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

DTRA-TR-21-19

TECHNICAL REPORT

Radiation Effects in Brain-Inspired Computing Systems

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

May 2021

HDTRA1-17-1-0035

Prepared by:
Subramanian Iyer

University of California, Los Angeles, CA 90095

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

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

1. REPORT DATE *(DD-MM-YYYY)*
02/17/2021

2. REPORT TYPE
Final Report

3. DATES COVERED *(From - To)*
4/18/2017 - 10/17/2020

4. TITLE AND SUBTITLE
Radiation Effects in Brain-Inspired Computing Systems

5a. CONTRACT NUMBER

5b. GRANT NUMBER
HDTRA1-17-1-0035

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHOR(S)
Subramanian Iyer - PI, UCLA
Jason Woo - Co-PI, UCLA
Michael Alles - Co-PI, Vanderbilt
Brian Sierawski - Co-PI, Vanderbilt

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UCLA - 10889 Wilshire Blvd, Suite 700, Los Angeles, CA 90095-1406
Vanderbilt University - 2301 Vanderbilt Place, Nashville, TN, 37240

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Threat Reduction Agency
8725 John J. Kingman Road
Fort Belvoir, VA 22060-6201

10. SPONSOR/MONITOR'S ACRONYM(S)
DTRA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
DTRA-TR-21-19

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
Final report for award HDTRA1-17-1-0035.

14. ABSTRACT
This work explores the effects of radiation in digital & analog hardware implementations of brain-inspired computing systems. The IBM TrueNorth Neurosynaptic System was utilized as a baseline system to evaluate the effects of radiation on neural network applications emulated by von Neumann hardware. The IBM TrueNorth was irradiated using 4 MeV proton irradiation at the Vanderbilt Pelletron. The effects of Single-Event Upsets (SEUs) on network accuracy and performance were negligible, reducing the accuracy of an MNIST-trained Convolutional Neural Network (CNN) with trinary weights from 98.82% to 98.78% after a 60 second exposure. Secondly, we also studied the effects of radiation on general analog in-mem

15. SUBJECT TERMS
Brain-inspired computing, IBM TrueNorth Neurosynaptic System, Spiking Neural Network (SNN), Proton Irradiation, Charge-Trap Transistor (CTT), Analog Synapse, Nonvolatile Memory (NVM), in-Memory Computing (iMC), 22nm FDSOI, 14nm FinFET, Total Ionizing Dose (TID), CTT-Hardware-based Inference Realistic Circuit Universal Simulator (CIRCUS)

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
Unclassified	Unclassified	Unclassified	SAR	81	Jacob Calkins

19b. TELEPHONE NUMBER *(Include area code)*
571-616-5946

Final Report: Award HDRTA1-17-1-0035

Project Title: Radiation Effects in Brain-Inspired Computing

PI Name:
Subramanian Iyer

Organization/Institution:
University of California, Los Angeles
and
Vanderbilt University

Table of Contents

Year 1…………………………………………………………….5

Year 2…………………………………………………………….17

Year 3…………………………………………………………….40

Year 4 (Extension period)……………………………………….58

Major goals of the project

The overall goal of the project is to advance the understanding of the radiation sensitivity of neuromorphic architectures and understand the implications of candidate emerging technologies.

Year 1

1. Fully analyze the impacts of radiation on brain-inspired computing systems emulated by von Neumann hardware. We will collaborate with IBM and perform studies on the IBM TrueNorth hardware to serve as a baseline in our studies.
2. Calibrate the simulation tool to understand the experimental data. Results will be used for evaluating the irradiation-induced damage of such systems.
3. Develop physics-based models for the radiation effects on brain-inspired computing systems emulated by von Neumann systems.
4. Examine the fundamental physics of charge-trap transistors (CTTs) as well as the effects of radiation on artificial synapses realized using CTTs.

Year 2

1. Design specific experiment vehicles based on the year one results to examine the radiation physics of the brain-inspired computing systems. Specifically, total dose effects and displacement damage (single-event upsets) will be studied. Radiation studies will be performed collaboratively between UCLA and Vanderbilt University.
2. Quantify the fundamental physics responsible for the radiation effects.
3. Develop a detailed physics-based model on the scaling of these brain-inspired computing systems under radiation environments.
4. Examine the fundamental physics of charge-trap transistors (CTTs) as well as the effects of radiation on artificial synapses realized using CTTs.

Year 3

1. Finalize the physics-based models on the radiation effects of brain-inspired computing systems.
2. Examine the design optimization strategy of brain-inspired computing systems for applications in radiation-harsh environments.
3. Evaluate a candidate technology for neuromorphic architectures, the charge-trap transistor. Memory arrays provided by UCLA will be tested by Vanderbilt University to investigate and compare the radiation performance to alternate memristor technology such as RRAM. The CTT array will be irradiated with x-rays to observe changes in performance with total ionizing dose. Further, the array will be exposed to ions to observe any susceptibility to ionization or displacement damage. The data collected in this task will provide insight into the radiation reliability of possible future implementations of neuromorphic architectures.
4. Utilize the IBM TrueNorth Compass simulator to construct configurations of the IBM TrueNorth chip to observe the generation of single-event effects (SEEs) in various components of the microarchitecture. We will determine how to purposefully program the network to direct transient faults to observable chip-level outputs. The chip will be exposed to an ion beam while running each of these configurations and the outputs will be monitored. Data collected in this task will indicate the relative contribution of microarchitectural components to the overall probability of error.
5. Utilize the IBM TrueNorth Compass simulator to construct configurations of the IBM TrueNorth chip to observe the propagation of faults through the neural network. We will determine how to purposefully program the network to determine how faults in neurons or axons affect the signaling of information. The chip will be exposed to an ion beam while running each of these configurations and the outputs will be monitored. The focus of these tests is to observe how the architecture responds to faults rather than observing a technology-specific result. The data in Task 2 will be used to separate the microarchitectural response.

Out-Years

1. Investigate how changes to neural networks affect radiation reliability.
2. Explore use more hidden layers within network and additional neurons to increase redundancy and determine to what extent does the inherent redundancy of networks provide reliability.

Year 1

Major Activities

- **Delid the TrueNorth chip**
  Due to the low penetration depth of the various sources at the Vanderbilt Pelletron, the TrueNorth chip encapsulation needed to be removed. A discrete, packaged TrueNorth chip was used to develop and test an etching recipe to delid the TrueNorth. After successfully delidding the standalone chip, a similar recipe was utilized to delid another TrueNorth chip attached to an NS1e board without damaging any of the other on-chip components, including FPGA and ARM core processor.

- **Develop an experimental protocol for irradiating the TrueNorth chip**
  The TrueNorth corelet programming environment (CPE) allows development and testing of different experimental programs. The environment can also be used to simulate the effect of SRAM bit flips by utilizing the TrueNorth Compass Simulator to simulate the performance of a modified TrueNorth model file; the neural network parameters are stored in SRAM, and the effect of random bit flips on each parameter were simulated by random fault injection into each parameter variable.

  The Vanderbilt Pelletron is capable of the following sources and corresponding maximum energies: protons (4MeV), alphas (6MeV), oxygen ions (14.3MeV), and chlorine ions (16.4MeV). Oxygen and chlorine ion penetration depths were too shallow to reach the active silicon area due to the ~8μm thick BEOL; thus, proton and alpha sources were selected for the initial set of experiments. A laptop with the programming environment was connected to the board via an ethernet and power supply feedthrough to the evacuated Pelletron chamber.

