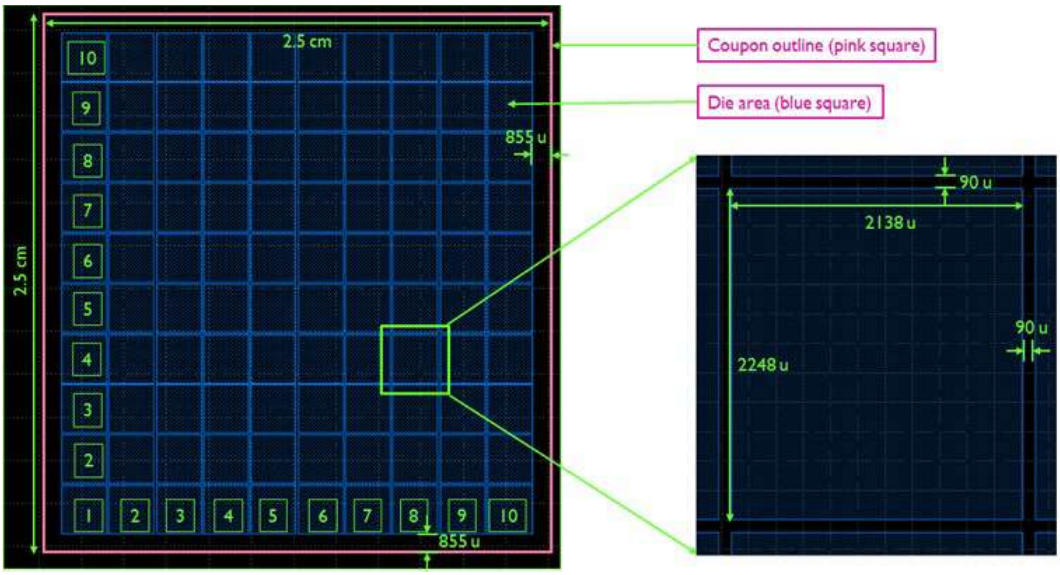
Since the margin of error (±2.5 µm) is larger than the separation between TE and BE (2 µm), one obvious issue that emerges is the overlap between the *Ag-$GeSe_3$* layer and the TE which should not be a problem so long as the *Ag-$GeSe_3$* layer does not completely cover the TE. This means that in the worst scenario, the final deposited *Ag-$GeSe_3$* layer cannot exceed its dimensions by 1.5 µm towards the left or bottom sides than what is outlined by its mask. Fig. 4.7 (b) shows the side-by-side comparison of masks for 2 sets of TE and BE stacks. Here, the inter-mask spacing between the *Ag* anode layers of two consecutive stacks is only 3.25 µm and thus care must be taken to prevent overlap. There are 8x7 such masks each for the *Ag* anode and the *Ag-$GeSe_3$* layers per die.


2.5 u
8 u
2.5 u
2.5 u
2.5 u
2.5 u
2.5 u
8 u
2.5 u
2.5 u

(a)


Mask for thin Ag dopant layer and ChG layer (yellow square)
Mask for thick Ag layer (red square)
18.75 u
3.25 u
14.5 u
14.5 u
10 µm

(b)



(c)

Fig. 133. Photolithography mask layouts for CBRAM: (a) mask layout for the Ag anode layer (red square), thin Ag dopant layer (yellow square), and the $GeSe_3$ layer (yellow square); (b) mask dimensions with respect to the metal routing; (c) coupon layout with the scribe, die and coupon dimensions.

Planarization

The processing was carried out on 1-inch by 1-inch square coupons diced from round MG wafers with a 6-inch diameter. MG wafers are the lowest grade wafers available and are only used for testing purposes. As such, these MG wafers do not have any FEOL CMOS and lower-level metal; only the top level. The processing work was carried out at the Nanofabrication facility of Arizona State University. The process flow steps for the planarization recipe is discussed below.

1 LOR 3A and SOG process flow

i. Wafer scribing and dicing

The 8-inch mechanical wafers were first scribed into 4 quarters before dicing into coupons. This is because the dicer available can only dice a wafer with a maximum diameter of 4 inches. Upon dicing, 8 1-inch by 1-inch coupons were obtained from each quarter. So, for one 6-inch wafer, 32 coupons can be obtained.

ii. Wafer cleaning and treatment

Ideally, an acid-based cleaning technique like Piranha clean or RCA clean is best for removing any metal ions or organic particles. But since the coupons contain metal layers on them as part of the TE and BE, acids would corrode the metal. Thus, a simpler clean is performed using Acetone, Methanol, and Isopropyl alcohol (AMI). Since the cleanroom has high humidity

(>40%), a dehydration bake is performed over a hot plate immediately before spinning at a temperature of 140°C for 10 minutes. The substrate is cooled to room temperature before starting the spin process and takes between 3-5 minutes for the 1-inch by 1-inch Si coupons. For the SOG, the cooling process should occur in an environment with low humidity (<40%) due to the hygroscopic nature of the wafer. It is also recommended that the spin cup humidity be maintained below 40% for the same reason, i.e., moisture adsorbed by the wafer surface can prevent the SOG from sticking properly to the wafer surface. The LOR 3A exhibits good adhesion to Si substrates and thus the cooling environment need not be a low humidity environment.

2 LOR 3A Spin Casting

The final coating profile after spinning and baking is influenced by many factors: - material properties like planarizing liquid's surface tension, viscosity, solvent evaporation rate; shrinkage due to high-temperature curing and defects on wafer are all key in determining whether the SOG or LOR process fills gaps and gives a planar surface. Similarly, spin speed, settling time, spin acceleration, the manner of dispensing, and the ambient during the coating process determine the quality and thickness of the coat on the substrate. But, unlike the SOG, the LOR is not a very difficult material to process and is designed to give very low defect coatings in addition to providing exceptional planarity.

As shown in Fig. 134/135, the overall LOR and SOG process flow has 6 main steps (not considering the wafer cleaning and treatment): - LOR 3A spinning, LOR 3A baking, LOR 3A dry etching, SOG spinning, SOG baking and SOG dry etching. According to the thickness versus spin speed curve, the LOR 3A film is about 225 nm thick when spun at 3000 RPM. Note that the film height is of no concern here if the LOR film is thick enough to reduce the step coverage to about 100 nm across the TE and BE. Accordingly, a technique is employed to increase the LOR settling time. After spinning each layer of LOR 3A, the substrate is left on the spin chuck for 5 minutes to allow for the spun LOR layer to 'settle'. During this time, the LOR 3A layer is still fluidic in nature due to the presence of solvents and is thus vulnerable to external factors. One such factor is the cleanroom airflow that can ruin the planarity of the spun resist layer. To avoid this, the substrate is covered with a beaker for the 5 minutes of 'settling' time.

It was found that 15 layers of LOR 3A (each layer is spun at 3000 RPM to a height of 225 nm) can be stacked on top of one another when individual films are baked at a constant high temperature. Any attempts to spin subsequent layers crack the layers at the bottom. But, if ramped baking (ramping the temperature up in steps over time) is used, even 20 layers can be stacked without cracking the film. Since this recipe uses ramped baking 18 stacked layers were spun (total predicted height is 4050 nm). But spinning such a huge number of layers is a time-consuming procedure, approximately 20 minutes per layer with a total of 6 hours for all 18 layers. It is worth noting that the LOR 10A (each layer is 1100 nm thick at 3000 RPM) by Kayaku Advanced Chemicals was considered as a substitute to lessen processing time but the results were highly unfavorable. The LOR 10A film cracked after the 9^{th} layer (total film thickness was about 10 μm) and despite the high thickness the film non-uniformity across the

TE and BE stacks was still about 200 nm (twice of that for LOR 3A). This proves that just increasing the total planarizer height thickness does not give better planarity; the number of layers also plays an important role.

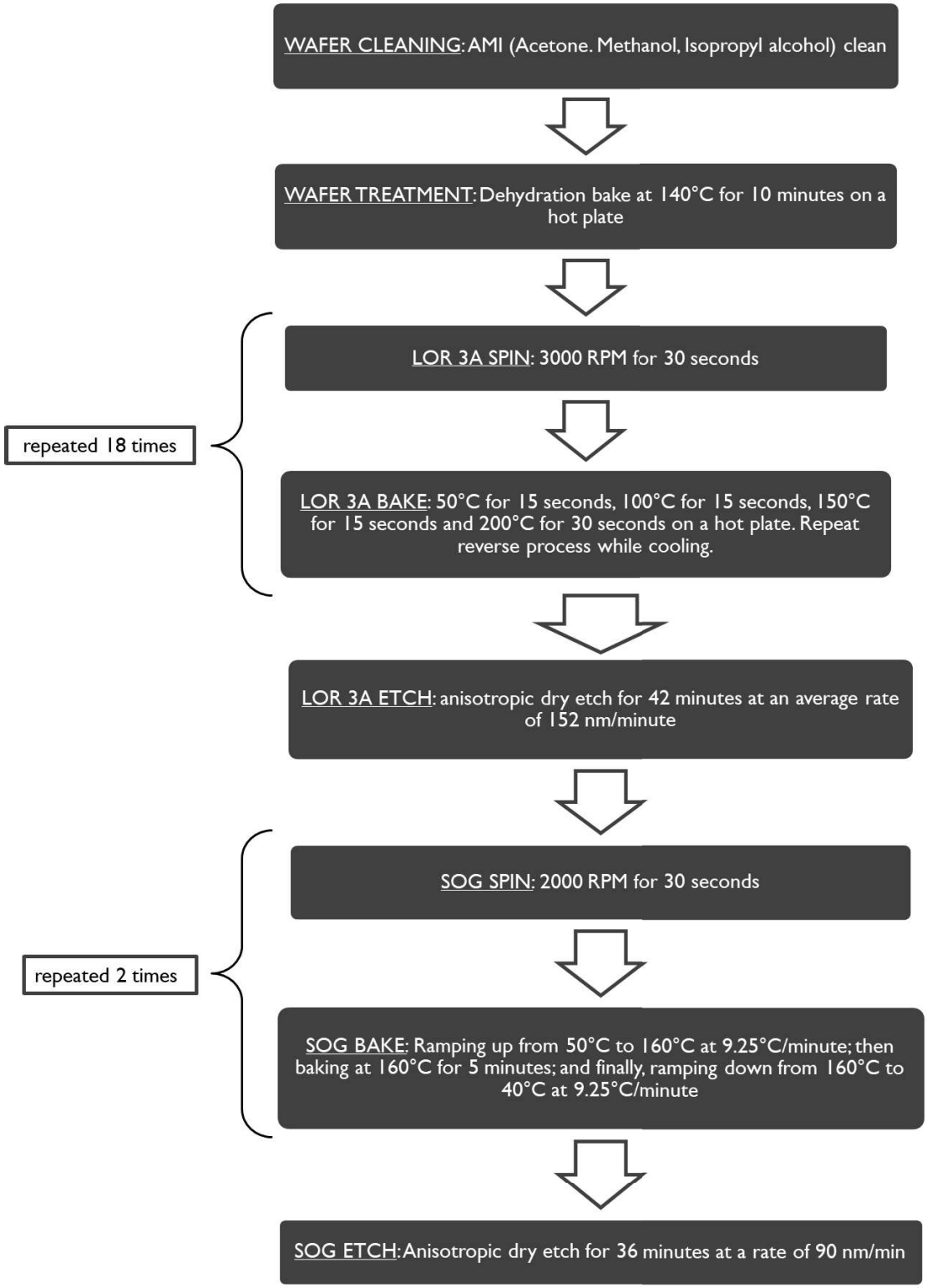


Fig. 134. SOG and LOR planarization process flow on MG Si wafers.

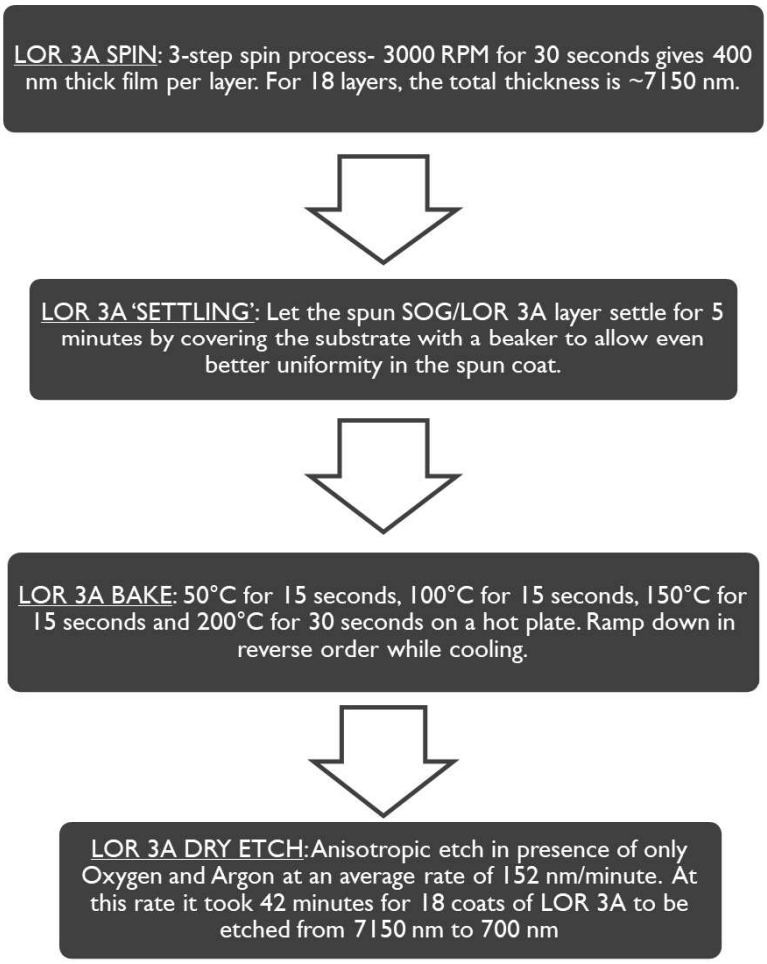


Fig. 135. LOR 3A process flow.

As shown in Table XX, the LOR 3A spin recipe is a 4-step process: - the dispense step, the spread/flow-in step, the spin-off step and finally, a decelerating step. The dispense step as the name suggests is when the LOR 3A is dispensed onto the substrate. It can be dynamic where the LOR 3A is dispensed while the substrate spins at a low speed of 100-300 RPM or static where the substrate does not spin. The spin coater tool used in this work does not allow dynamic dispensing. It should be noted that dynamic dispense gives more uniform and planar film for patterned wafers with 125mm or larger diameter whereas static dispense produces higher quality films on smaller wafers. This is because static dispense would require larger dispense time on larger wafers and PR sitting on a static wafer for long periods results in solvent evaporation causing spin defects. The substrates used in this work being 1-inch by 1-inch, static dispense works better. The spread/flow-in step causes the dispensed LOR 3A to spread over the entire wafer evenly by pushing the excess LOR 3A to the edges. The flow-in

step is short (5-7 seconds) and slow (500-1000 RPM/second) acting as a transition between dispense and spin-off. The spin-off step allows the excess resist to be spun off the edge of the substrate leaving a more planar layer of desired thickness. This step runs at a much higher speed (determined by the film thickness versus spin speed chart in [12]) and for a longer time than the previous two steps. Finally, the deceleration step is an intermediate step before the spin process comes to a halt. A solvent-rich environment is needed in the spin cabinet for good film uniformity and to prevent the spun layer from drying too fast. Squirting about half a pipette's worth of LOR 3A into the cabinet right before spinning is adequate.