- **Experimentally verify radiation-induced upsets are observed on-chip**
  The intentionally fragile and deterministic "Randomly Connected Spiking Neural Network" (RCSNN) corelet (program)—typically used for post-fabrication defect detection and hardware verification—was provided by IBM and was used to verify that upsets were being induced. The structure of this corelet is such that if a parameter in the network is altered, the output of the system will change. The output of chip is monitored and if it no longer matches the known output spike file, the system is designed to halt. The estimated time to failure was less than 1 second, although the uncertainty in starting and stopping the Vanderbilt Pelletron is on the order of 3 seconds; therefore, an accurate time-to-failure is not obtainable. However, the intention of the corelet was to identify whether irradiation affects the SRAM stability, which was verified.

- **Study the effects of radiation on a variety of programs on the IBM TrueNorth**
  A variety of spiking neural network (SNN)-based corelets were developed and tested under irradiation, including convolutional neural networks (CNNs) trained to classify the MNIST dataset. Single-chip, deterministic CNN models were trained for performing classification on each of these respective datasets. A fresh model file was loaded onto the chip during each experiment. Prior to irradiation, the chip was programmed to evaluate all test patterns once to confirm that the output prior to irradiation was identical between runs. Afterward, the chip was iteratively exposed to periods of radiation. After each period of irradiation, the chip was evaluated and classification labels as well as output neuron magnitudes were recorded for each test pattern.

  The first set of experiments were run using the MNIST dataset. The MNIST database consists of 60,000 training and 10,000 test images of handwritten digits ranging from “0” to “9”. While the number of additional classification errors introduced by radiation were minimal, the relative increase in error rates for some runs were quite large—in particular, the error rate increased by 4.24% (1.18% to 1.23%) during one run with a 120 second total exposure.

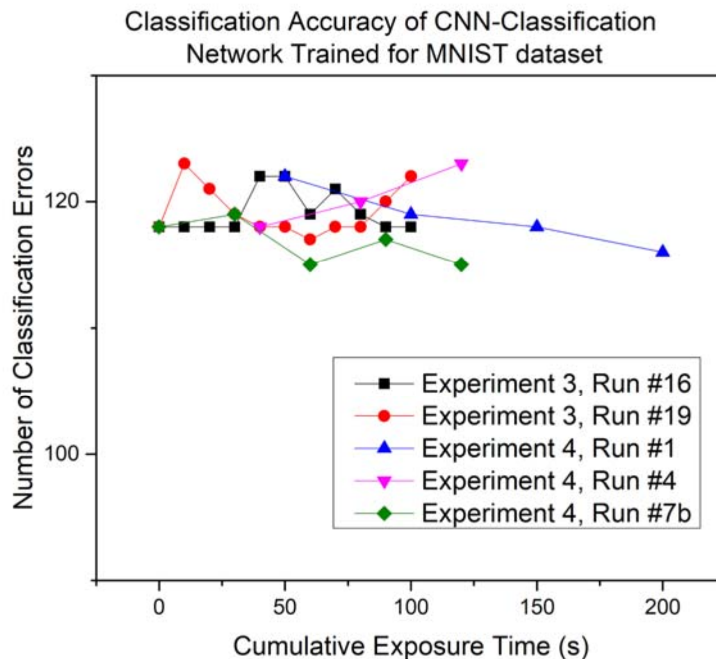


Figure 1. MNIST Classification Error Rate for 5 separate trials with variable exposure durations to 4MeV protons with a flux of 2.5×10^5 $cm^{-2}s^{-1}$. Total dataset consists of 10,000 test patterns. Worst-case error rate occurs during Experiment 4, Run #4 where the accuracy reduces from 98.82% to 98.77%—corresponding to a 4.24% increase in the error rate.

- **Analyze potential mechanisms for system crash and critical bits of the network**
  Several possible mechanisms of failure were identified for the network. The most severe mechanism of failure or change in network structure, currently identified, is the introduction of soft errors into particular bit locations (“critical bits”) which result in a

change in neuron addressing parameters. Each neuron on chip is configured to transmit output spikes to a fixed location, determined at compile time. If the address is modified at run-time due to radiation-induced upsets, the network graph is permanently altered. This neuron addressing parameter is configured to support a 16-chip TrueNorth system (known as the NS16e system) arranged on a PCB in a 4x4 array. In the most severe of circumstances, the address can be altered to transmit spikes off-chip, which results in system crash for single chip (NS1e) system. The system waits to transmit the data packet without receiving a "handshake" signal from another chip. After receiving several packets to be transmitted off-chip, the system crashes as it runs out of memory since it is designed to not delete untransmittable data packets.

- **Establish robust environment for wafer-scale study of CTT devices**
  We have configured and set up a Cascade S300 wafer prober for studying Global Foundries (GF) technology wafers, specifically 14LPP and 22FDX discrete devices. A set of analysis equipment was also purchased and configured, namely a Keysight B1500A semiconductor parameter analyzer and an E4980A precision LCR meter. The B1500A is equipped with special Waveform Generator Fast Measurement Units (WGFMU's) capable of 10ns pulse resolution and extremely fast measurement for such applications as BTI and charge pumping. This analyzer is also equipped with precise measurement units for fA-resolution, and couples well with the S300's ability to do temperature-dependent experiments using the cooled & heated chuck. Celadon ceramic probe cards were fabricated in order to probe the 25-pad test macros on the wafers, coupled with a Keysight switching matrix to multiplex the pads during testing and allow for wafer-scale reliability studies.

- **Establish programming characteristics of the CTT for use as an analog synapse**
  Various programming schema were developed for use in programming and verification of the CTT devices as implemented analog synaptic weights. The Pulsed gate Voltage Ramp Sweep (PVRS) method was verified on both 14LPP and 22FDX technologies. These results gave the NeuroCTT chip designers reasonable parameters for gate and drain voltages in order maximize the potential dynamic range of the device while preserving repeatability and reliability. Using only short-pulse programming with varied pulse voltages and total number of pulses, fine-tuning windows were established to provide a linear shift in the synaptic weight over as wide a tuning window as possible. The variation and stochastic characteristics of programming between like-devices was elucidated, as well as between different device types (regular threshold, ultra-low threshold, etc.).

- **Explore physics-based modeling of CTT devices**
  Developed qualitative model using Sentaurus TCAD software that demonstrates the programming and erase characteristics seen in experimental data, including finite retention time of traps. However, only a simple singly-ionized trap level was modeled at a discrete energetic level, rather than multiple trap levels distributed across a distribution of energies. The spatial distribution of traps was considered to be initially uniform

throughout the high-K dielectric layer, although the model also predicts the asymmetry of the trap distribution under an asymmetric bias condition (e.g., V_{GS} and $V_D \neq 0$, $V_S = 0$). Generation of new trap sites within the oxide will be incorporated into the model as well, in order to model the lack of saturation of the threshold voltage shift during programming. Trap identities (type & oxidation state), energies, spatial and energetic distributions, and capture/emission coefficients will be experimentally measured and incorporated into the model as well. A new software, Ginestra, was just recently acquired to develop an accurate model of trapping kinetics, and will be heavily used over the next year.

- **Design a neuromorphic system (NeuroCTT) using CTT-based analog "synapses"**
  The UCLA circuit design team completed a chip design in May 2018 using Global Foundries 22nm FDSOI technology. The chip uses 2 charge-trap transistors to differentially implement a synapse. The design implements a single, fully-connected layer network with 1024 inputs and 10 output neurons. If functional, it will be capable of classifying the MNIST dataset with a predicted classification accuracy greater than 95%.

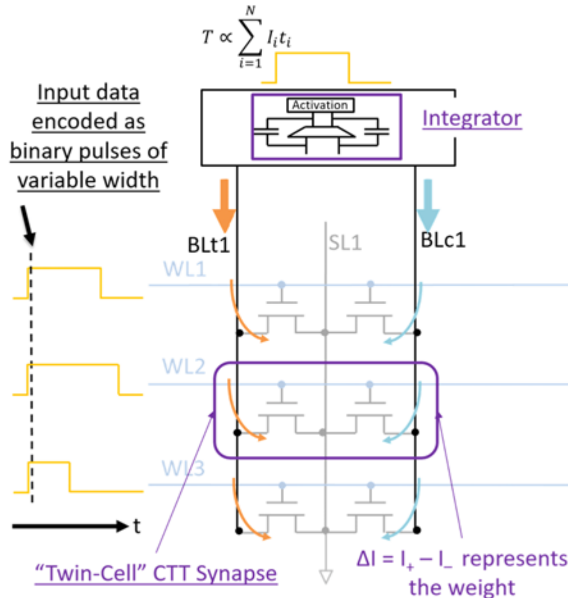


Figure 2. NeuroCTT concept of Single Neuron Implemented using charge-trap transistors (CTTs). Input patterns (e.g. input images for classification) are supplied to each word line as a digital-to-time converted, constant-amplitude voltage waveform. Synaptic weights are encoded in the devices by modifying the channel conductance (inversely proportional to threshold voltage) using the charge-trapping and charge-detrapping phenomenons. "Twin-cell" CTT synapse implements a single synapse differentially to increase precision and to allow for "zero" weights to be stored in the network as well. The neuron effectively computes the weighted sum of the inputs by integrating the charge collected across the capacitor inside the neuron: $q = \int I = \int \frac{dq}{dt} = \int \sum_i \sigma_i V_i$. Classification is performed by comparing the magnitude of the charge collected on the capacitor for each respective neuron (total of 10 in this design).

2.) Specific Objectives

- Analyze effect of single event upsets on the IBM TrueNorth by means of radiation
- Develop test corelets and train convolutional neural networks on the TrueNorth to study the change in classification accuracy as a function of total irradiation exposure
- Determine and investigate mechanisms for system crash, such as critical bit flips and sensitive analog, router, and other on-chip circuitry
- Simulate kinetics of how soft errors impact overall network classification accuracy
- Collaborate with Vanderbilt to develop a rigorous testing platform for brain-inspired computing systems, including GPU-based systems
- Develop physics-based models to predict the underlying effect of radiation on brain-inspired computing emulated by von Neumann systems
- Characterize the charge-trap transistor (CTT) for use as artificial synapses, and examine the fundamental physics of the CTT
- Design a neuromorphic system using analog synapses and irradiate the system to study the impacts of radiation on a fully-functional analog-based system

3.) Significant Results

- **IBM TrueNorth Radiation**

  Corelets (programs) such as the “Randomly Connected Spontaneously Connected Neurons” (RCSSN) are able to immediately detect soft errors and halt. The RCSSN corelet is not a functional corelet in the sense that it doesn’t provide any useful computation. It was originally intended for determining if a chip had any fabrication defects by comparing the output of the network to a known output “spike” file. Similarly, it is also useful for fault detection in general. This corelet halted within the first couple of seconds of radiation during each trial we conducted.

  In general, the IBM TrueNorth is capable of performing more useful calculations including a spiking neural network (SNN) implementation of a convolutional neural network (CNN) for image classification using the IBM EEDN (Energy Efficient Deep Neuromorphic Networks) training algorithm. Using the IBM TrueNorth Compass Simulator, we were able to modify a trained model file, introduce “bit errors” into each of network parameters, and simulate the effect of each of these “soft” errors. Initial simulation results across the entire parameter space indicate that as many as 0.1% of the bits in the on-chip 410Mb SRAM may be affected by soft errors before seeing any significant changes in classification accuracy for the SRAM—ignoring radiation effects on sensitive analog circuitry, the spike routing network, and other non-SRAM radiation effects.

  Simulation results were verified by a running trained model file for the MNIST dataset while simultaneously irradiating the TrueNorth NS1e board inside a vacuum chamber with an ion beam. The board was irradiated with 4MeV protons with a flux of ~2.5×10^5

$cm^{-2} s^{-1}$. Classification errors vs. cumulative exposure time for MNIST were shown previously in Fig. 1. ΔError Rate for each run vs. the cumulative exposure time is summarized in Figure 3. The worst case ΔError Rate was 4.237% and occurred after 120 second exposure during MNIST Run #4—the total number of classification errors increased from 118 to 123 out of 10,000 total test patterns.

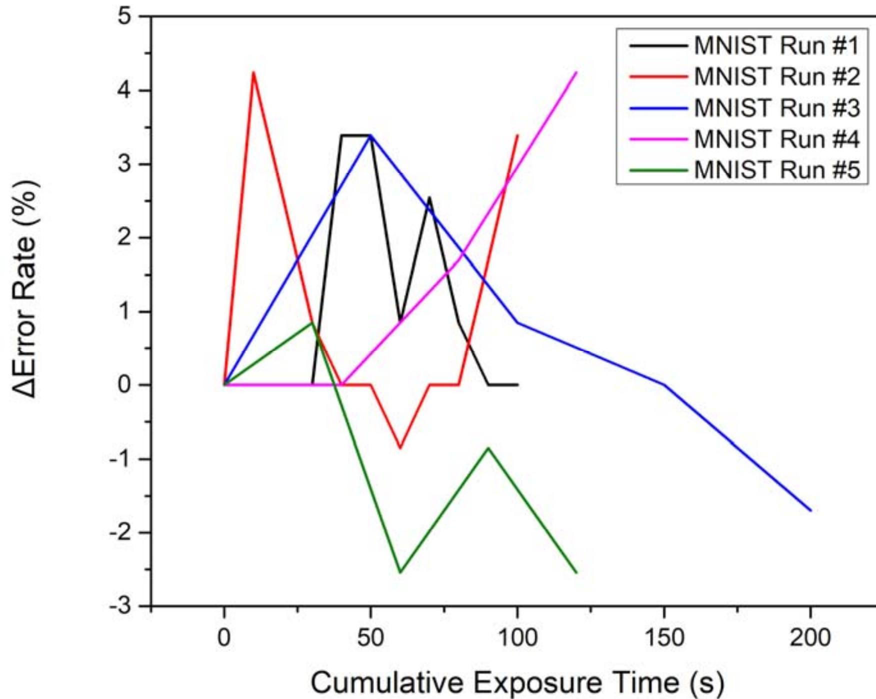


Figure 3. Summary of ΔError Rate vs. Cumulative Exposure Time for each MNIST trial. Some trials observed a relatively large ΔError Rate since the classification accuracy was initially very high (98.82%). Results for 5 different experimental runs with 4MeV protons at a flux of ~2.5 x 10^5 $cm^{-2} s^{-1}$ on the MNIST dataset are shown.