Table XX
LOR 3A spin recipe

Steps	Spin Speed (RPM)	Ramp rate (RPM/second)	Time (seconds)	Exhaust
Dispense	0	0	< 20	-
Spread/Flow in	500	500	5	5%
Spin-off	3000	500	30	80%
Deceleration	300	500	7	80%

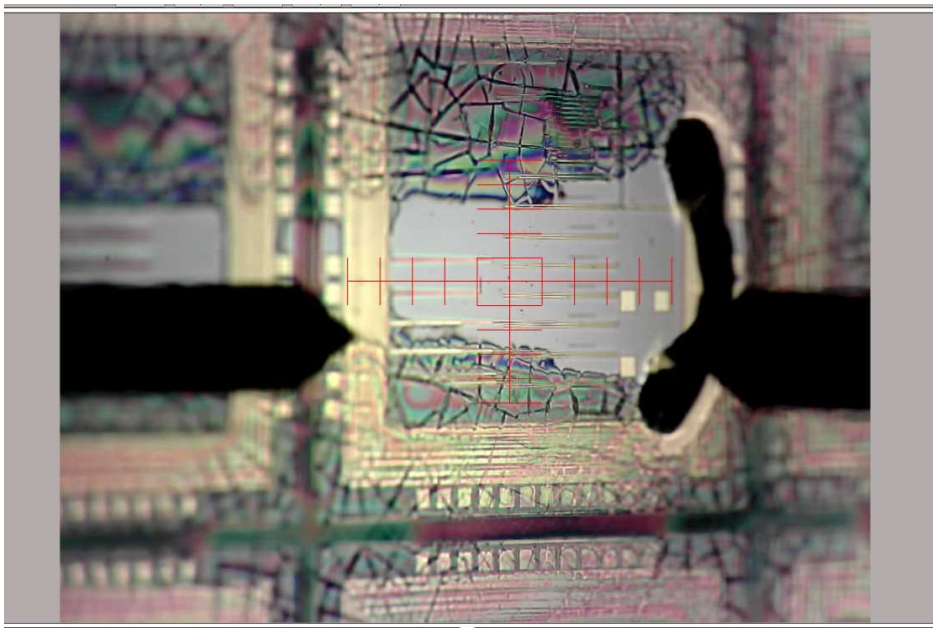
Another step is usually included between the spin-off and deceleration steps to push away the edge beads off the edge or at least as close to the coupon edge as possible. This recipe does not use one because edge beads are not of primary concern. Besides since a square coupon is used it is impossible to eliminate edge beads as there will always be some left at the square corners. Edge bead formation is a build-up of spun liquid along the outer edge of a wafer caused by surface tension during spin coating. In this step, the substrate is spun at speeds much higher than spin-off (~6000-10,000 RPM) for a short time (~5-7 seconds). Alternatively, an Edge Bead Remover (EBR) solvent can be used to manually strip off any edge beads.

3 LOR 3A Post-spin bake

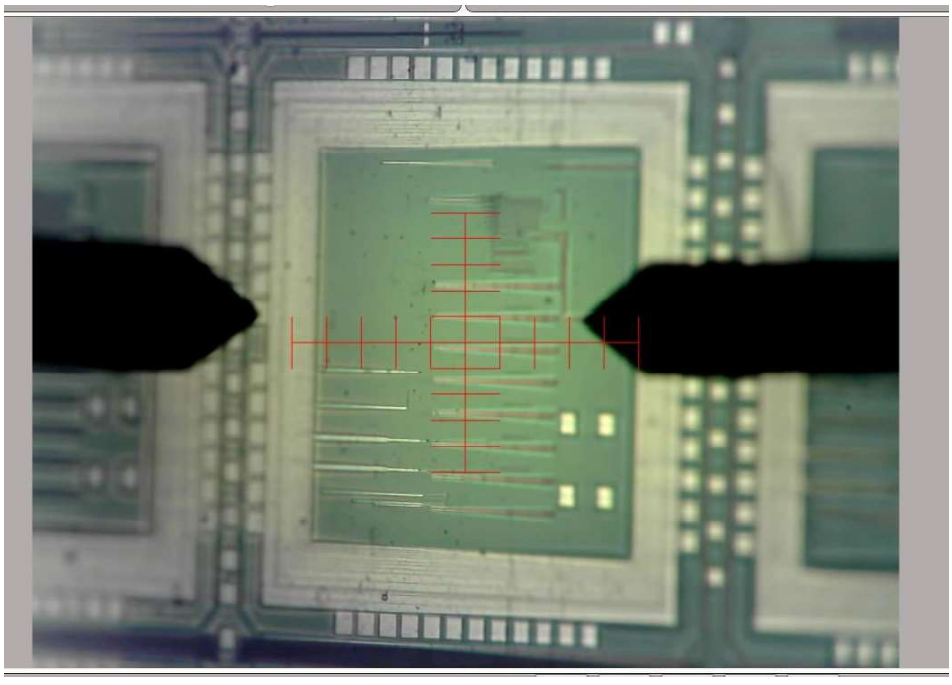
Owing to the propensity of cracking after a high number of spun-on layers, a ramped bake is performed for LOR 3A where the temperature was ramped from 50°C to 200°C in 4 steps and again back down to 50°C in 3 steps. The steps are: - after letting the substrate 'settle' for 5 minutes it was first subjected to 50°C for 15 seconds, then 100°C for 15 seconds, followed by 150°C for 15 seconds and finally, the target temperature of 200°C for 30 seconds. The steps were retraced in the same manner for cooling: - 150°C for 15 seconds, 100°C for 15 seconds and finally, 50°C for 15 seconds. The reason for choosing a target temperature of 200°C is that it renders the LOR 3A optically undevelopable. This means that any subsequent exposure and resist development process steps as part of CBRAM deposition will not affect the film stack in any way.

This method of baking enabled a completely crack-free film stack. This is because the lower the LOR 3A layer sits in the film stack, the greater number of times it gets subjected to baking; the bottommost layer gets baked 18 times, the layer above it gets baked 17 times and so on. The heat travels through the bottom layers every time to be able to bake the topmost layer subjecting them to unnecessary stress. To make matters worse, it is imperative that the bottom

2-3 layers of LOR 3A remain in perfect condition because the top 15-16 LOR 3A layers are to be etched away and only serve to provide the desired planarity to the topography. Fig. 136(a) shows a LOR film that cracked and ‘peeled-off’ after etching due to stress when the films were baked at a constant high temperature whereas the film in Fig. 136(b) is completely crack-free due to ramped baking.



(a)



(b)

Fig. 136. 18-layer LOR 3A film stack when baked at (a) constant temperature of 200°C and (b) ramped temperatures.

It was found that each LOR layer was about 400 nm thick after baking instead of the predicted 225 nm. This is because of the 'settling' process, i.e., as some of the resist liquid from the thicker edge beads flow inward while 'settling' thereby increasing the film thickness. Thus, the total thickness of the 18-layer film stack is 7150 nm as shown in Fig. 137.

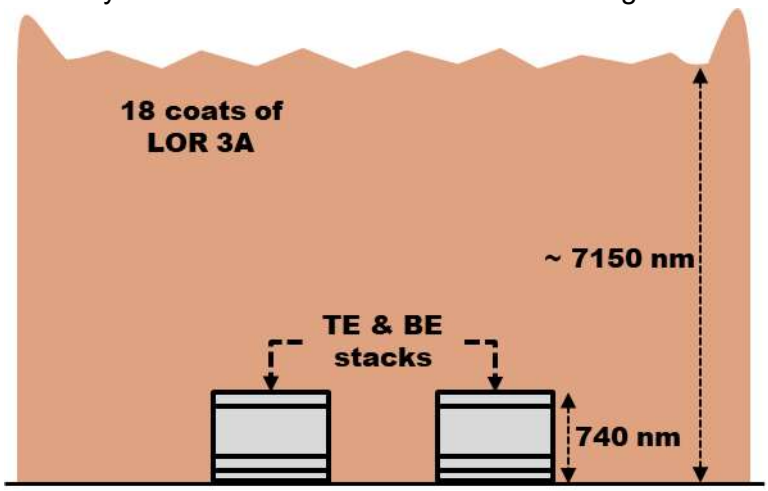


Fig. 137. LOR 3A film stack after baking

It should be noted that even though ramped baking in discrete steps of temperature increment works fine, the best way is to continuously increase the temperature slowly from 50°C to 200°C and back down to 50°C again. But since it would take an enormous amount of time to perform the bake for 18 layers, this is not done.

4 LOR 3A etch-back

After baking, the 7150 nm thick LOR 3A film stack was etched for 42 minutes in presence of oxygen and argon gases to bring the thickness down to 700 nm as shown in Fig. 138. Now, the etch time of 42 minutes means that the etch rate is 152 nm/minute. But it was found that this etch rate varies between 150 minutes to 165 minutes and thus a fixed etch time does not exist. The best way is to perform the etch in 2 steps. Firstly, perform an initial etch for a 'safe' time such that over etching is impossible, calculate the etch rate and measure the thickness of the LOR 3A left. Secondly, perform a final etch-back based on the time needed to bring down the thickness to 700 nm at the calculated etch rate. The etch time can differ between etching tool types and etch chamber conditions. The etch time is 'safe' when it is minimum due to the etch rate being maximum. Thus, the 'safe' time assuming a maximum etch rate of 165 minutes is 39 minutes for the Plasmatherm 790 Reactive Ion Etch (RIE) tool in the ASU NanoFab. This is an anisotropic etch meaning that the LOR 3A etches only in the direction of the gas flow. Other parameters of importance are the chamber pressure of 11 mTorr, dc bias of 515 V and the

forward and reflected powers of 175 W and 3W, respectively. Table XXI shows the entire etch recipe.

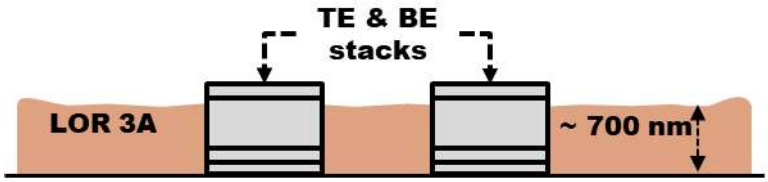


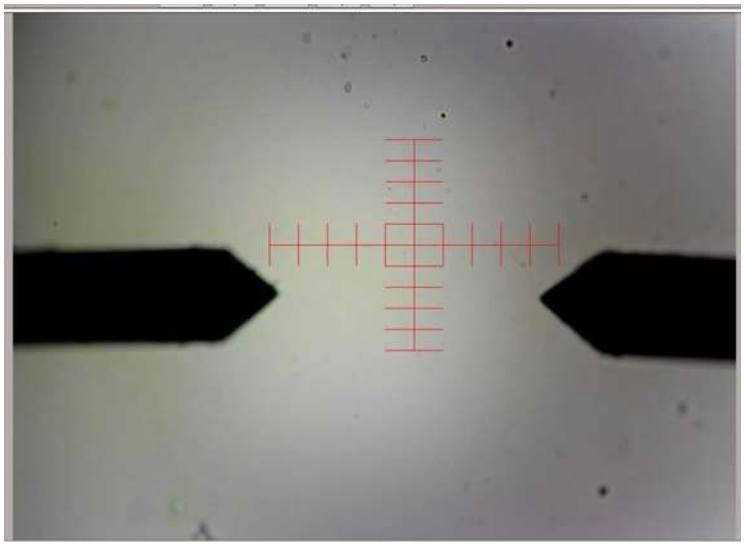
Fig. 138. LOR 3A film stack after dry etching

Table XXI
LOR 3A dry etch process parameters

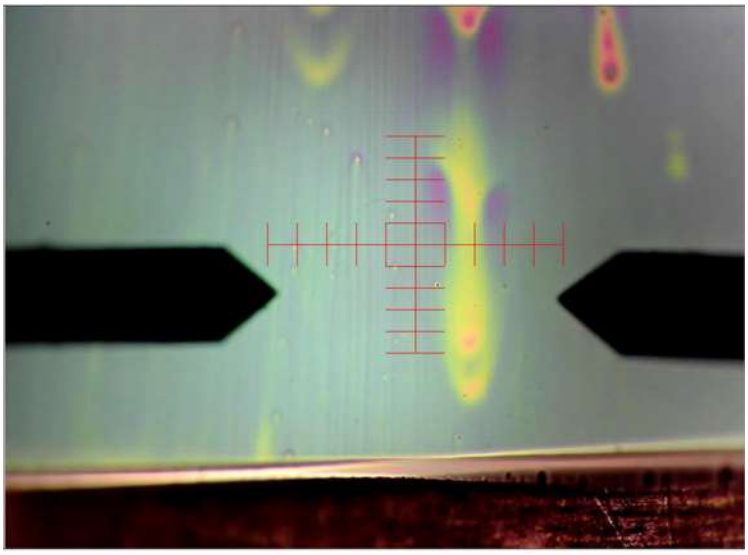
Gas	**Oxygen (O_2)**	**Argon (Ar)**
Gas flow rate	8	8
Etch time (minutes)	42	
Etch rate (nm/minute)	152	
Forward Power (W)	175	
Reflected Power (W)	3	
DC Bias (V)	515	
Chamber Pressure (mTorr)	11	

5 SOG Spin Casting

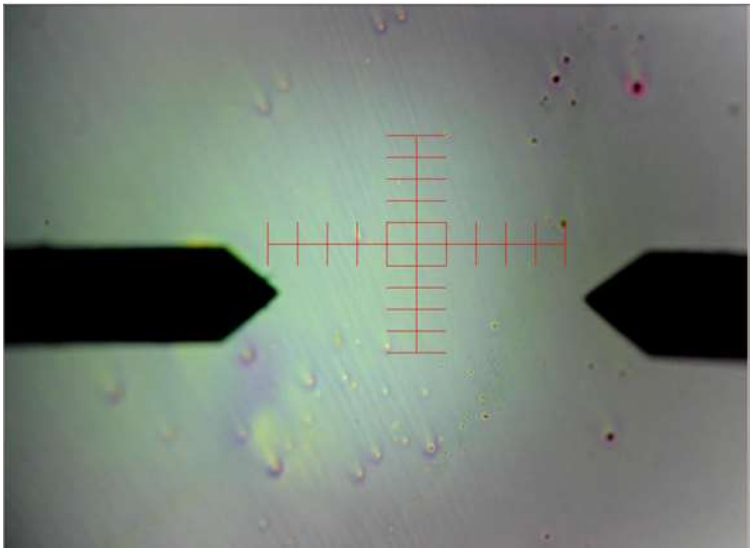
The SOG is a very difficult material to process due to several issues like film cracking, defects, film hazing, etc. Thus, before describing the spin process, certain precautions are to be followed for best SOG coats. It is recommended to store the SOG at 4°C to prolong its shelf life. If done so, the SOG needs to be warmed to room temperature before spinning it onto substrates. This is to be done by allowing the SOG to settle for about 24 hours at room temperature. Cold SOG can give poor quality films with high particle levels and form haze on the spun film thereby making the film cloudy as shown in Fig. 139 (a).



(a)



(b)



(c)

Fig. 139. Image of a SOG film under a microscope: (a) a cloudy SOG film due to high humidity, (b) and (c) particles on a substrate causing the SOG to accumulate around the particles while spinning giving a comet-like appearance.