- **Charge-Trap Transistor (CTT) Device Characterization**
The 14LPP and 22FDX devices' programming characteristics were measured in order to determine the most efficient and effective methodologies for use as an analog memory component in brain-inspired computing applications. Using the PVRS method, a window of drain and gate bias conditions was determined for achieving the maximum dynamic range in threshold voltage shift while preserving device reliability. On 22FDX devices with 20 nm gate length and 170 nm gate width, gate pulses of 10 milliseconds were applied with successively larger gate biases from 0 to 3V in 50mV steps. A drain voltage of 1.4V and a gate voltage of 2.8V yielded a maximum threshold shift of 325mV while still maintaining low gate leakage and erasability. Although gate leakage did increase by roughly two orders of magnitude for this substantial threshold voltage shift, it is still 5 orders of magnitude below the drain current, quite usable in application. This method explores the effect of lattice temperature (self-heating) on the efficiency of the trap filling mechanism. The ability to perform multiple programming and erase cycles was verified, although the window of threshold voltage still shifts upwards over time; that is, complete

detrapping is not occurring and new trap states are being generated over successive cycles. Another method of programming is to use the same V_{DS} and slowly increase V_G and the number of pulses over successive cycles; that is, apply 50 pulses at (V_G, V_{DS}), 100 pulses at (V_G + 100mV, V_{DS}), and so on. This methodology avoids the saturation in the threshold voltage shift and provides a more linear increase in the threshold voltage.

Physical modeling of the CTT was performed using a TCAD simulation platform, Sentaurus. Capture of carriers in the high-κ dielectric layer was modeled considering elastic and phonon-assisted tunneling, and the emission rate was determined using the principle of detailed balance. By considering a uniformly distributed single-level trap (E_T = 1.8 eV below conduction band), the simulated programming and erase mechanisms behaved closely to the experimental results. The trapped electron density initially jumps as the gate pulse is applied, some detrapping occurs during an idle period, and the erase pulse displays a characteristic time constant and inability to fully depopulate the traps, as seen in experimental characterization. The trapped electron spatial distribution also demonstrates only minor dependence on source-gate-drain bias asymmetry, indicating nearly uniform charge injection across the majority of the channel length. This is due to the minor difference in the vertical electric field from one source to drain, which dominates the tunneling rate. Temperature-dependance of the lattice has also been incorporated, and the model agrees with experimental results that the charge trapping mechanism is more efficient with more device self-heating. This is due to the strong temperature dependance of electron energy and the phonon-assisted tunneling process.

4.) Key Outcomes

The overarching objective of this work is to determine the robustness and resiliency of brain-inspired computing platforms implemented in non-von Neumann architectures, as compared to traditional von Neumann architectures. By simulating the effect random permutation of stored parameters has on the overall TrueNorth classification accuracy, an estimate of the minimum number of radiation-induced bit flips required to measurably affect the output was determined. After training MNIST networks on the chip and establishing a baseline, the chip was delidded and we began conducting initial radiation testing using the Vanderbilt Pelletron accelerator. The MNIST network proved robust, as it is already highly clustered; the error rate fluctuated roughly 0.04% around the initial 1.18% misclassification rate. One interesting characteristic observed in the MNIST test performed was that the distribution of types of classification errors changed with radiation; however, the overall number of cases there were errant prior to irradiation but correct after exposure was offset by those that were correct prior to but errant after exposure. Additional experiments are required determine if negligible changes in error rates hold for more complex datasets such as CIFAR and CIFAR100.

Initial wafer characterization of the 14LPP and 22FDX GF devices provided substantial information regarding the variation in devices and their type, as well as crucial data to guide the design of the NeuroCTT chip. Two programming methodologies were considered and provided

the optimal conditions for maximal dynamic range and reliability in both 14LPP and 22FDX, as well as the development of a method to achieve linear fine-tuning of threshold voltage with the widest cycling window. The initial model developed incorporates various aspects of the dielectric charge trapping, capturing the qualitative programming and erase characteristics, dependance on lattice temperature, and influence of trapped charge spatial variation. The model will soon be expanded to incorporate bias-induced trap generation, the effect of multiple discrete trap levels, and consideration of trap energetic distribution within the bandgap of the high-K dielectric.

The UCLA circuit design team also generated a neuromorphic chip as a proof-of-concept using Global Foundries 22nm FDSOI technology in May 2018. It utilizes the charge-trap transistor (CTT) to implement analog synapses. The circuit operates as an analog fully-connected layer computation engine, useful for large fully-connected neural network computations. If functional, it is predicted that it will be able to classify the MNIST dataset with classification accuracy greater than 95%. The NeuroCTT chip will allow us to perform radiation experiments on a candidate technology for performing neural network computations outside of the typical digital domain.

What do you plan to do during the next reporting period to accomplish the goals?

We will proceed with the current plan with more radiation studies, simulations, and modeling to be conducted during the next reporting period. Radiation studies will be conducted on the IBM TrueNorth, CPU, and GPU systems to study the kinetics of soft errors and their effects on neural networks—with a comparison between different physical architectures. In addition, exact mechanisms for system crash will be studied with an emphasis on understanding the underlying architecture(s). We will verify experimental verified by detailed, fault-injection simulations for each respective hardware.

We will also conduct experiments on discrete charge-trap transistors (CTTs) using Global Foundries 22nm FDSOI and 14nm LPP to study the effect of radiation (e.g. change in conductance) on analog "synapses" implemented using single high-k-metal-gate logic transistors. Radiation effects on memristors (e.g. RRAM)—an alternative candidate technology for implementing analog "synapses"—will be evaluated as well and compared against CTT-device performance under radiation.

Additional conference and journal publications will be targeted to further disseminate information to communities of interest.

Appendix – CTT Physics and Simulation

The TCAD simulation platform, Sentaurus, was used for modeling the 14LPP and 22FDX CTT devices and their charge trapping dynamics. This tool will aid in developing physical models for understanding and predicting the effects of radiation on the device for brain-inspired computing systems.

A high-κ dielectric layer is incorporated in the gate stack, which is where the majority of trapped carriers will reside. The capture and emission of carriers in this high-κ layer is primarily due to elastic and inelastic (phonon-assisted) tunneling mechanisms. In general, the trap occupation rate, f, is determined by the capture and emissions rates:

$$\frac{df}{dt} = (1-f)C - fE$$

The capture rate, C, can be determined by the different tunneling processes, and the emission rate, E, can be obtained from C and the principle of detailed balance. A higher capture rate indicates a smaller trapping time constant, leading to more easily filled trap states. For the elastic tunneling process, the initial and final states will have the same energy, as demonstrated in Fig. 1 below. As the gate bias increases, the crossing point between the trap level E_T and Fermi level E_F in the high-κ layer will become closer to the interfacial layer and silicon channel. Under this condition, the trapping will occur closer to the interface due to direct tunneling with a shorter tunneling distance, smaller barrier height, and higher capture rate. The capture rate can be modeled[1] as [1]:

$$C_{el}(E,z) = D(E)f(E)W_{el}(E,z)$$

where: $$W_{el} = \frac{2\pi}{\hbar}T(E,z)^2\delta(E_T - E)$$

In these equations, $D(E)$ is the density of states, $f(E)$ is the Fermi occupancy function, and $T(E,z)$ is the tunneling rate calculated by the WKB approximation:

$$T(E,z) = e^{-2\int_0^z dz\sqrt{\frac{2m}{\hbar^2}(V(z)-E))}}$$

Here, z is the distance between the channel and trap, and $(V(z) - E))$ represents the barrier height along the tunneling path.

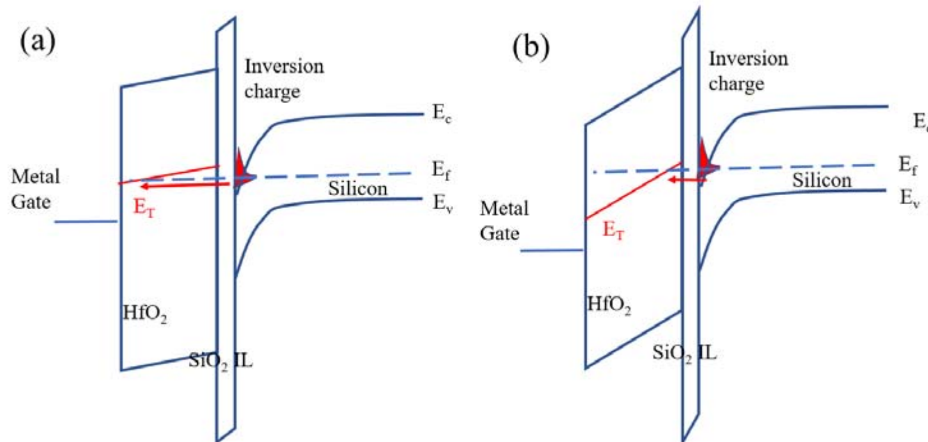


Figure 1. Elastic (direct) tunneling process at (a) low gate bias (b) high gate bias.

The process of phonon-assisted tunneling is displayed in Fig. 2. In order for a carrier in the channel to tunnel to a trap of a lower energy than itself, it must emit one or more phonons of energy $\hbar\omega$. Therefore, the capture rate is a function of phonon energy, the number of emitted phonons (or energetic difference

between the carrier and trap), and the tunneling probability. This process is also quite thermally-dependent, due to the larger range of trap energies accessible with phonon emission. Using the single energy phonon approximation[2], the transition rate between a carrier and the trap state can be found to be:

$$W_{ph} = \frac{\pi}{\hbar} S \left(1 - \frac{p}{S}\right)^2 G(E_T, \hbar\omega) T(E)^2$$

In this equation, p is the number of emitted phonons, S is the Huang-Rhys factor, and $G(E_T)$ is a function of trap energy, temperature, and phonon energy[3].

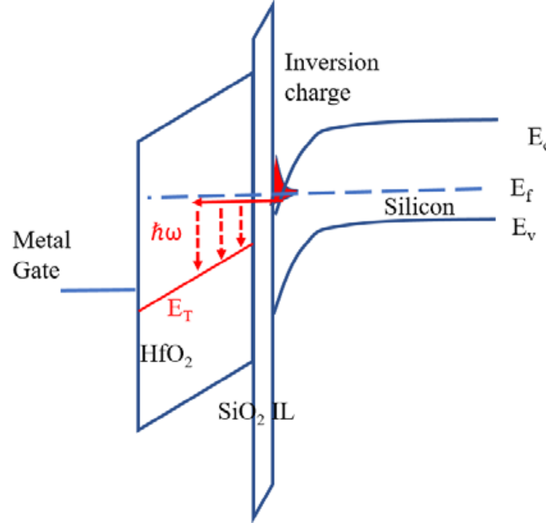


Figure 2. Inelastic tunneling assisted by phonon emission with energy ħω.

The non-local trapping model[3] described above was used in the TCAD simulation tool to capture the initial trapping physics of the CTT. The trap energy level, spatial distribution, and density can be defined across the two dielectric layers. Trap volume density, interface trap density, capture cross-section, and tunneling effective mass can be used as fitting parameters when calibrating to experimental data. The simulated gate stack consists of SiO_2/HfO_2 with respective thicknesses of 1 nm and 1.5 nm, and the device channel length was 20 nm. Traps are defined as single-level acceptor-type with $E_T = 1.8\ eV$ from the conduction band of HfO_2 ($\sim 0.2\ eV$ above the silicon conduction band during flat-band conditions.) The trap distribution is defined to be uniformly throughout the HfO_2 layer with a volume density of $2x10^{20}\ cm^{-3}$.

A program & erase cycle was simulated, shown below in Fig. 3. During the program and erase periods, a different characteristic time constant can be seen for both, and the erase period does not fully depopulate all traps, qualitatively capturing the behavior seen in experiment. In Fig. 4, the variation in trapping efficiency with lattice temperature can be seen, along with temperature dependence of ΔV_T saturation. Since the tunneling rate strongly depends on the applied electric field, trapping does not occur until a certain point and saturates within a $\sim 0.5\ V$ window.

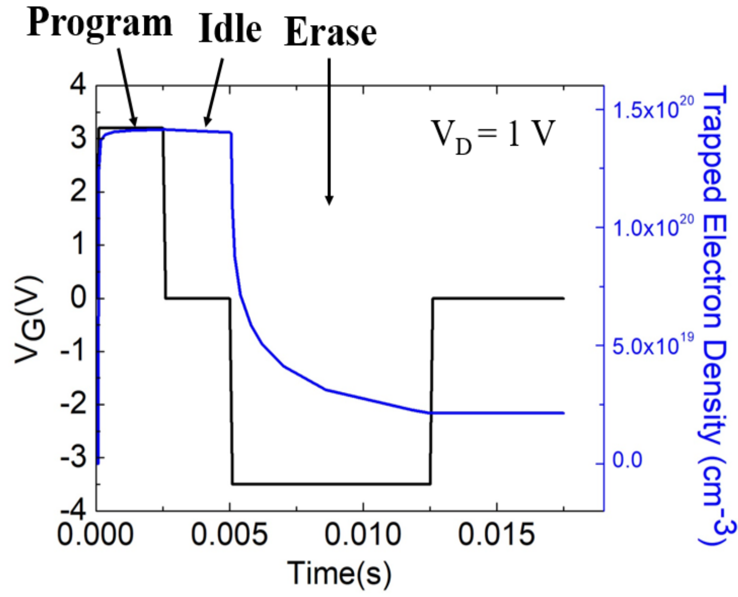


Figure 3. Simulated program & erase cycle with finite trap retention time

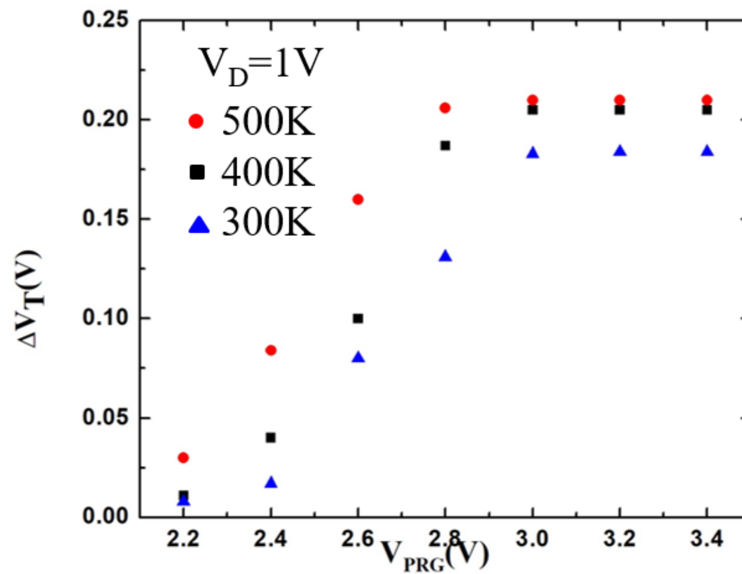


Figure 3. Variation of trapping efficiency with lattice temperature.

References

[1] F. Jiménez-Molinos et al., *Journal of Applied Physics*, vol. 91, no. 8, pp. 5116–5124, 2002

[2] Sentaurus Device User Guide 2011, p. 405

[3] A. Palma *et al*., *Physical Review B*, vol. 56, no. 15, pp. 9565–9574, 1997.