The SOG can be rinsed and removed by isopropyl alcohol (IPA), denatured alcohol or ethyl acetate if the SOG solvent has not completely evaporated. But, once the solvent evaporates and the SOG layer sets in, only hydrofluoric acid (*HF*) or Buffered Oxide Etchant (BOE) can etch it. And the SOG dries very fast so all equipment that is prone to exposure to the SOG must be made from polyvinyl or polypropylene as these materials wet well with rinsing solvents. Good wetting ensures complete removal of residual SOG thereby preventing them from being a source of particles in future processing work. In this process, IPA was used for rinsing and it worked well given that the equipment was rinsed immediately after spinning.
A plastic pipette was used for dispensing the SOG onto the substrate. The dispensing should be done immediately after taking the pipette out of the SOG bottle as the solvent dries very fast and can cause particles to form within the pipette. These particles would then transfer to the SOG layer on the substrate while dispensing. The same can happen if the pipette tip touches the SOG layer while dispensing and so sufficient distance must be maintained. Similarly, to avoid the transfer of the same particles to the SOG solution in the bottle, the pipette is to be disposed of after using only once and along with the SOG that may remain unused in the pipette. Fig. 72 (b) and (c) show the effect of particles on the spun and dried SOG layer.

At least 80% of the substrate must be covered with SOG while dispensing so that the spun layer gives uniform coverage over the substrate. It is recommended to complete the dispensing within 5 seconds and start the spin process immediately after. A longer dispense time can cause the layer to be thicker at the substrate center due to solvent evaporation.

Note that the 700 nm of the topography is covered by LOR 3A while only 40 nm is left to be planarized by the SOG. The reason for this is two-fold. Firstly, the SOG can only be stacked up to 2 layers for a maximum total height of 2200 nm at a spin speed of 2000 RPM before the film cracks. And secondly, SOG layers are prone to defects that come from the material itself and thus are impossible to eliminate meaning that the more the layers are stacked, the more is the defect count in the film stack. Unlike LOR 3A spin casting, a 3-step spin recipe is used here as shown in Table 5.3. The spin-off step spin speed was chosen to be 2000 RPM because going any higher would only decrease total film thickness thereby increasing non-uniformity due to bad step coverage and going any lower also decreases film planarity. Any significant uniformity variation is easily visible to the naked eye in the form of color variation along the plane of the film as shown in Fig. 140.

Table XXII
SOG spin recipe

Steps	**Spin Speed (RPM)**	**Ramp rate (RPM/seconds)**	**Time (seconds)**	**Exhaust**
Dispense	0	0	< 10	-
Spread/Flow in	500	500	5	5%
Spin-off	2000	1000	30	80%

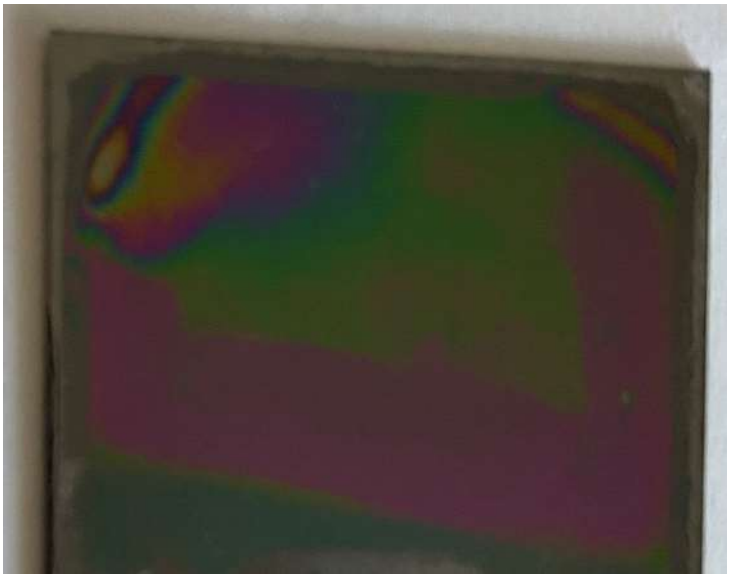


Fig. 140. Visible color variation in SOG film due to non-uniformity.

Even though we predict the SOG film thickness to be 1100 nm per layer at 2000 RPM, the final SOG layer is somewhat higher, 1225 nm per layer. Again, similar to the SOG, this is because of the ‘settling’ process as some of the resist liquid from the thicker edge beads flow inward while ‘settling’ thereby increasing the film thickness. The spin time depends on the humidity level of

the cleanroom. If the cleanroom humidity is less than 35%, a spin time of 30 seconds is sufficient. Otherwise, a higher spin time between 30 seconds to 1 minute is required to allow for more solvent evaporation while spinning. This is because if the environment is too humid, the SOG absorbs water while performing post-spin bake and the film becomes cloudy.

The SOG tends to form a mist of liquid droplets above the substrate while spinning which produces particles or bubble-like defects in the films. Sufficiently high exhaust can evacuate the solvent vapor and keep the mist level below the plane of the wafer surface. Besides this, it was noticed that using insufficient exhaust causes non-uniformity in the spun film. Thus, the exhaust must be turned on and maximized during all process steps. In this work, the exhaust was kept at 80% because that was the maximum allowable limit. Besides the exhaust, a solvent rich environment is also needed within the spin cabinet for film uniformity and to prevent the SOG layer from drying too fast. For best results, squirting about one pipette's worth of SOG into the cabinet is recommended. But, since the SOG can be costly using one of the component casting solvents of the SOG instead would suffice. Accordingly, in this process, IPA was sprayed in the cabinet as it is one of the casting solvents of the Desert Silicon NDG-7000R SOG and is readily available in a cleanroom. Also, while dispensing with a pipette, care should be taken such that no air bubbles form on the dispensed SOG film otherwise, defects will form in the baked layer in the form of voids. It is recommended to use the highest possible acceleration for the spin step that sets the SOG thickness. Due to being limited by the lithography capabilities of ASU NanoFab, one second is the minimum possible spin speed ramp rate that could be achieved. The entire process flow of SOG processing is outlined in Fig. 141.

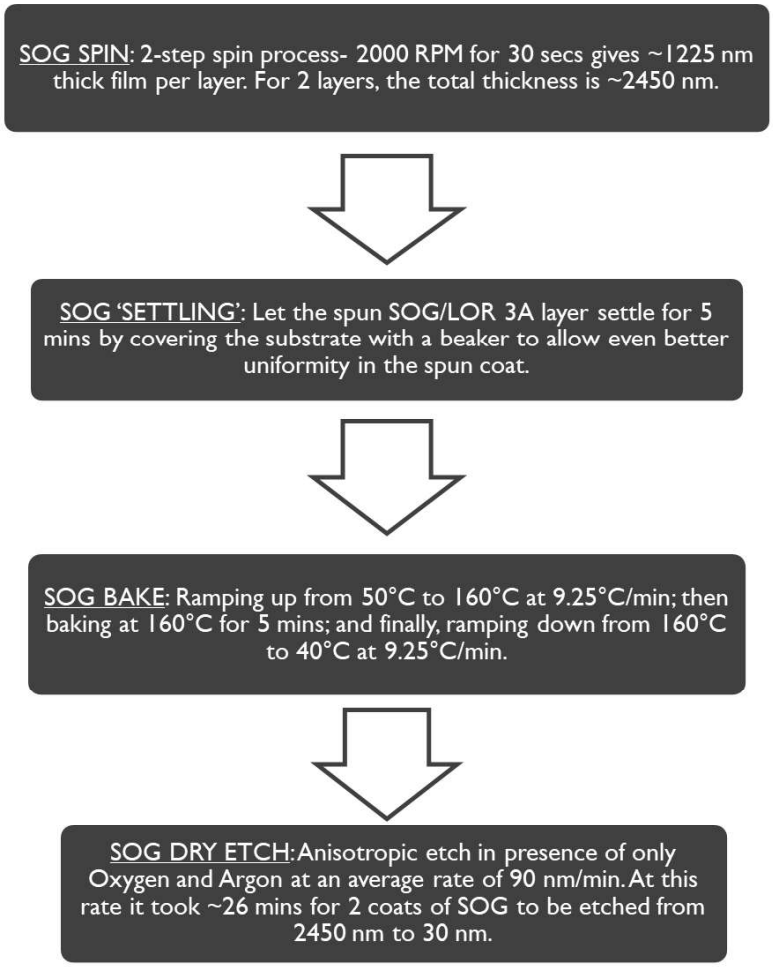


Fig. 141. SOG process flow

6 SOG Post-spin bake

The vendor recommended maximum bake temperature for a double coat of NDG-7000R SOG is 225°C when spun at 2000 RPM. But this SOG is highly sensitive to temperature and may crack at 225°C when spun at lower spin speeds (thicker films). The SOG is also very moisture sensitive and clouds up in highly humid environments (>35%) if not immediately baked (humidity level in the ASU NanoFab cleanroom is >40%). And similar to LOR 3A, the SOG film cracks when subjected to a constant high temperature due to induced film stress. In fact, the situation is even worse as performing a ramped bake in steps is not enough but rather a continuous ramped bake is needed. Despite this, the film still cracks if a third layer is stacked. Considering all these factors, the coupon is placed onto a hot plate immediately after spinning the SOG and the temperature is ramped continuously from 50°C to 160°C at a constant rate of 9.25°C/minute; the substrate then sits at 160°C for 5 minutes before ramping down the temperature at the same rate of 9.25°C/minute to 50°C.

Usually, post-spin baking is followed by curing the SOG film at a very high temperature of around 400°C or higher to remove any remnant solvents and further densify the film. But, since the Desert Silicon NDG-7000R can only withstand a temperature of 225°C before cracking, curing is not feasible. Hence, to maximize solvent evaporation the baking temperature is kept as close to the upper limit as possible given that it does not crack the film. The final double layer SOG coat is around 2450 nm after baking is shown in Fig. 142.

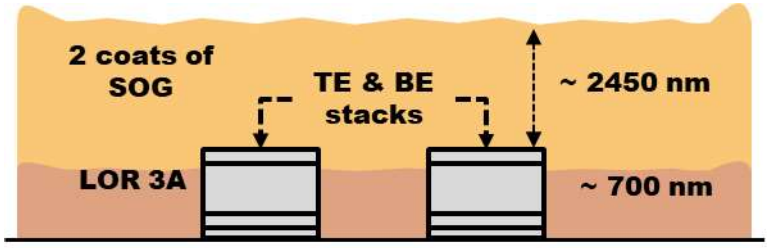


Fig. 142. 2-layer SOG stack

7 SOG etch-back

After baking, the 2450 nm thick LOR 3A film stack was etched anisotropically for 27 minutes in presence of oxygen (O_2) and fluoroform/trifluoro methane (CHF_3) gases to bring the thickness down to 40 nm as shown in Fig. 143. The etch time is 27 minutes which means that the etch rate is around 90 nm/minute. But the etch rate, similar to the situation for LOR 3A, can vary by ±10 minute and so the best way is to first perform the 'safe' etch considering the maximum possible etch rate of 100 minutes for 24 minutes (specific to Plasmatherm 790 RIE tool in the ASU NanoFab), then calculate the new etch rate, and perform the final etch for the time needed to make the SOG 40 nm thick based on the new etch rate. Etch time can differ between etching tool types and etch chamber conditions. Other parameters of importance are the chamber pressure of 41 mTorr, dc bias of 543 V and the forward and reflected powers of 246 W and 1 W, respectively. Table XXIII shows the entire etch recipe.

Table XXIII
SOG dry etch recipe

Gas	**Oxygen (O_2)**	**Fluoroform (CHF_3)**
Gas flow rate	3	40
Etch time (minutes)	27	
Etch rate (nm/minute)	90	
Forward Power (W)	246	
Reflected Power (W)	1	
DC Bias (V)	543	
Chamber Pressure (mTorr)	41	

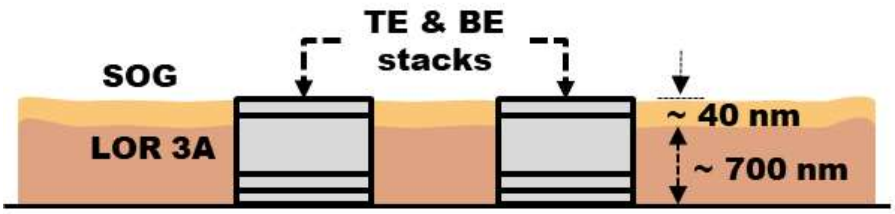


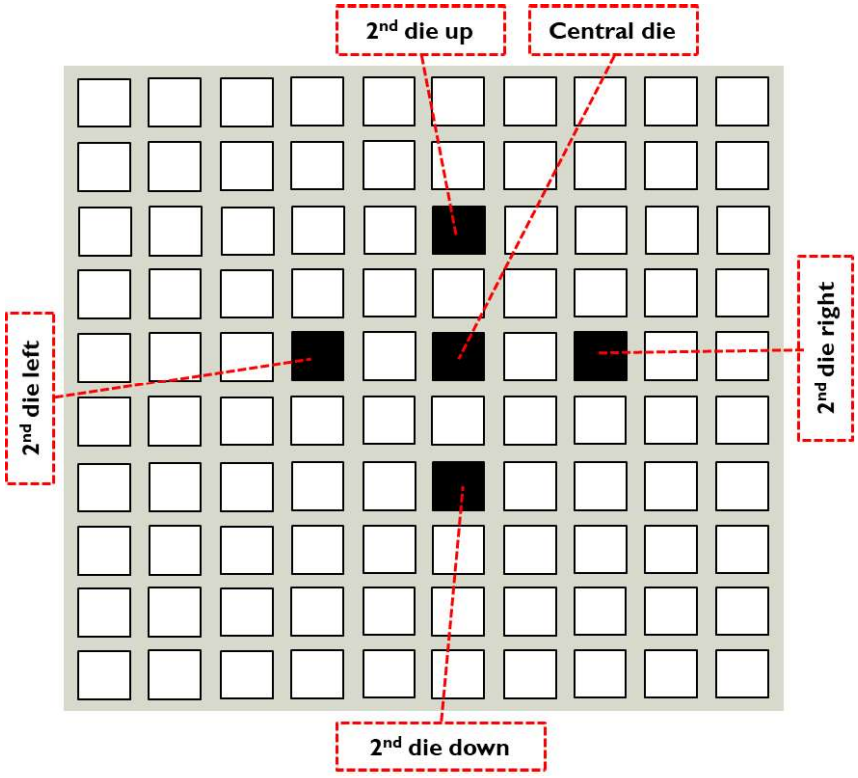
Fig. 143. SOG film stack after dry etching

8 LOR 3A and SOG film characterization

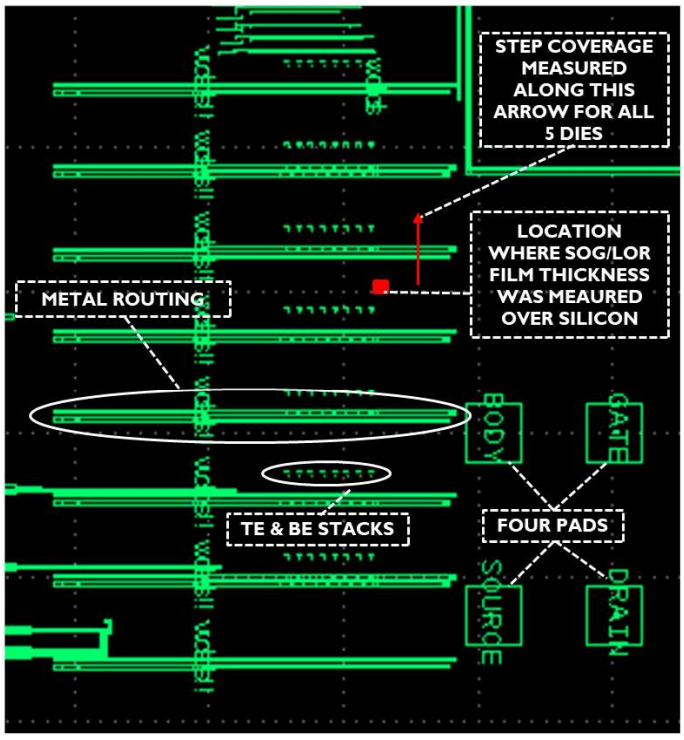
Material characterization is done using profilometry (Bruker Dektak XT Stylus profilometer in ASU NanoFab) and spectroscopic reflectometry (Filmetrics F40 in ASU NanoFab). Profilometer as the name suggests measures the surface profile of films. Here, the profilometer tool Bruker Dektak XT is a contact-based profilometer meaning it has a very sharp stylus that measures the 'hills' and 'valleys' as it runs over the film.