Year 2

Major Activities

- **Establish robust environment for wafer-scale study of CTT devices**

  We have configured and set up a Cascade S300 wafer prober for studying Global Foundries (GF) technology wafers, specifically 14LPP and 22FDX discrete devices. A set of analysis equipment was also purchased and configured, namely a Keysight B1500A semiconductor parameter analyzer and an E4980A precision LCR meter. The B1500A is equipped with special Waveform Generator Fast Measurement Units (WGFMU's) capable of 10ns pulse resolution and extremely fast measurement for such applications as BTI and charge pumping. This analyzer is also equipped with precise measurement units for fA-resolution, and couples well with the S300's ability to do temperature-dependent experiments using the cooled & heated chuck. Celadon ceramic probe cards were fabricated in order to probe the 25-pad test macros on the wafers, coupled with a Keysight switching matrix to multiplex the pads during testing and allow for wafer-scale reliability studies.

- **Establish programming characteristics of the CTT for use as an analog synapse**

  Various programming schema were developed for use in programming and verification of the CTT devices as implemented analog synaptic weights. The Pulsed gate Voltage Ramp Sweep (PVRS) method was verified on both 14LPP and 22FDX technologies. These results gave the NeuroCTT chip designers reasonable parameters for gate and drain voltages in order maximize the potential dynamic range of the device while preserving repeatability and reliability. Using only short-pulse programming with varied pulse voltages and total number of pulses, fine-tuning windows were established to provide a linear shift in the synaptic weight over as wide a tuning window as possible. The variation and stochastic characteristics of programming between like-devices was elucidated, as well as between different device types (regular threshold, ultra-low threshold, etc.).

- **Explore physics-based modeling of CTT devices**

  Developed qualitative model using Sentaurus TCAD software that demonstrates the programming and erase characteristics seen in experimental data, including finite retention time of traps. However, only a simple singly-ionized trap level was modeled at a discrete energetic level, rather than multiple trap levels distributed across a distribution of energies. The spatial distribution of traps was considered to be initially uniform throughout the high-K dielectric layer, although the model also predicts the asymmetry of the trap distribution under an asymmetric bias condition (e.g., VGS and VD $\neq 0$, VS $= 0$). Generation of new trap sites within the oxide will be incorporated into the model as well, in order to model the lack of saturation of the threshold voltage shift during programming. Trap identities (type & oxidation state), energies, spatial and energetic distributions, and capture/emission coefficients will be experimentally measured and incorporated into the model as well. A new software, Ginestra, was just recently acquired to develop an accurate model of trapping kinetics, and will be heavily used over the next year in conjunction with Sentaurus.

- **Explore physics-based modeling of RRAM devices**

  Preliminary work has been done using Ginestra to study the physics of the RRAM HfO_x device, in order to understand the material system and relations to the CTT as device physics

as well. More precisely, the effects of radiation-induced lattice damage to the RRAM device are considered, involving the creation of oxygen vacancy V_O and mobile O^- ion pairs. These vacancies alter the filamentary formation process and result in a net difference of defects, which in turn alter the SET and RESET process curves significantly. This would be detrimental to the device in analog operation, as the expected transition curve from R_{ON} to R_{OFF} is modified. However, digital operation could still be functional, as the R_{ON}/R_{OFF} ratio did not change appreciably in simulation.

- **Develop experimental radiation setup and irradiate CTT devices**

  A GlobalFoundries 22FDX wafer was quartered and diced to isolate several device test macros onto singulated dies, which were then bonded out to high speed packages using Au wire. Approximately 50 macros each of various types were provided (single devices with width scaling, devices with length scaling, metal-oxide-metal capacitors, and ganged-gate devices) and will be continually used for testing into the next year. Preliminary irradiation studies were performed at Vanderbilt using the ARACOR X-ray irradiator. Studies indicated that significant trapping was occurring in the buried oxide (BOX) of the devices, which is highly coupled to the channel in FDSOI nodes. The induced trapping in the BOX caused a significant V_{th} shift in different devices, which could potentially be mitigated by employing circuitry to supply a feedback back-bias to the devices.

- **Develop experimental radiation setup to irradiate commercial RRAM chip**

  A commercial RRAM chip from Fujitsu has been acquired and configured with a stand-alone testing setup for radiation experiments. The chip has CMOS overhead and the RRAM array is behind decoder logic, but the physical resistances and associated digital values can be read and modified to conduct various experiments. Of interest is how the device characteristics of a large RRAM array will shift with radiation dose, and how robust an off-the-shelf commercial RRAM chip is with the integrated CMOS overhead.

- **Investigate mitigation techniques for analog networks**

  Both hardware and software mitigation techniques can be employed to reduce the impact of ionizing radiation to analog network. In hardware, feedback circuitry can be employed to monitor for drift in device characteristics over time and correct the shift back to the intended value. A twin-cell architecture of the CTT was employed in the neuromorphic analog chip taped out last year using 22FDX. In this architecture, analog weights are encoded as the differential conductance between two devices, which mitigates both process variation and BOX irradiation, since both cells will shift by the same amount. Software methods for mitigation are also being investigated, particularly the modification of the stochastic gradient descent method to incorporate device variation. This software mitigation technique modifies learning algorithms, and thus can be employed on any analog neural network, regardless of the type of device used in the architecture.

2) Specific Objectives

- Establish robust electrical characterization system for measuring CTT characteristics and determine behavior as an analog memory
- Develop preliminary charge trapping models with experiments to understand the physical mechanisms governing analog memory behavior, such as weight granularity, stored weight retention, and endurance

- Explore initial physical mechanisms governing the operation of RRAM cells using TCAD simulation and analyze a commercial RRAM chip's resiliency
- Irradiate single CTT devices to understand the impact of x-ray radiation on individual device characteristics and ultimately the stored network weights
- Analyze the impact of the shift of weights on analog neural networks of various complexity using simulation

3) Significant Results

- **Establish characteristics of CTT as an analog memory**
  A Keysight B1500A parameter analyzer equipped with high-speed waveform generator / fast measurement units was used in conjunction with a Cascade S300 wafer prober to make electrical measurements. 22FDX wafers from GlobalFoundries were provided for testing, and a neuromorphic inference engine was taped out in 22FDX, which contained several test macros as well. To program these devices, a non-zero V_{DS} is applied while pulses of V_{GS} are supplied to the gate, causing channel current to flow and electrons to tunnel into oxygen vacancy states within the HfO_2 gate dielectric. This alters the electrostatic control of the gate over the channel, causing a substantial shift in the device threshold voltage V_{th}. Similarly, V_{GS} pulses of negative polarity can be supplied with no source-drain bias to de-trap charge (erase) and restore the shifted V_{th}. This has been explored previously for application as a digital memory, but the programming methodology must change for an analog device.
  In order to realize a continuous range of analog weights (device conductances) which can be reversibly fine-tuned, a method of programmed known as Pulsed Voltage gate Ramp Sweep (PVRS) was used. This method utilizes pulses of increasing V_{GS} magnitude to achieve a linear shift in V_{th}, as seen in Fig. 1(c) below.
  Devices of various threshold voltage implants and widths were subjected to PVRS programming and erasing. During programming, the V_{DS} of the devices was 1.2 V, and gate pulses were supplied from 1.4 V to 2.7 V in increments of 100 mV; from 2.4 V to 2.7 V, the step-size was reduced to 50 mV. Additionally, at each V_{GS} level three pulses were supplied before moving to the next V_{GS} bias level, for a total of 51 pulses for the entire program sequence in one cycle. For erasing the device, V_{DS} was set to zero and V_{GS} pulses of equal magnitude and opposite polarity to that of the program sequence are applied. The characteristics of programming and erasing across four cycles on six devices are shown below in Fig. 2(a), with a comparison to constant amplitude pulses in 2(b). It is apparent that the V_{th} shift is linear and reversible using this method, and avoids substantial jumps in threshold voltage, as well as the saturating behavior.