Spectroscopic reflectometry, on the other hand, uses reflection of light waves at film surfaces and interfaces to characterize film thicknesses. When light is incident on a film, some of it is reflected from the surface and some transmit across the film. This same happens to this transmitted light at the interface between this film and another film of a different material sitting underneath. This keeps going until the bottommost layer is reached. Depending on the phase difference between the light reflected from the film surface and the interface between the films, the combined light can create varying levels of constructive and destructive interference. This phase difference depends on the light wavelength and the optical distance the light travels. Finally, a model is fit to the reflected intensity versus light wavelength data to extract the film thickness. Here, the film thickness is varied in the model to minimize the error between the fitted model and measured data. Note that the accuracy of spectroscopic reflectometry deteriorates after more than three layers of different materials are stacked. Spectroscopic ellipsometry is a better option if the optical properties of the film material are known. Even though the optical constants of the LOR 3A were known, that for the SOG were not known and ellipsometry could not be used.

To get accurate and consistent results, the location of measurement on a coupon was kept constant. Out of the 100 dies on a coupon, both film thickness and profile were measured at 5 locations: - the centermost die and 2nd die up, down, left, and right of it as shown in Fig. 144 (a). This is because the dies around the center have the least non-uniformity as will be discussed later in this section. But the die dimensions are in thousands of microns while the profilometer measures film profile accurately over only a few hundreds of microns. Thus, the measurement can only be taken over a small location on a die. This location is near the 4th row from the bottom and 8th column from the left of the TE and BE stacks when the coupon is oriented such that the four pads in any die are at the bottom right corner as shown in Fig. 144 (b).


2nd die up
Central die
2nd die left
2nd die right
2nd die down

(a)


STEP COVERAGE MEASURED ALONG THIS ARROW FOR ALL 5 DIES
LOCATION WHERE SOG/LOR FILM THICKNESS WAS MEAURED OVER SILICON
METAL ROUTING
BODY
GATE
TE & BE STACKS
FOUR PADS
SOURCE
DRAIN

(b)

Fig. 144 (a) dies used for measurement and (b) location of measurement on a die is marked in red color (thickness is measured over the red square and step coverage is measured the red arrow).

To get accurate and consistent results, the location of measurement on a coupon was kept constant. Out of the 100 dies on a coupon, both film thickness and profile were measured at 5 locations: - the centermost die and 2nd die up, down, left, and right of it as shown in Fig. 144 (a). This is because the dies around the center have the least non-uniformity as will be discussed later in this section. But the die dimensions are in thousands of microns while the profilometer measures film profile accurately over only a few hundreds of microns. Thus, the measurement can only be taken over a small location on a die. This location is near the 4th row from the bottom and 8th column from the left of the TE and BE stacks when the coupon is oriented such that the four pads in any die are at the bottom right corner as shown in Fig. 144 (b).

Fig. 145 shows the film thickness measured on the five die locations after post-spin bake or before etching and Fig. 146 shows the same after etching of LOR 3A. Fig. 147 (a) shows the 18-layer LOR 3A film stack profile on the centermost die right after baking or before etching and Fig. 146 (b) shows the step height difference for all 5 dies. Fig. 148 (a) and (b) depicts the same information but after etching. Similar to LOR 3A, the SOG film thickness as measured by Filmetrics F40 before and after etching is shown in Fig. 149 and Fig. 150 respectively. And Fig. 151 shows the 18-layer LOR 3A film stack profile before etching whereas Fig. 152 shows the LOR 3A stack profile right after etching.

Note that a SOG film thickness of 40 nm is quite thin compared to the 700 nm of LOR 3A film sitting underneath it. So, an obvious concern is that the SOG top layer non-uniformity might be larger than 40 nm. In such a case, some areas of the SOG will be completely etched away and the LOR 3A underneath will be exposed. But averaging the non-uniformity values for the 5 dies in Fig. 5.19 (b), the average film non-uniformity comes out to be 22.7 nm. Thus, an extra 18 nm of leeway is left since after etching the SOG film stack it needs to be at least 40 nm thick to cover the TE and BE stacks perfectly. Additionally, if the SOG film non-uniformity was so high that it was completely etched away in certain locations, then the Filmetrics F40 would not have been able to measure the thicknesses correctly. This is because the model fitting in F40 requires that the layer materials and their estimated thicknesses be specified beforehand. If the SOG was completely etched due to high non-uniformity but was specified in the F40 model, the goodness of fit (GOF) of the model would be very low. But despite doing multiple measurements across many locations across the 5 dies and even other dies, the GOF remained above 98% consistently.

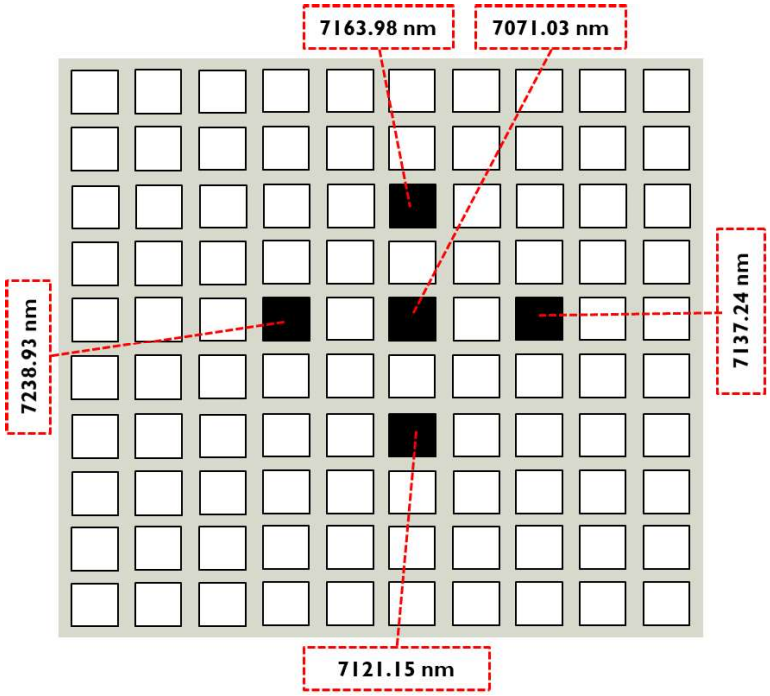


Fig. 145. LOR 3A film thicknesses on 5 dies before etching.

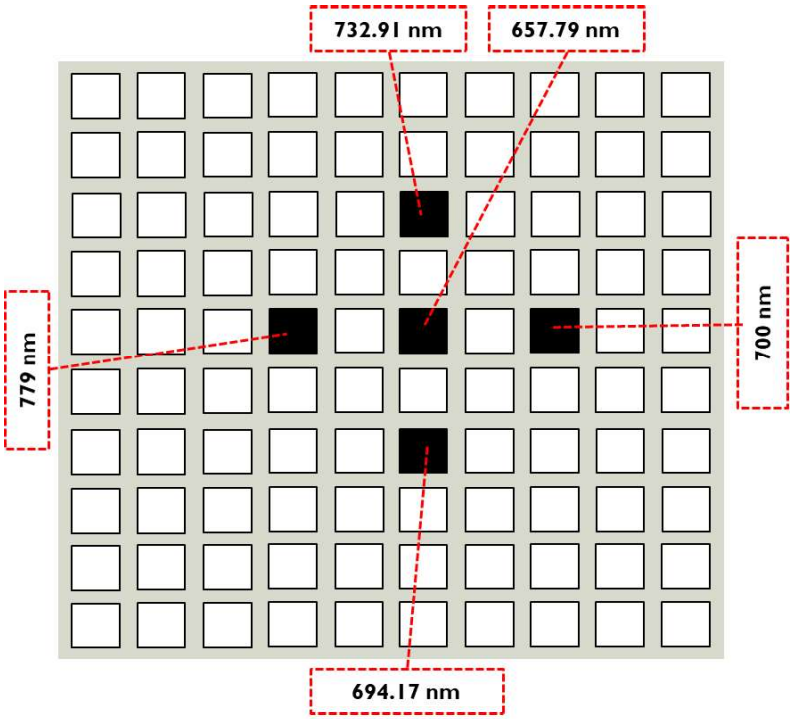
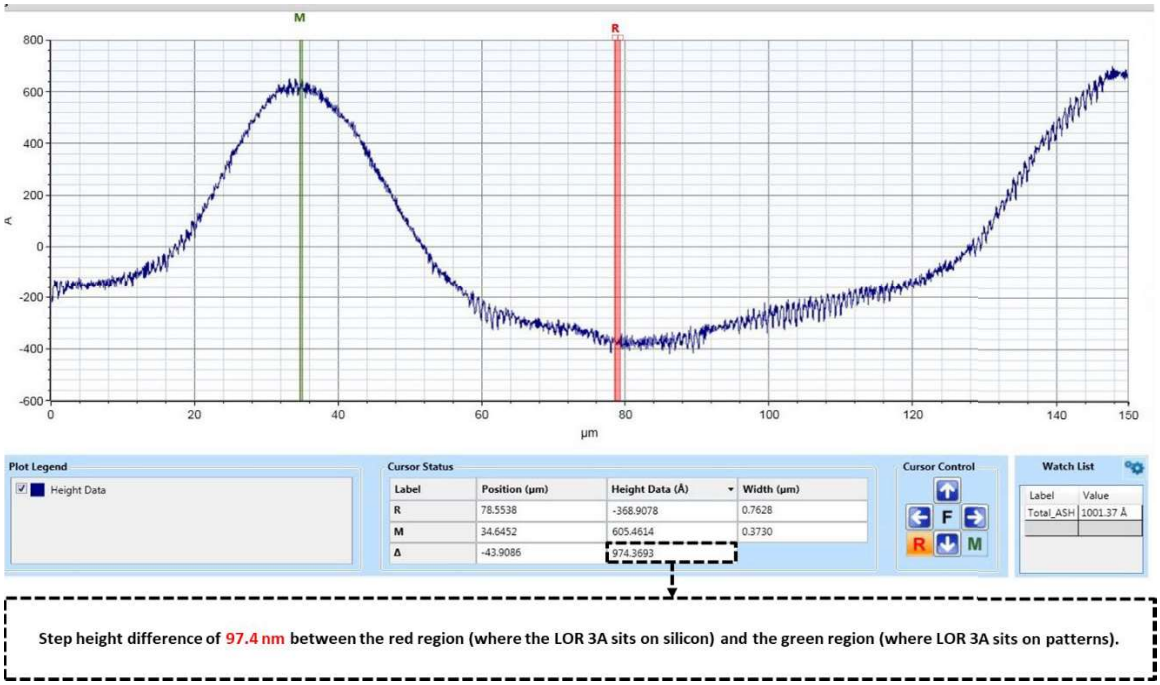
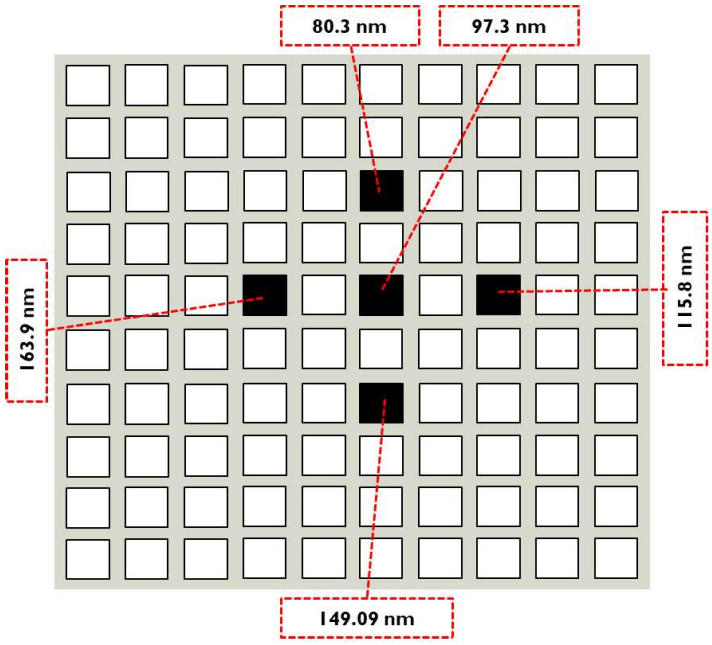


Fig. 146. LOR 3A film thicknesses on 5 dies after etching.

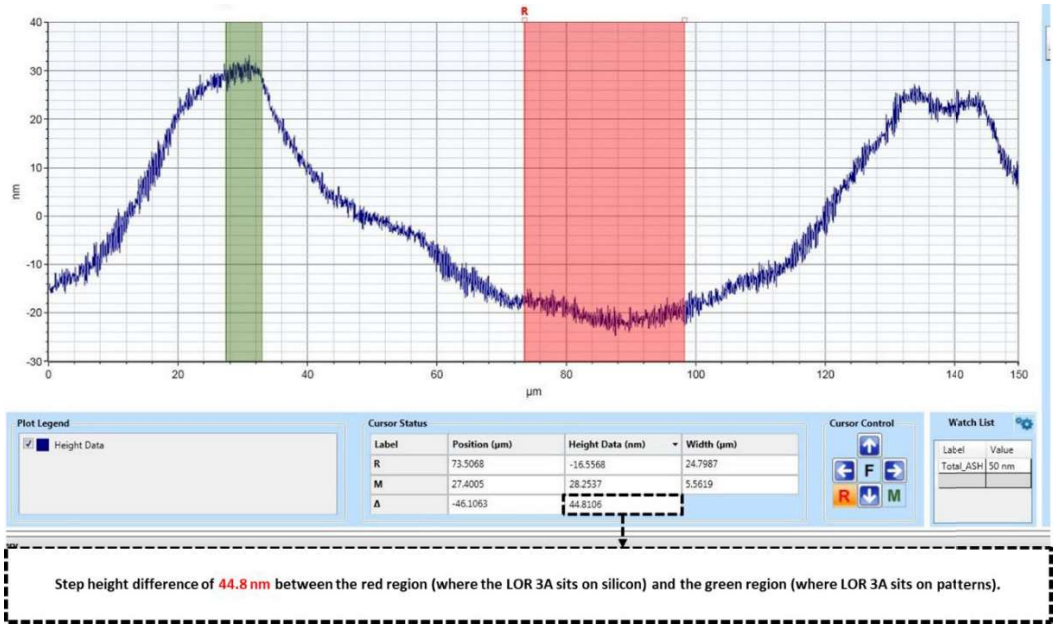


(a)

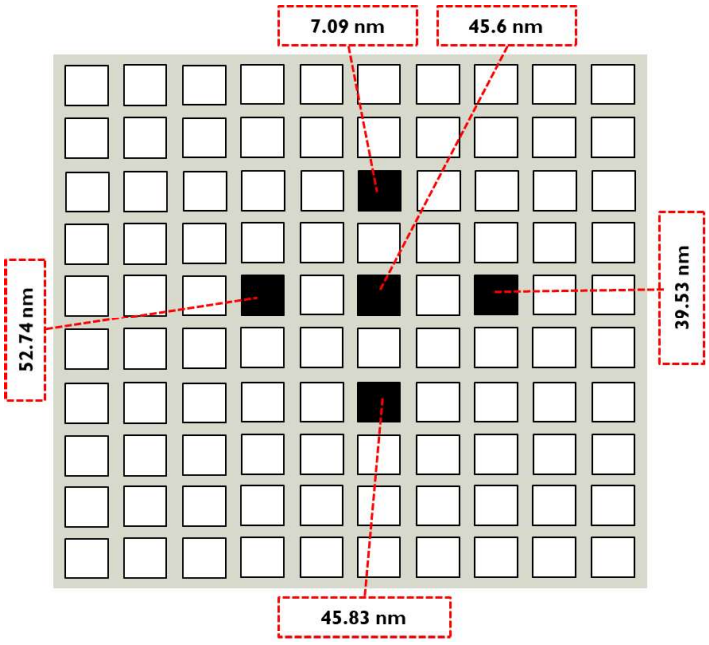


(b)

Fig. 147. LOR 3A film profile before etching on the (a) centermost die, (b) LOR 3A film profile for all 5 dies- centermost die and 2nd dies to the right, left, above and below it.