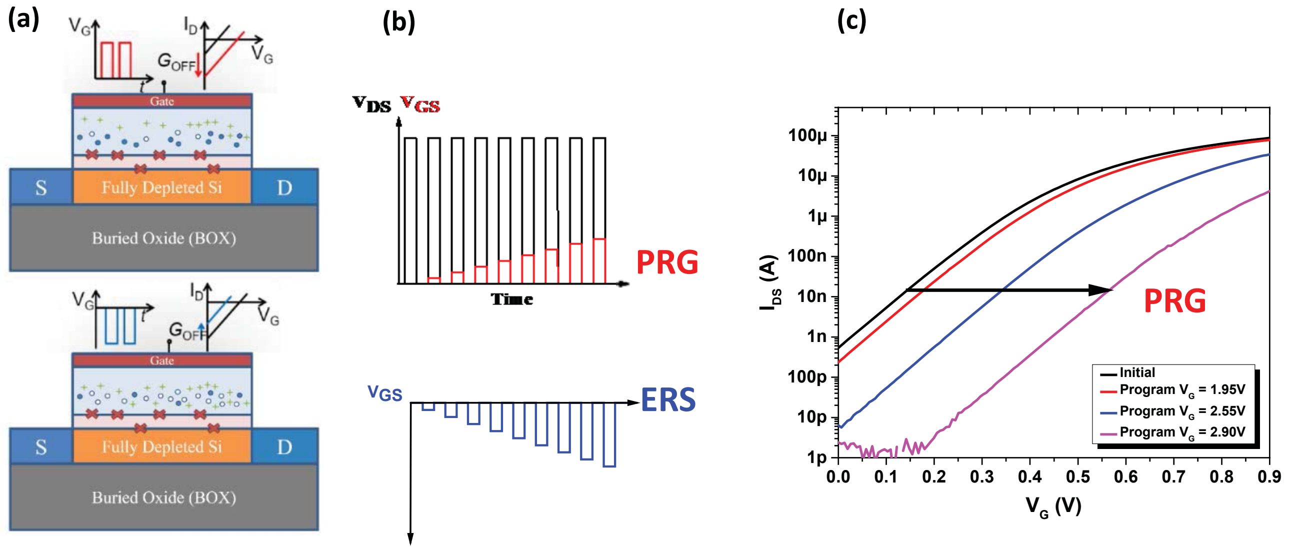


Figure 1: ***(a)*** *Pulses of positive V_{GS} trap electrons and pulses of negative V_{GS} de-trap electrons;* ***(b)*** *PVRS methodology of programming, with increasing V_{GS} pulse magnitude;* ***(c)*** *Larger V_{GS} biases cause more significant V_{th} shifts*

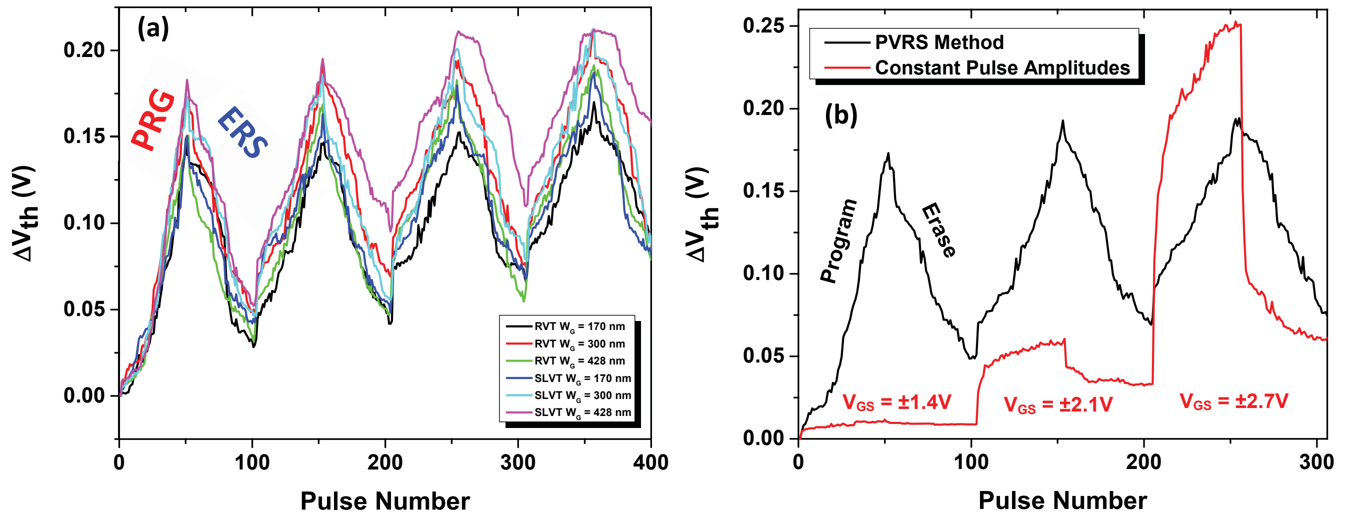


Figure 2: ***(a)*** *Four cycles of programming and erasing of several devices with different threshold implants and widths;* ***(b)*** *Comparison of constant pulse programming with the PVRS method, which can achieve linear reversible shifts in V_{th}*

The twin-cell architecture effectively measures the differential conductance across two CTT devices, corresponding to different device currents. This not only great improves process variation mitigation, but allows for bipolar weights, a necessary component of an effective analog neural network. In Fig. 3(a) below, the differential currents (conductances) of an array of twin-cells are measured after initial fabrication to demonstrate the random stochastic variation across an array of cells. The cells are randomly assigned bipolar target differential currents linearly mapped between -600

nA and 300 nA in order to show the programming accuracy in Fig. 3(b). The 3σ variation in device threshold voltage is roughly 50 mV, and the standard deviation of programming achievable is within 1.0% of the target value. The traps responsible for this V_{th} shift can be either energetically deep or shallow, with shallow traps causing issues for retention and programming accuracy. Luckily, the mapping of V_{th} relaxation over time has been characterized, and the shift can be anticipated based on the target V_{th}. Since this shift can be predicted, the device can be correspondingly overprogrammed to anticipate the shift that will occur, increasing programming accuracy.