(a)



(b)

Fig. 148. LOR 3A film profile after etching on the (a) centermost die, (b) LOR 3A film profile for all 5 dies- centermost die and 2nd dies to the right, left, above and below it.

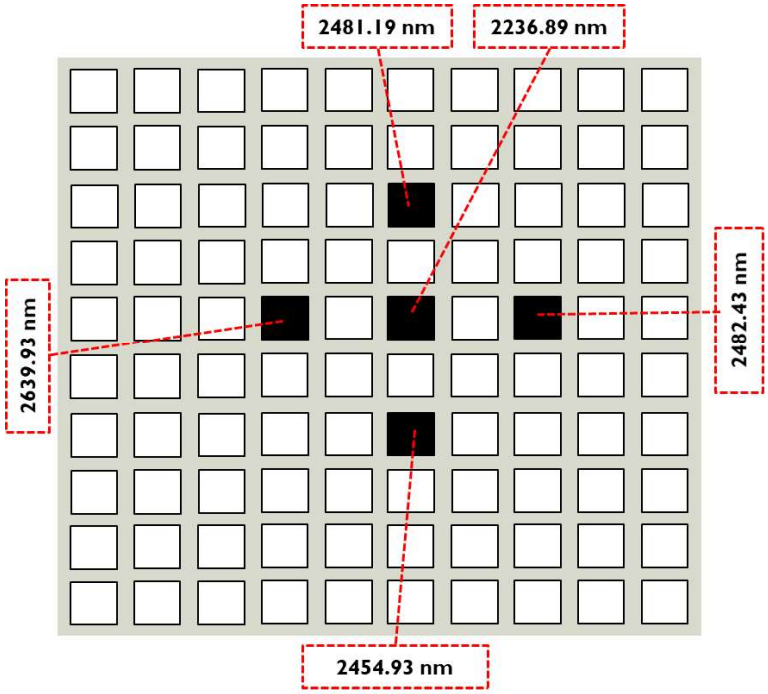


Fig. 149. SOG film thicknesses on 5 dies before etching.

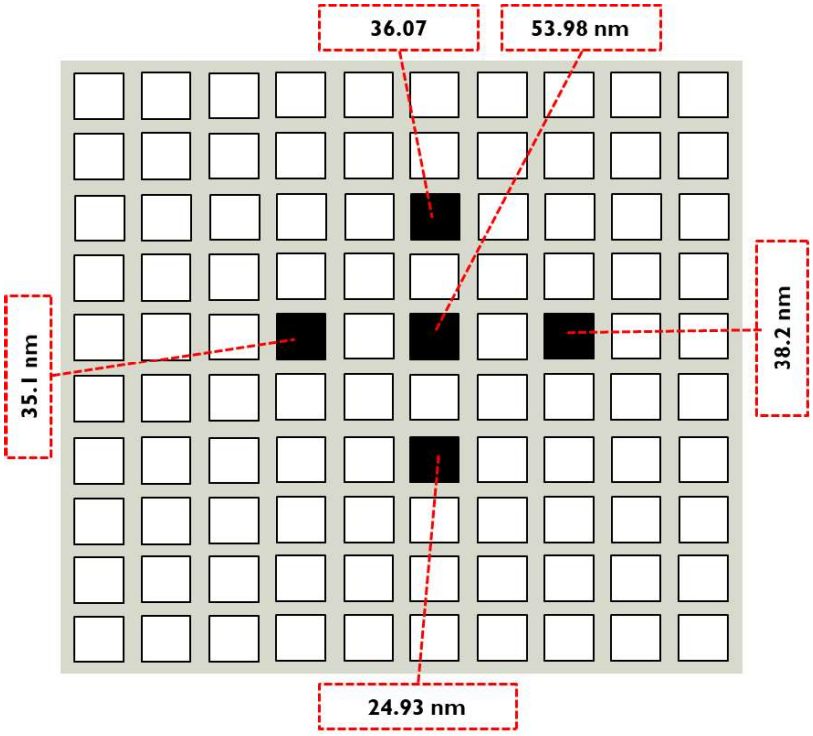
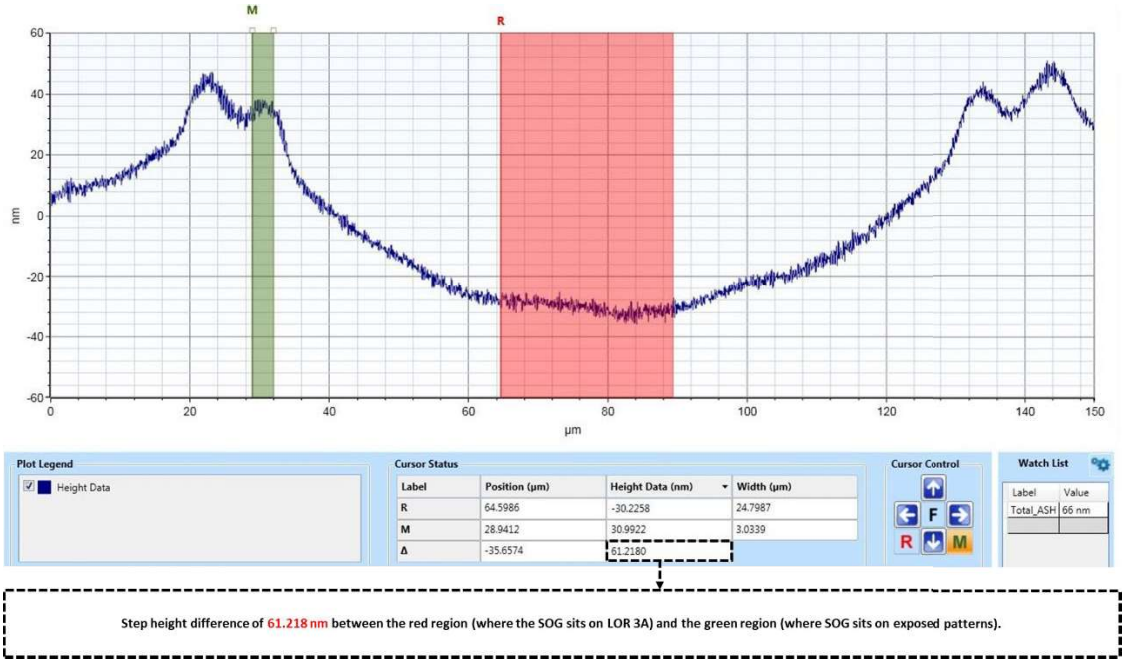
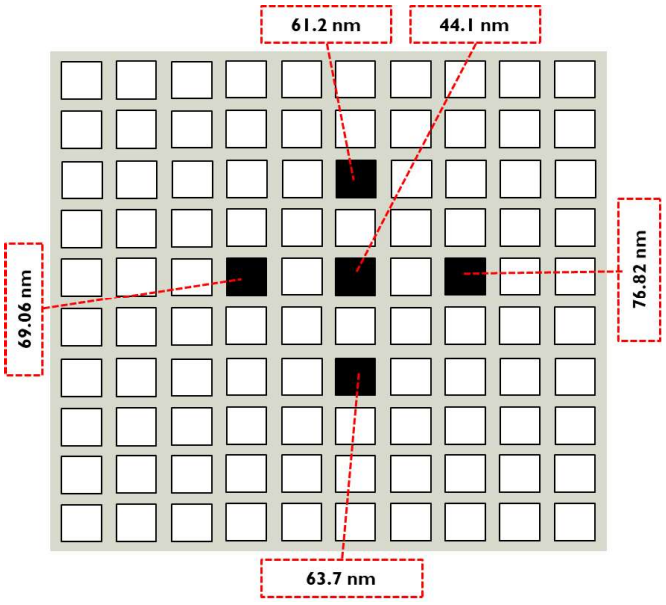


Fig. 150. SOG film thicknesses on 5 dies after etching.

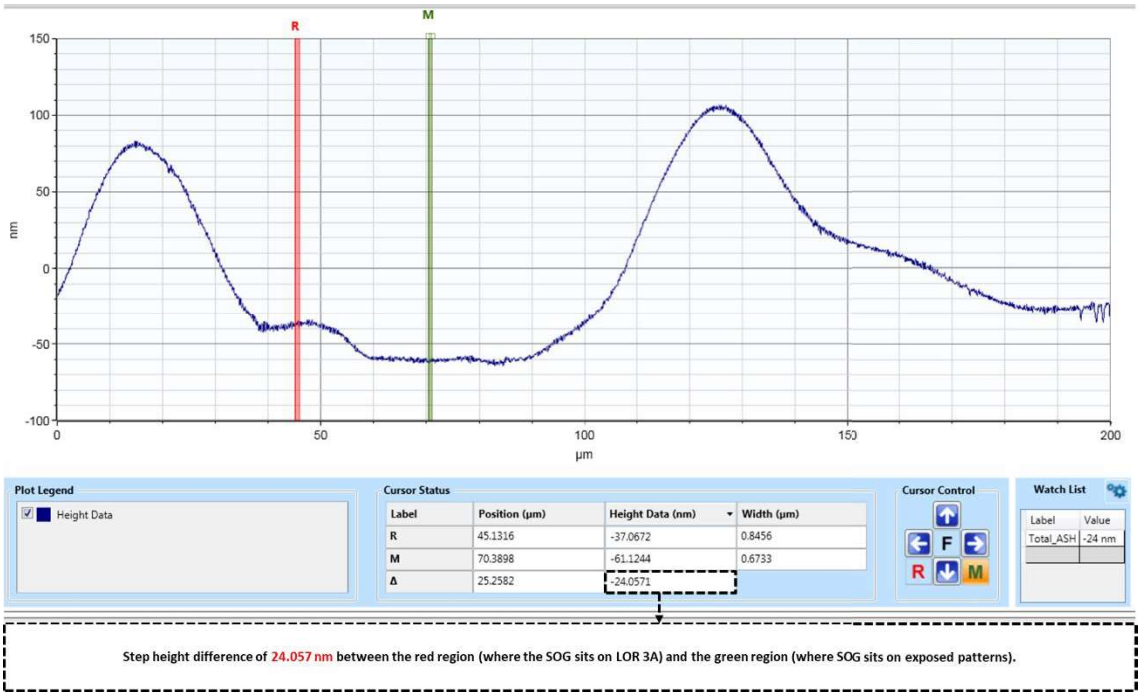


(a)

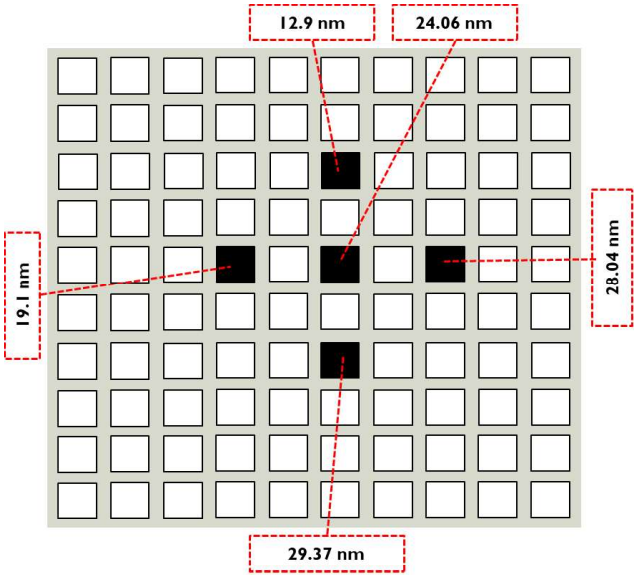


(b)

Fig. 151. SOG film profile before etching on the (a) centermost die, (b) SOG film profile for all 5 dies- centermost die and 2nd dies to the right, left, above and below it.



(a)



(b)

Fig. 152. SOG film profile after etching on the (a) centermost die, (b) SOG film profile for all 5 dies- centermost die and 2nd dies to the right, left, above and below it.

Another point to note is that it is only possible to measure SOG and LOR 3A thickness in locations where the layers sit right over the silicon and not over the metal patterns because of two reasons. Firstly, if the polished silicon wafer underneath is not the only reflector and if there are patterns in the area where the thickness is measured then the TiN ARC, Al-5Cu, Ti and the

silicon wafer all act as perfect reflectors thereby making it difficult to measure the layer thicknesses. Secondly, considering each of the layers in the patterns there are a total of 6 layers that are measured and reflectometry fails for more than 3 layers. But, the SOG thickness over electrode patterns can be calculated analytically using Eqns 21 and 22 before SOG etch-back. The variables in Eqs. 21 and 22 are depicted in Fig. 153.

$$h_{over\ metal\ stack} = \left(h_{step} + h_{over\ Si}\right) - h_{stack} \tag{21}$$

$$h_{over\ Si} = \left(h_{LOR\ 3A\ over\ Si} + h_{SOG\ over\ Si}\right) \tag{22}$$

where

$h_{over\ metal\ stack}$ = SOG film stack thickness over metal patterns.
h_{step} = step coverage across metal pattern as measured by profilometer.
$h_{over\ Si}$ = SOG and etched LOR 3A total thickness over Si wafer.
h_{stack} = height of the patterns from the Si substrate (740 nm).
$h_{LOR\ 3A\ over\ Si}$ = etched LOR 3A thickness over Si wafer.
$h_{SOG\ over\ Si}$ = SOG thickness over Si wafer.

Now, if $h_{over\ metal\ stack}$ is calculated after SOG etch-back and comes to be negative, this means that the electrode is exposed as distance cannot be a negative value.

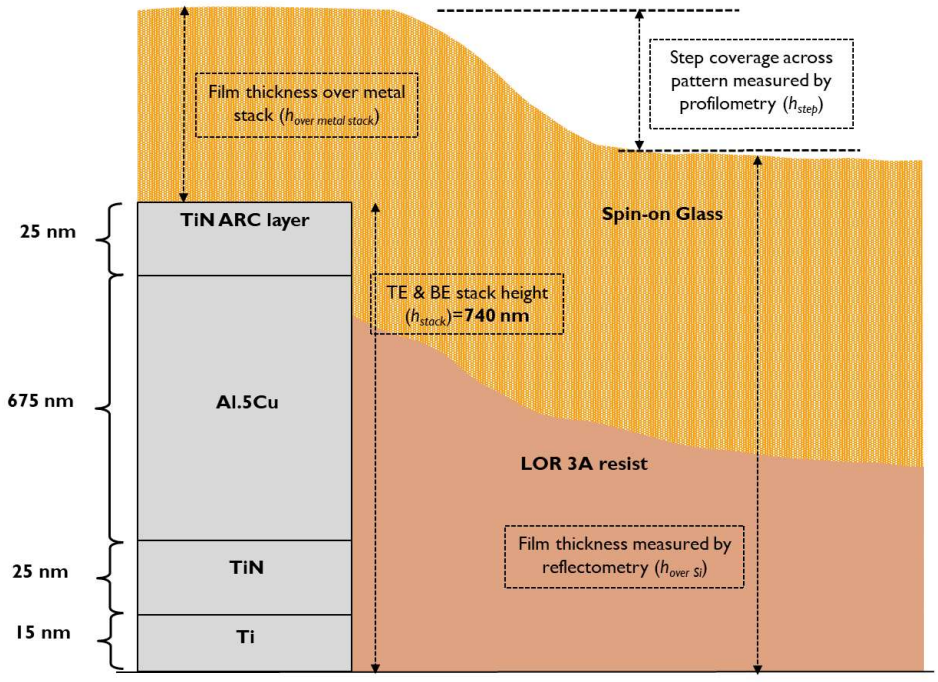
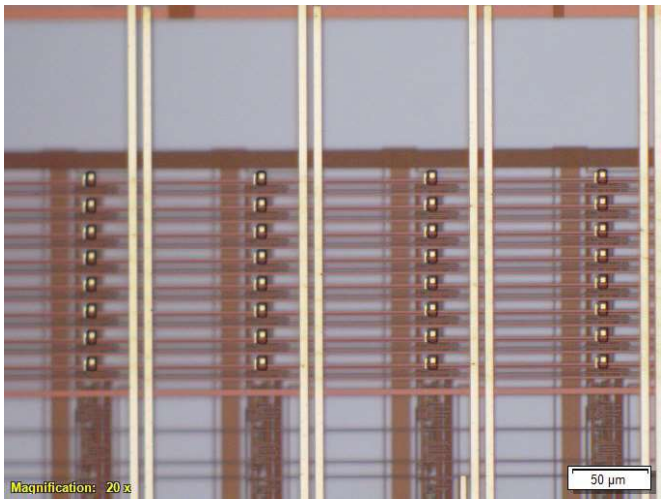


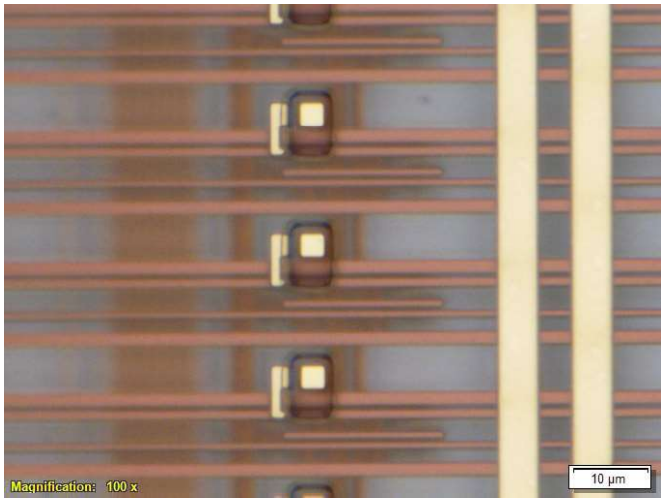
Fig. 153. SOG and LOR 3A profile over a pattern before performing SOG etch-back

Sandia National Laboratory (SNL) Process

A process recipe was also developed at Sandia National Laboratories using the same mask set as shown in Fig. 133. While the process was not fully developed due to this program ending in the third year, SNL staff at the MESA facility were able to complete mask steps on several prototype parts. Because MESA is equipped with a Chemical Mechanical Polish (CMP) tool, the SNL process did not have to follow the SOG process described above. Fig 154 shows a photomicrograph of the top of the wafer after etching after Mask 2. At this point the wafer would be ready for solid-state electrolyte deposition.



(a)



(b)

Fig. 154. (a) Top view and (b) closer view of VMM wafer after Mask 2 etch step perform at SNL MESA facility.

Perform tests to evaluate susceptibility to TID

The tested parts are a prototype application specific integrated circuit (ASIC) was designed and fabricated in a 350nm CMOS process technology to carry out small vector-matrix multiply (VMM) operations using a 8x8 crossbar architecture. The VMM operation is key linear algebra computational primitive that underlie many machine learning algorithms. As such, the VMM ASIC functions as an analog neuromorphic inference accelerator. For testing purposes, and in anticipation of delays in integrating emerging memories, the VMM ASIC was designed such that the last column of the crossbar array is populated with polysilicon resistors of known resistance in place of resistive memory elements. This enables the analog accelerator to compute a known and repeatable VMM value during inference. Fig. 155 shows the VMM ASIC layout view (left) and die micrograph (right).

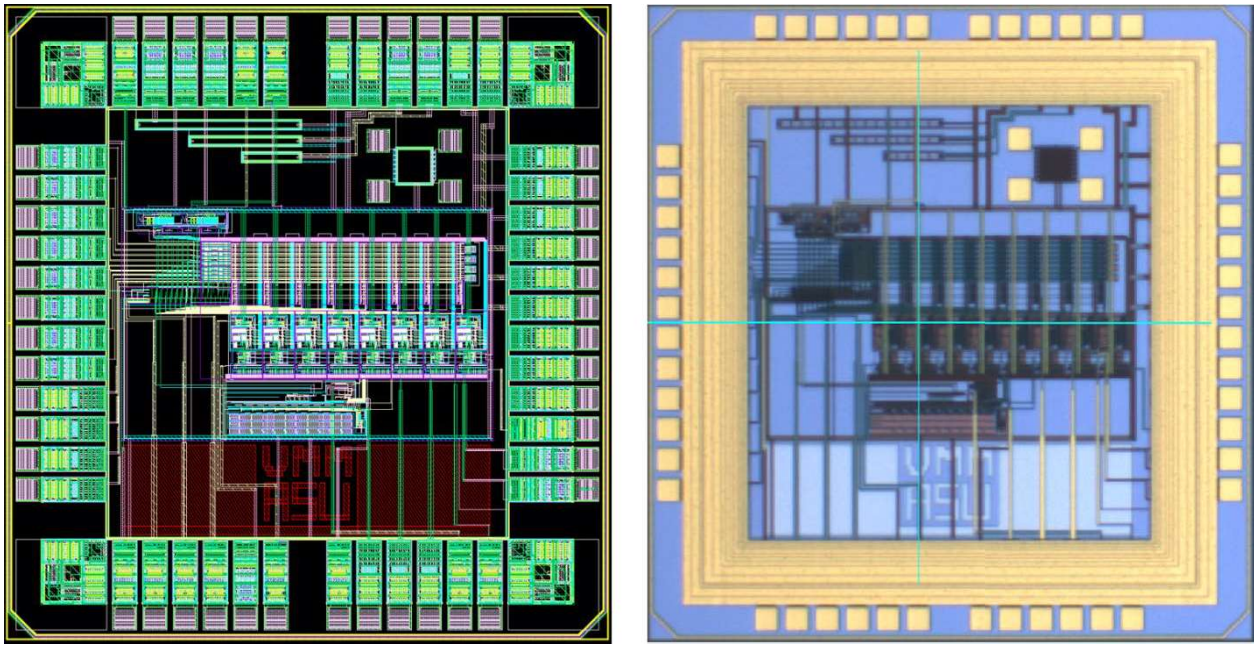


Fig. 155. VMM ASIC layout (left) and die micrograph (right).

For this radiation test, the TID resilience of the accelerator architecture was tested by measuring changes in the VMM value output from the fixed polysilicon resistor column of the crossbar for increasing dose levels. The neuromorphic accelerator has 8 analog inputs, which are input into the ASIC using an FPGA that connects to a custom interface PCB. Inference is achieved through a transimpedance amplifier (TIA) stage which sums and converts row device currents to voltage for subsequent ADC operation. The ADC is comprised of a single comparator and counter per column via a shared ramping digital-to-analog (DAC). It works by ramping a DAC incrementally through a range of values while 10-bit counters increment from zero; the column comparator trips when the ramp DAC equals the TIA output voltage (scaled dot product). The counter is then stopped, and the 10-bit digital count value is held for readout. Fig. 156 provides a block diagram of the neuromorphic accelerator design.

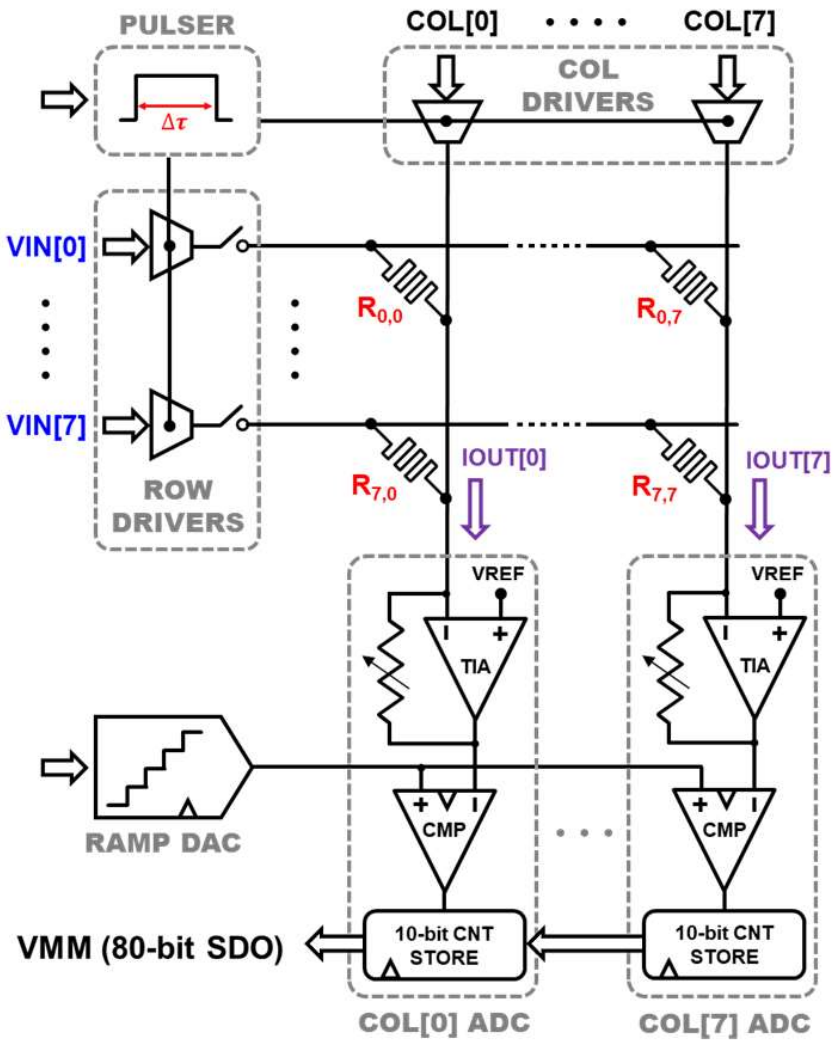


Fig. 156. Neuromorphic accelerator (VMM ASIC) block diagram.

The TIA used in the ASIC design has programmable gain (feedback resistor) to scale the dot product or changes in it into a range that spans more least significant bits (LSBs) of the ADC. There are 10 selectable gain settings. For the purposes of this test, only the read (inference) functionality of the accelerator design is employed. However, once they are integrated, the design can be conFig.d into modes to readily SET and RESET emerging resistive memory devices.

To control dataflow to and from the VMM ASIC, a Python interface was created to work in combination with an Opal Kelly field-programmable gate array (FPGA). Instructions are encoded in the Python interface on a personal computer (PC) and sent via a universal serial bus (USB) interface to the FPGA. The instructions are then decoded in the FPGA's main control unit (MCU). A printed circuit board (PCB) was designed to interface the digital signals output from the FPGA and input to the ASIC. The custom PCB has board level DAC components controlled through a Serial Peripheral Interface (SPI) which enable the analog inputs to the ASIC. The interface described here is worth mentioning as it enabled *in-situ* measurements, i.e. real-time analysis of the VMMs output from the ASIC during testing. More information at: https://www.osti.gov/servlets/purl/1846087. The ASIC was designed and fabricated in On Semiconductor's I3T25 foundry process, which is a 350nm CMOS technology. Details on the I3T25 PDK are available at: https://www.onsemi.com/products/product-taxonomy/custom-foundry-services/i3t25-process-technology

Test Methodology

In this test, two different individual ASIC parts have been inserted into a Cobalt-60 source producing gamma radiation. Depending on the test configuration, parts were either a) measured frequently during the activate irradiation (*in-situ* measures) and removed from the chamber and measured during breaks from the source (*ex-situ* measures), or b) only measured outside of the chamber. The two parts, hereafter A1 and A2, were subjected to protocol a) and b) , respectively. For both in-situ measurement and ex-situ measurement, total ionizing dose was computed based on the total time spent under the active sources. Every read in or out of the chamber was assigned a total time stamp and a TID value. In addition, 5 different gain settings on the TIA were interposed into the testing flow. The TIA feedback resistors gains are selected by loading data into an on-chip shift register chain; a 10-bit control signal is serially shifted into the TIA to connect one of the 10 different feedback resistors options. The TIA gain is of interest since it notably effects the shape of the output ADC curve. In principle, higher gains may help or hurt depending on the magnitude of offset bias desired for the VMM operation.

Key Results

Effect on characteristic ADC output

The target measurement during testing is the VMM values output by the last crossbar array column when the analog input voltages are swept from the negative (VSS) to positive (VDD) voltage rails. Each input voltage level produces a 10-bit digital output count corresponding to the computed VMM. In a neuromorphic crossbar array, the computed VMMs can change as different input data is applied, or if the synaptic weights in the array have been updated during

training. Since this test is measuring the output computed from a column of fixed resistors (fixed synaptic weights) the accelerator produces the same repeatable set of computed VMMs when sweeping the input voltages. The digital output counts are converted in software (Python interface) into decimal values between VSS (-1.6V) and VDD (1.6V). It should be noted that the column TIAs in this architecture use operational amplifiers which do not have a rail-to-rail input common mode range. Consequently, input voltage levels outside the input common mode range of the column TIA yield incorrect VMM values. This can be seen in Figs 158-160 where the column ADC is operating correctly when its output has linearly decreasing slope; the TIAs are inverting and thus the ADC will store a lower 10-bit digital value when there is a higher column current. For example, in Fig. 6 the ADC shows a correct output for an input voltage range of approximately -0.2V to 0.7V.

Our results indicate that TID doesn't notably affect the characteristic VMM output across the studied voltage range until a critical threshold. For part A1, which was read in-situ, this threshold was around 215 kRad, whereas for part A2, which was not, it was around 235 kRad. This thresholded TID response is visible in Fig. 157, where the y axis represents the sum of absolute value of error on the output vector over the voltage space, as compared to a gold standard result obtained at 0kRad (control). As visible in the Fig. 157, gain has little overall effect, and the only notable difference is between parts.

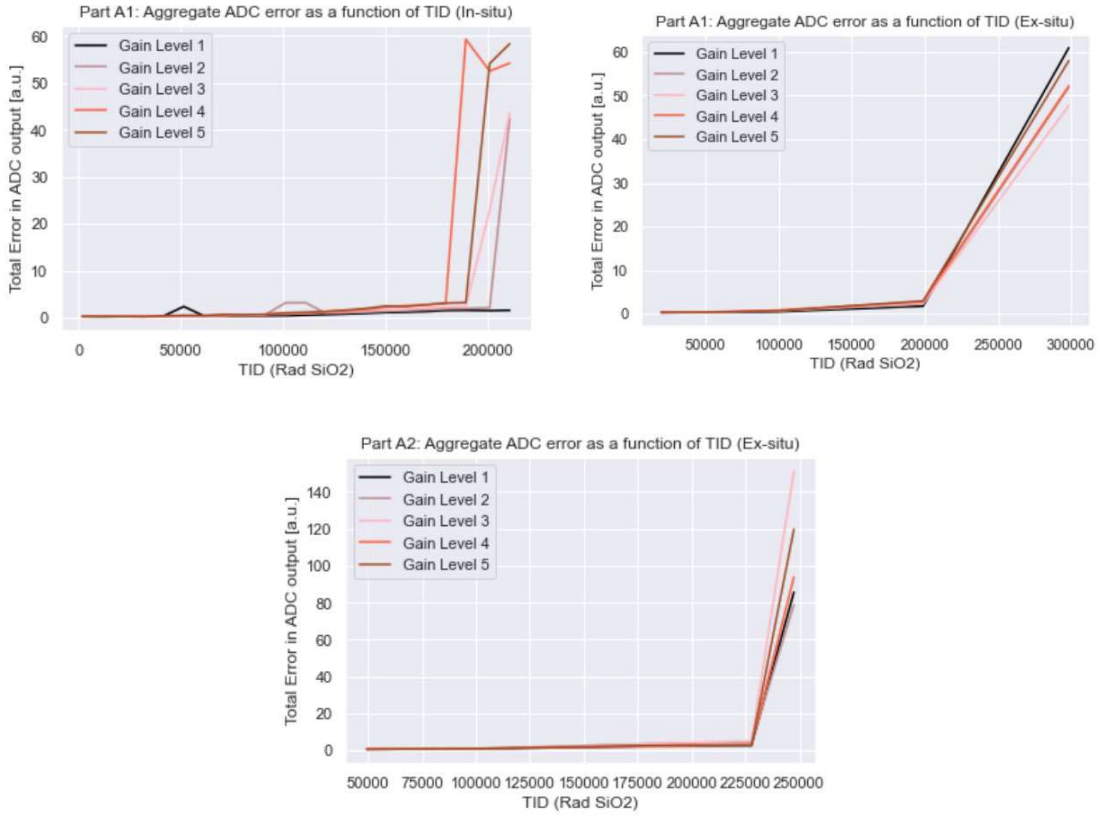


Fig. 157. TID threshold behavior for the 2 studied parts, over all 5 gain levels

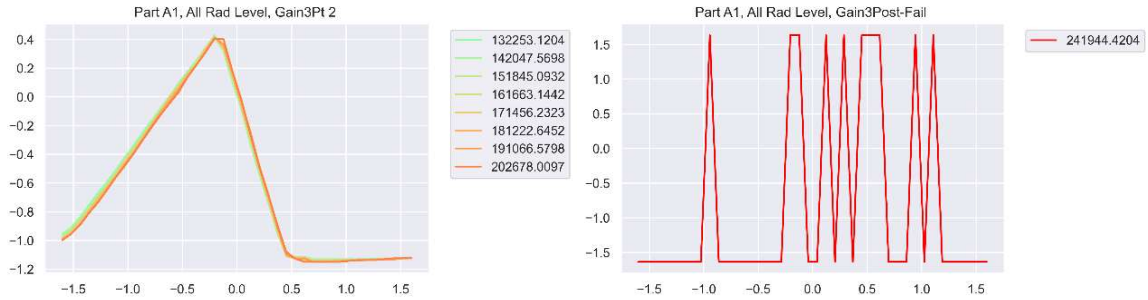


Fig. 158. Characteristic ADC Output over Input Voltage Space for A1 part – Some of In-situ measures (25 total completed per gain)

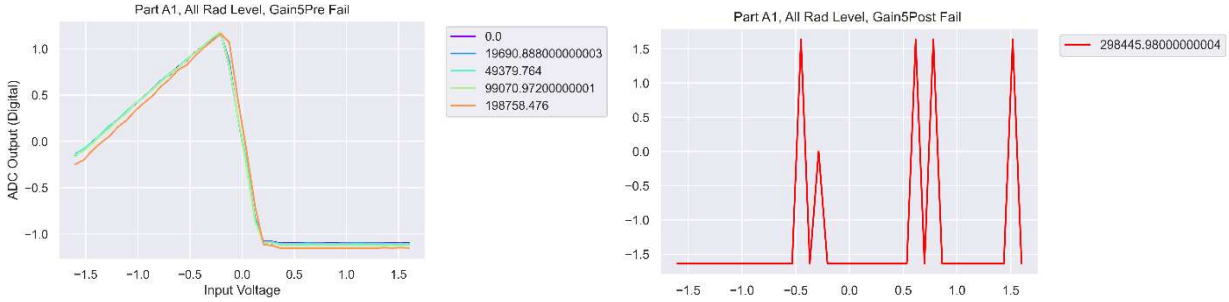


Fig. 159. Characteristic ADC Output over Input Voltage Space for A1 part – Ex-situ measures (4 total)

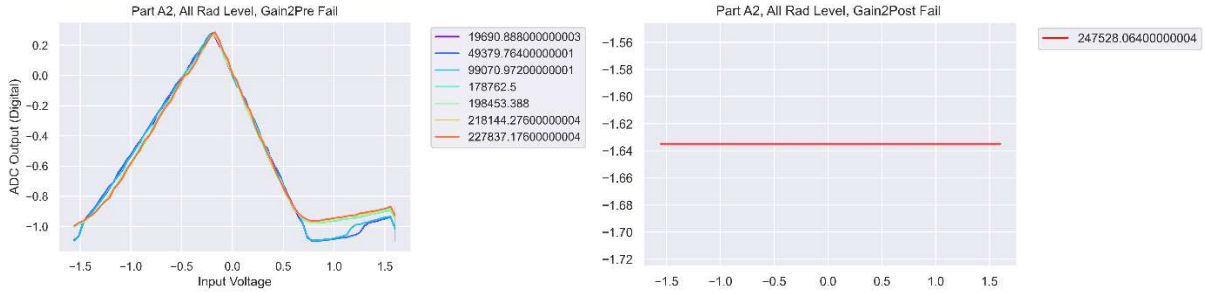


Fig. 160. Characteristic ADC Output over Input Voltage Space for A2 part – Ex-situ measures only (7 total)

As visible in Figs. 158-160 , the characteristic ADC output varies part-to-part and for every gain setting. But, for a given gain setting, there is minor change over the TID range up to the threshold , after which the ADC output is catastrophically affected (single curves showing complete failure depicted in the right pane of each Fig.). One exception is a change in the higher voltage range (Fig 6), which shifts for some gains on some parts.

Effect on supply current used by part

Another result is a fairly monotonic increase in supply current taken by the part during inference mode. For Part 1, this increased from 27.9 mA on Supply 1 and 32.5 mA on Supply 2 at 0kRad, to 41.3 mA and 45.3 mA respectively at 2667 exposure time (approx. ~190kRad), up to a final value of 53.6 mA and 56.8 mA at 4695 sec (post part failure). For A2 , similar numbers were

seen although with a slightly higher power burn as a baseline: S1 31.5mA, S2 40 mA at 0kRad ; 52.0 mA S1, 60.4 mA at 200k ; and 55.3 mA, 63.6mA at 230kRad (immediately post failure).

.

What opportunities for training and professional development has the project provided?
If the research is not intended to provide training and professional development opportunities or there is nothing significant to report during this reporting period, state "Nothing to Report." Describe opportunities for training and professional development provided to anyone who worked on the project or anyone who was involved in the activities supported by the project. "Training" activities are those in which individuals with advanced professional skills and experience assist others in attaining greater proficiency. Training activities may include, for example, courses or one-on-one work with a mentor. "Professional development" activities result in increased knowledge or skill in one's area of expertise and may include workshops, conferences, seminars, study groups, and individual study. Include participation in conferences, workshops, and seminars not listed under major activities.

Arizona State University

- Kiraneswar Muthuseenu completed his Ph.D degree from ASU in December 2020
- Blayne Tolleson completed his Ph.D qualifying exam at ASU. He is currently active in the PHD program. He does not have a graduation date scheduled. Blayne is a graduate research assistant currently working on the neuromorphic circuit chip design. He has served as a Teaching Assistant for EEE 335, Analog and Digital Circuits, at ASU.
- Priyanka Apsangi completed her Ph.D in July 2022 and is currently employed at ON Semiconductor.
- Xu Han completed her Ph.D at ASU in June 2022 and is currently employed at Qualcom.
- M. Musisi-Nkambwe is currently active in the PHD program. She does not have a graduation date scheduled. Mirembe is currently working on neuromorphic architectures.
- Matt Spear is currently active in the ASU PHD program. He does not have a graduation date scheduled. Matt is a graduate research assistant currently working on CMOS modeling, design, and test
- Prabaha Biswas completed is MS degree from ASU in May 2021
- Don Wilson is currently active in the ASU PhD program. He does not have a graduation date scheduled. Don is currently working on neuromorphic circuit test and evaluation.
- Matt Morton completed his BS in Electrical Engineering December 2020.
- Aaron Saltou completed his BS in Electrical Engineering May 2022.
- Jesse Short completed his BS in Electrical Engineering May 2022, and is currently active in the ASU graduate program. He does not have a graduation date scheduled. Jesse is currently working on neuromorphic circuit design
- Katherine Pearson completed her BS in Electrical Engineering May 2022 and is currently active in the ASU graduate program. She does not have a graduation date scheduled. Katherine is currently working on RF circuit design
- Trace Wallace completed his BS in Electrical Engineering May 2022, and is currently active in the ASU PhD. program. He does not have a graduation date scheduled. Trace is currently working on radiation effects in electronics
- Zachary Giorno completed his BS in Electrical Engineering May 2022, and is currently active in the ASU MS program. He does not have a graduation date scheduled. Zachary is currently working on radiation effects in electronics.
- Jennifer Taggart graduated with a Ph.D. in Electrical Engineering in December 2018. Jennifer is now working at the Aerospace Corporation in Los Angeles, CA.
- Rui Liu graduated with a Ph.D. in Electrical Engineering in December 2018.
- Zhilu Ye completed her MSEE (non-thesis) in Electrical Engineering in May 2019.

Sandia National Laboratories

- Christopher Bennett is currently serving as staff at Sandia National Laboratories. He is currently working on the neuromorphic circuit chip design.
- T. P. Xiao Bennett is currently serving as staff at Sandia National Laboratories. He is currently working on the neuromorphic circuit chip design.
- Patrick Finnegan is currently serving as staff at Sandia National Laboratories. He is currently working in the SNL MESA facility
- Robin Jacobs-Gedrim is currently serving as a post-doctoral researcher at Sandia National Laboratories. Robin presented a paper at the 2017 IEEE International Conference on Rebooting Computing (ICRC) in Washington, DC. He is author of an abstract submitted and accepted for presentation at the 2018 IEEE Nuclear and Space Radiation Effects Conference in Kona HI.
- John Niroula served as an undergraduate intern at Sandia National Laboratories in the summer and fall of 2017. He graduated from MIT in May 2018. John was co-author on a paper presented at the 2017 IEEE International Conference on Rebooting Computing (ICRC) in Washington, DC.
- Diana Garland is serving as an undergraduate intern at Sandia National Laboratories in the summer 2018. She is an undergraduate student at the University of Arizona.

How have the results been disseminated to communities of interest?
If there is nothing significant to report during this reporting period, state "Nothing to Report." Describe how the results have been disseminated to communities of interest. Include any outreach activities that have been undertaken to reach members of communities who are not usually aware of these research activities, for the purpose of enhancing public understanding and increasing interest in learning and careers in science, technology, and the humanities.

Refereed Archival Journal Papers

1. T. P. Xiao, C. H. Bennett, S. Agarwal, D. R. Hughart, H. J. Barnaby, H. Puchner, A. A. Talin, and M. J. Marinella, "Single event effects induced by heavy ions in SONOS charge trapping memory arrays," IEEE Transactions on Nuclear Science, 2021.
2. X. Han, A. Privat, K. Holbert, J.-s. Seo, S. Yu, and H. Barnaby, "Total Ionizing Dose Effects on Multistate HfO_x-Based RRAM Synaptic Array," IEEE Transactions on Nuclear Science, vol. 68, no. 5, pp. 756-761, 2021.
3. M. Musisi-Nkambwe, S. Afshari, H. Barnaby, M. N. Kozicki, and I. S. Esqueda, "The Viability of Analog-based Accelerators for Neuromorphic Computing: a Survey," Neuromorphic Computing and Engineering, 2021.
4. T. Patrick Xiao, C. H. Bennett, S. Agarwal, D. R. Hughart, H. J. Barnaby, H. Puchner, V. Prabhakar, A. Alec Talin, and M. J. Marinella, "Ionizing radiation effects in SONOS-based neuromorphic inference accelerators," Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 0018-9499, 2021.
5. T. P. Xiao, C. H. Bennett, S. Agarwal, D. R. Hughart, H. J. Barnaby, H. Puchner, V. Prabhakar, A. A. Talin, and M. J. Marinella, "Ionizing Radiation Effects in SONOS-Based Neuromorphic Inference Accelerators," IEEE Transactions on Nuclear Science, vol. 68, no. 5, pp. 762-769, 2021.
6. T. P. Xiao, C. H. Bennett, B. Feinberg, S. Agarwal, M. J. Marinella, "Analog architectures for neural network acceleration based on non-volatile memory", Applied Physics Reviews 7(3), 2020. doi 10.1063/1.5143815
7. Y. Li, T. P. Xiao, C. Bennett, E. Isele, A. Melianas, H. Tao , M. Marinella, A. Salleo, E. Fuller, A. A. Talin, "In-situ Parallel Training of Analog Neural Network using Electrochemical Random-Access Memory", Frontiers in Neuroscience, 2021, doi 10.3389/fnins.2021.636127
8. M. J. Marinella, "Radiation Effects in Advanced and Emerging Nonvolatile Memories," in IEEE Transactions on Nuclear Science, vol. 68, no. 5, pp. 546-572, May 2021, doi: 10.1109/TNS.2021.3074139.
9. Y. Luo, X. Han, Z. Ye, H. Barnaby, J.-S. Seo, and S. Yu, "Array-level programming of 3-bit per cell resistive memory and its application for deep neural network inference," IEEE Transactions on Electron Devices, vol. 67, no. 11, pp. 4621-4625, 2020. Doi 10.1109/TED.2020.3015940
10. M. Musisi-Nkambwe, S. Afshari, H. Barnaby, M. N. Kozicki, I. S. Esqueda, "The Viability of Analog-based Accelerators for Neuromorphic Computing: a Survey," Neuromorphic Computing and Engineering, 2021. Doi 10.1088/2634- 4386/ac0242
11. K. Muthuseenu, H. Barnaby, A. Patadia, K. Holbert, and A. Privat, "Ionizing radiation tolerance of stacked Si3N4-SiO2 gate insulators for power MOSFETs," Microelectronics Reliability, vol. 104, p. 113554, 2020.
12. S. Yin, Y. Kim, X. Han, H. Barnaby, S. Yu, Y. Luo, W. He, X. Sun, J.-J. Kim, and J.-S. Seo, "Monolithically Integrated RRAM-and CMOS-Based In-Memory Computing Optimizations for Efficient Deep Learning," IEEE Micro, vol. 39, no. 6, pp. 54-63, November, 2019.
13. I. P. Livingston, I. S. Esqueda, and H. J. Barnaby, "Explicit approximation of the surface potential equation of a dynamically depleted silicon-on-insulator MOSFET for performance and reliability simulations," Solid-State Electronics, vol. 160, p. 107609, October, 2019.

14. Z. L. Ye, R. Liu, J. L. Taggart, H. J. Barnaby, S. M. Yu, "Evaluation of Radiation Effects in RRAM-Based Neuromorphic Computing System for Inference," IEEE Transactions on Nuclear Science, vol. 66, no. 1, pp. 97-103, Jan 2019.
15. Z. Ye, R. Liu, H. J. Barnaby, and S. Yu, "Evaluation of Single Event Effects in SRAM and RRAM Based Neuromorphic Computing System for Inference," 2019 IEEE International Reliability Physics Symposium (IRPS), pp. 1-4, 2019.
16. R. B. Jacobs-Gedrim, D. R. Hughart, G. Vizkelethy, E. S. Bielejec, B. L. Vaandrager, S. E. Swanson, K. E. Knisely, J. L. Taggart, H. J. Barnaby, M. J. Marinella, "Training a Neural Network on Analog TaOx ReRAM Devices Irradiated With Heavy Ions: Effects on Classification Accuracy Demonstrated With CrossSim," IEEE Transactions on Nuclear Science, vol. 66, no. 1, pp. 54-60, Jan 2019.
17. J. L. Taggart, R. B. Jacobs-Gedrim, H. J. Barnaby, E. S. Bielejec, W. Hardy, M. J. Marinella, M. N. Kozicki, K. Holbert, "Failure Thresholds in CBRAM Due to Total Ionizing Dose and Displacement Damage Effects," IEEE Transactions on Nuclear Science, vol. 66, no. 1, pp. 69-76, Jan 2019.
18. I. P. Livingston, I. S. Esqueda, and H. J. Barnaby, "Explicit approximation of the surface potential equation of a dynamically depleted silicon-on-insulator MOSFET for performance and reliability simulations," Solid State Electronics, vol. 160, 2019.
19. Matthew J. Marinella, Robin B. Jacobs-Gedrim, Chris Bennett, Alex Hsia, David R. Hughart, Conrad D. James, George Vizkelethy, Ed Bielejec, Sapan Agarwal, Elliot J. Fuller, Alec A. Talin, Jennifer L. Taggart, Hugh J. Barnaby, “Energy Efficient Neuromorphic Algorithm Training with Analog Memory Arrays,” in Proceedings of the Government Microcircuits and Critical Technologies Conference (GoMACTech), Albuquerque NM, 2019. (acknowledged DTRA)
20. C. H. Bennett, D. Garland, R. B. Jacobs-Gedrim, S. Agarwal and M. J. Marinella, "Wafer-Scale TaOx Device Variability and Implications for Neuromorphic Computing Applications," 2019 IEEE International Reliability Physics Symposium (IRPS), Monterey, CA, USA, 2019, pp. 1-4. doi: 10.1109/IRPS.2019.8720596
21. S. Agarwal, MJ Marinella, "Using Floating-Gate Memory to Train Ideal Accuracy Neural Networks," in IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, vol. 5, no. 1, pp. 52-57, June 2019.doi: 10.1109/JXCDC.2019.2902409
22. Elliot J. Fuller, Scott T. Keene, Armantas Melianas, Zhongrui Wang, Sapan Agarwal, Yiyang Li, Yaakov Tuchman, Conrad D. James, Matthew J. Marinella, J. Joshua Yang, Alberto Salleo, A. Alec Talin, "Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing," Science, vol. 364, no. 6440, pp. 570-574, 2019.
23. C. H. Bennett. "Contrasting Advantages of Learning With Random Weights and Backpropagation in Non-Volatile Memory Neural Networks," in IEEE Access, vol. 7, pp. 73938-73953, 2019. doi: 10.1109/ACCESS.2019.2920076
24. 1. J. L. Taggart, W. Chen, Y. Gonzalez-Velo, H. J. Barnaby, K. Holbert, and M. N. Kozicki, "In Situ Synaptic Programming of CBRAM in an Ionizing Radiation Environment," IEEE Transactions on Nuclear Science, vol. 65, pp. 192-199, Jan 2018.
25. J. L. Taggart, R. Fang, Y. Gonzalez-Velo, H. J. Barnaby, M. N. Kozicki, J. L. Pacheco, et al., "Resistance State Locking in CBRAM Cells Due to Displacement Damage Effects," IEEE Transactions on Nuclear Science, vol. 64, pp. 2300-2306, Aug 2017.
26. W. Chen, N. Chamele, Y. Gonzalez-Velo, H. Barnaby, and M. Kozicki, "Low Temperature Characterization of Cu-Cu: silica Based Programmable Metallization Cell," IEEE Electron Device Letters, vol. 38, pp. 1244-1247, Sep 2017.
27. W. Chen, S. Tappertzhofen, H. J. Barnaby, and M. N. Kozicki, "SiO2 based conductive bridging random access memory," Journal of Electroceramics, vol. 39, pp. 109-131, Dec 2017.
28. Y. Gonzalez-Velo, H. J. Barnaby, and M. N. Kozicki, "Review of radiation effects on ReRAM devices and technology," Semiconductor Science and Technology, vol. 32, Aug 2017.

29. A. Mahmud, Y. Gonzalez-Velo, H. J. Barnaby, M. N. Kozicki, M. Mitkova, K. E. Holbert, M. Goryll, T. L. Alford, J. L. Taggart, W. Chen, "A Comparative Study on TID Influenced Lateral Diffusion of Group 11 Metals into GexS 1-x and Ge x Se 1-x Systems: A Flexible Radiation Sensor Development Perspective," IEEE Transactions on Nuclear Science, vol. 64, no. 8, pp. 2292 – 2299, Aug. 2017.
30. M. J. Marinella, S. Agarwal, A. Hsia, I. Richter, R. B. Jacobs-Gedrim, J. Niroula, S. J. Plimpton, E. Ipek, C. D. James, “Multiscale Co-Design Analysis of Energy, Latency, Area, and Accuracy of a ReRAM Analog Neural Training Accelerator,” IEEE Journal on on Emerging and Selected Topics in Circuits and Systems, vol. 8, no. 1, Mar 2018.

National Conference Proceedings Refereed Papers

1. C. H. Bennett, T. P. Xiao, R. Dellana, B. Feinberg, V. Agrawal, V. Prabhakar, K. Ramkumar, L. Hinh, S. Saha, V. Raghavan, R. Chettuvetty, S. Agarwal, M. Marinella, “Device-aware inference operations in SONOS non-volatile memory arrays”, IEEE International Reliability Physics Symposium (IRPS), 2020. doi 10.1109/IRPS45951.2020.9129313
2. K. Muthuseenu, E. C. Hylin, H. J. Barnaby, P. Apsangi, M. N. Kozicki, G. Schlenvogt, and M. Townsend, "TCAD Model for Ag-GeSe 3-Ni CBRAM Devices," in 2019 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), September, 2019: IEEE, pp. 1-4.

National Conference Presentations

1. X. Han, M. Spear, J. Seo, D. Wilson T. Wallace, O. Forman, J. Solano, M. Turowski, M. Marinella, H. J. Barnaby, “Total Ionizing Dose Effects in FDSOI SRAM-Based XNOR IMC Synaptic Array,” *presented at the IEEE Nuclear and Space Radiation Effects Conference,* Provo, UT, July, 2022.
2. P. Apsangi, H. J. Barnaby, N. Chamele, Y. Gonzalez-Velo, K. Holbert, M. N. Kozicki, “Impact of Total Ionizing Dose on Analog Synaptic Characteristics of CBRAM,” *presented at the IEEE Nuclear and Space Radiation Effects Conference*, Santa Fe, NM, Dec. 1-4, 2020
3. X. Han, A. Privat, K. Holbert, J. Seo, H. J. Barnaby, S. Yu, “Total Ionizing Dose Effects on Multi-state HfOx-based RRAM Synaptic Arrays,” *presented at the IEEE Nuclear and Space Radiation Effects Conference,* Santa Fe, NM, Dec. 1-4, 2020.
4. T.P. Xiao, C.H. Bennett, S. Agarwal, D. Hughart, H. Barnaby, H. Puchner, A.A. Talin, M. Marinella, “Single event effects induced by heavy ions in SONOS charge trapping memory arrays “”, Accepted, NSREC 2021
5. C.H. Bennett, Donald A. Robinson2, T. Patrick Xiao1 Armantas Melianas3, David Hughart1, Sapan Agarwal2, Eliot J. Fuller 2, Yiyang Li3, Alberto Salleo3,, Mark .D. Allendorf2, A. Alec Talin2, and Matthew J. Marinella1, “Effects of Total Ionizing Dose on PEDOT/PSS Organic Memory Devices”, Accepted , NSREC 2021
6. P. Apsangi, H. J. Barnaby, N. Chamele, Y. Gonzalez-Velo, K. Holbert, M. N. Kozicki, “Impact of Total Ionizing Dose on Analog Synaptic Characteristics of CBRAM,” accepted at the IEEE Nuclear and Space Radiation Effects Conference, Santa Fe, NM, Dec. 1-4, 2020
7. X. Han, A. Privat, K. Holbert, J. Seo, H. J. Barnaby, S. Yu, Total Ionizing Dose Effects on Multi-state HfOx-based RRAM Synaptic Arrays,” accepted at the IEEE Nuclear and Space Radiation Effects Conference, Santa Fe, NM, Dec. 1-4, 2020.
8. M. Cannon, A. Rodrigues, D. Black, J. Black, L. Bustamante, B. Feinberg, M. McLain, S. Aggarwal, M. Marinella, Sandia National Laboratories, USA, L. Clark, J. Brunhaver, H. Barnaby, “Multiscale System Modeling of Single Event Induced Faults in Advanced Node Processors,” accepted at the IEEE Nuclear and Space Radiation Effects Conference, Santa Fe, NM, Dec. 1-4, 2020.
9. K. Muthuseenu, H. J. Barnaby, A. Patadia, K. Holbert, and A. Privat, "Ionizing Radiation Tolerance of Stacked Si3N4-SiO2 Gate Insulators for Power MOSFETs," presented at the IEEE Nuclear and Space Radiation Effects Conference, San Antonio, TX, July 8-12, 2019.

10. (Invited) M.J. Marinella S. Agarwal, R. Jacobs-Gedrim, D. Hughart, C. Bennett, A. Hsia, E. Fuller, A. Talin, H.J. Barnaby “Efficient Neuromorphic Computing in SWaP-Constrained Environments,” Presented at the Microelectronics Qualification and Reliability Conference (MRQW), Los Angeles, 2019. (acknowledged DTRA)
11. R. B. Jacobs-Gedrim, D. R. Hughart, G. Vizkelethy, E. S. Bielejec, B. L. Vaandrager, S. E. Swanson, K. E. Knisely, J. L. Taggart, H. J. Barnaby, M. J. Marinella, “Effect of Radiation on the Classification Accuracy of a Neural Network Trained on Analog TaOx Resistive Memory Arrays,” IEEE Nuclear and Space Radiation Effects Conf., Kona, HI, July 2018.
12. J. L. Taggart, R. B. Jacobs-Gedrim, H. J. Barnaby, E. S. Bielejec, W. Hardy, M. J. Marinella, M. N. Kozicki, K. Holbert, “Separation of Total Ionizing Dose and Displacement Damage Effects in CBRAM Cells,” IEEE Nuclear and Space Radiation Effects Conf., Kona, HI, July 2018.
13. Z. Ye, R. Liu, J. L. Taggart, H. J. Barnaby, S. Yu, “Evaluation of Radiation Effects in RRAM based Neuromorphic Computing System,” IEEE Nuclear and Space Radiation Effects Conf., Kona, HI, July 2018.
14. K. Muthuseenu, H. J. Barnaby, A. Patadia, K. Holbert, and A. Privat, "Ionizing Radiation Tolerance of Stacked Si3N4-SiO2 Gate Insulators for Power MOSFETs," (accepted for presentation) IEEE Nuclear and Space Radiation Effects Conference, San Antonio, TX, 2019.
15. K. Muthuseenu et al., "TCAD Model for Ag-GeSe3-Ni CBRAM Devices," (accepted for presentation) SISPAD Conference, Udine, Italy, 2019
16. (invited) M. J Marinella et al., “Toward an Analog Neural Accelerator with 10 fJ per Operation using Resistive Synaptic Devices,” presented at the 2017 Fall Materials Research Symp., Boston MA, Nov 2017.
17. (invited) M. J. Marinella et al., “Nanoionics for Efficient Neuromorphic Computing,” presented at the 21st Intl. Conf. on Solid State Ionics, Padua Italy, June 2017.
18. (invited) M. J. Marinella, “Energy Efficient Neural Algorithm Training and Execution with Analog Crossbar Accelerators,” Short course section, 2nd IEEE Electron Devices Tech and Manuf. Conf, Kobe, Japan, March 2018. http://ewh.ieee.org/conf/edtm/2018/tutorial/short_courses.html
19. (invited) M. N. Kozicki, “Fabrication and Characterization of Copper-Silica-Metal Resistive Switching Devices,” 232nd ECE Meeting, National Harbor, MD, October 2017.
20. (invited) M. N. Kozicki, “Memristive Devices - From Fundamentals to Applications,” Symposium H.3, XXVI International Materials Research Congress (IMRC 2017), August, 2017, Cancun, Mexico.
21. (invited) M. N. Kozicki, “Simplified Resistive Memory for CMOS Integration,” Symposium FM, XXVI 8th Forum on New Materials (CIMTEC 2018), June, 2018, Perugia, Mexico.
22. (invited) H. J. Barnaby, “Resistive Memory Technology and Applications,” Symposium FM, 8th Forum on New Materials (CIMTEC 2018), June, 2018, Perugia, Mexico.
23. J. L. Taggart, W. Chen, Y. Gonzalez-Velo, H. J. Barnaby, K. Holbert, M. N. Kozicki, "In Situ Synaptic Programming of CBRAM in an Ionizing Radiation Environment,” IEEE Nuclear and Space Radiation Effects Conf., New Orleans, LA., July 2017.
24. R. B. Jacobs-Gedrim, D. R. Hughart, G. Vizkelethy, E. S. Bielejec, B. L. Vaandrager, S. E. Swanson, K. E. Knisely, J. L. Taggart, H. J. Barnaby, M. J. Marinella, “Effect of Radiation on the Classification Accuracy of a Neural Network Trained on Analog TaOx Resistive Memory Arrays,” IEEE Nuclear and Space Radiation Effects Conf. (accepted),” Kona, HI, July 2018.
25. J. L. Taggart, R. B. Jacobs-Gedrim, H. J. Barnaby, E. S. Bielejec, W. Hardy, M. J. Marinella, M. N. Kozicki, K. Holbert, “Separation of Total Ionizing Dose and Displacement Damage Effects in CBRAM Cells,” IEEE Nuclear and Space Radiation Effects Conf. (accepted),” Kona, HI, July 2018.

Patent Applications

1. U.S. Provisional Patent Application No. 62/945,683, filed December 9, 2019, “PHYSICAL UNCLONABLE FUNCTIONS WITH SILICON-RICH SiOx DIELECTRIC DEVICES,” Inventor: Michael Kozicki

Ph.D. Thesis

1. Kiraneswar Muthuseenu, “Development of Radiation Hardened High Voltage Super-Junction Power MOSFET,” December 2020
2. Priyanka Apsangi, “Conductive Bridge Random Access Memory (CBRAM) as an Analog Synapse for Neuromorphic Computing,” July 2022
3. Xu Han, “Radiation Effects on In-Memory Computing Architectures.”June 2022
4. Jennifer Taggart, “Space Radiation Effects in Conductive Bridging Random Access Memory.” December 2018
5. Rui Liu, “Semiconductor Memory Applications in Radiation Environment, Hardware Security and Machine Learning System” December 2018

MS Thesis

1. Prabaha Biswas, “Spin-on Glass and LOR 3A Resist-based Planarization Technique for Neuromorphic CMOS Chips,” April 2021
2. Zhilu Ye completed her MSEE (non-thesis) in Electrical Engineering in May 2019.