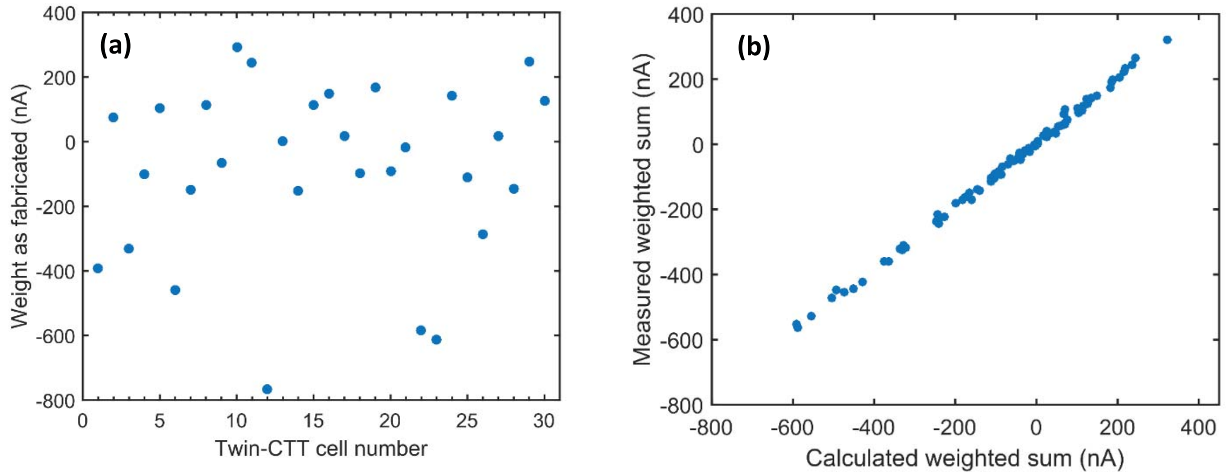


***Figure 3: (a)** As-fabricated weights (differential currents) of an array of twin-cell CTT's; **b)** The array's weights after programming along a linear differential current mapping from -600 nA to 300 nA*

- **Develop preliminary charge trapping model for CTT operation**
  The TCAD software Ginestra was used to investigate the underlying physical mechanisms governing the kinetic charge trapping within the CTT. This software accounts for numerous mechanisms of charge transport, and can simulate charge trapping characteristics accurately. The effect of a short pulse of V_{GS} = 2.4 V bias with an applied V_{DS} = 1.4 V is shown below in Fig. 4(a), demonstrating the characteristic electron emission time constant after the V_{GS} bias is removed and shallow traps are free to tunnel out of the oxide traps. The corresponding energy band diagram of the gate cross-section before and after the applied pulse are shown in Fig. 4(b); the majority of trapping occurs near the SiO_2-HfO_2 interface, as expected. The trap occupancy is plotted in color from 0 to 1, and interestingly enough it is also the shallow states near this interface that de-trap the fastest at room temperature, leading to variation in the programmed device weight. Traps further from this interface are more difficult to fill, but remain trapped for a longer time. The electron capture time constant admittedly still needs to be tuned, as the trapping response is nearly-instantaneous with the applied bias, whereas in reality this time constant is on the order of microseconds. A spatial visualization of the trapped charge density within the gate stack is displayed in Fig. 5, showing more trapped charge near the SiO_2-HfO_2 interface and slightly more source-side trapping.

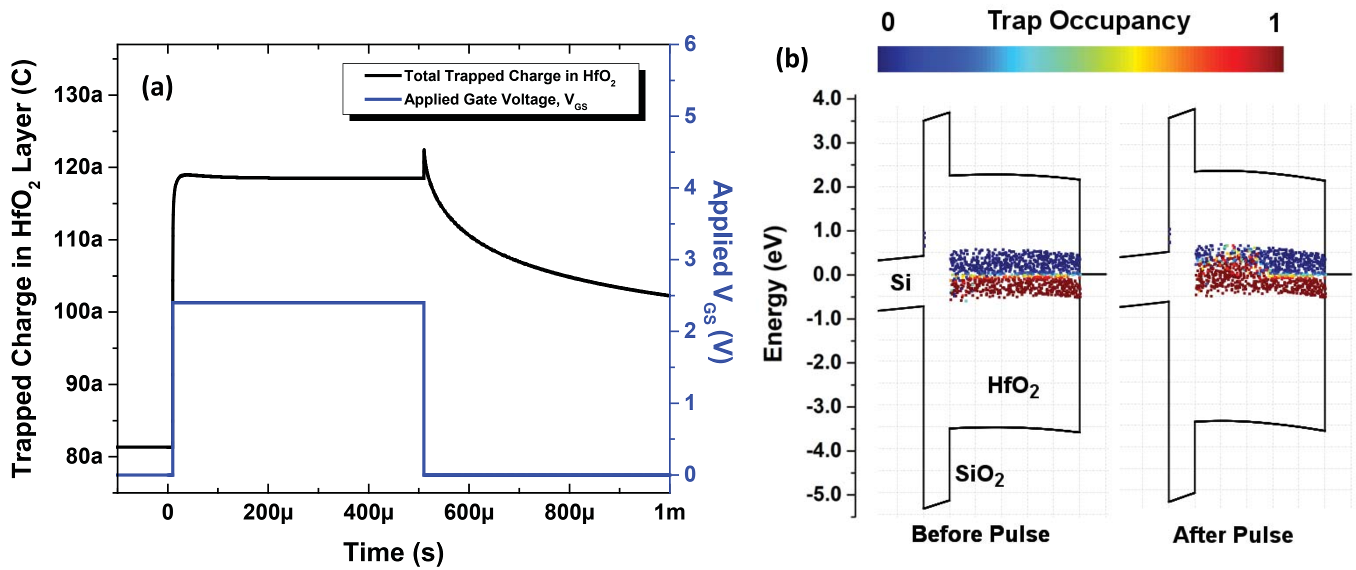


Figure 4: *(**a**) Simulated response of total trapped charge within the HfO_2 gate dielectric to an ideal applied gate pulse of 2.4 V, V_{DS} = 1.4 V; (**b**) Energy band diagram of the gate cross-section before and after the applied pulse, with trap occupancy colored. V_O states within the dielectric are displayed within the HfO_2*

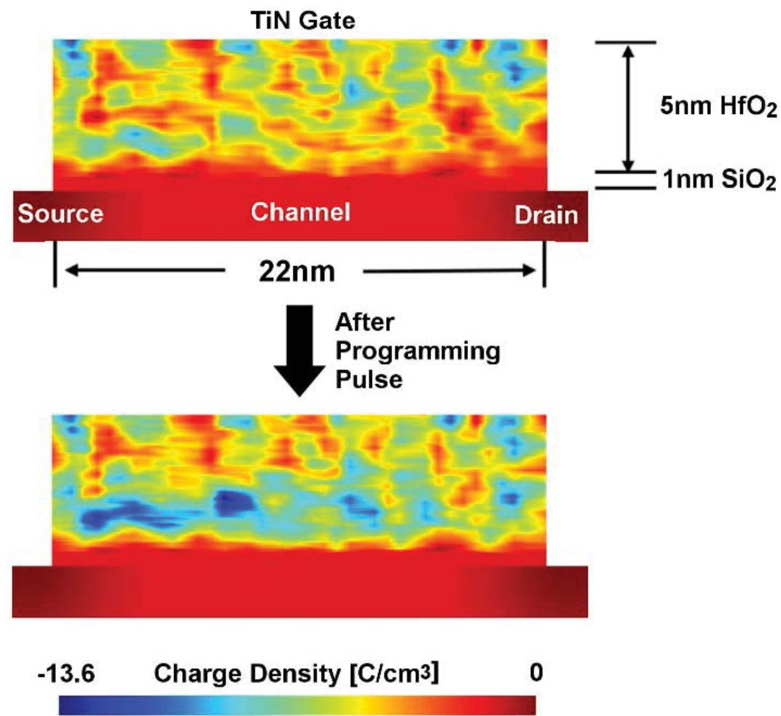


Figure 5: *Cross-section of the gate stack before and after programming pulse. The charge density is plotted in color, with blue being the higher density of electronic charge. Most charge is trapped near the SiO_2-HfO_2 interface, and slightly more source-side trapping is seen due to the larger magnitude of vertical field at the source-side.*

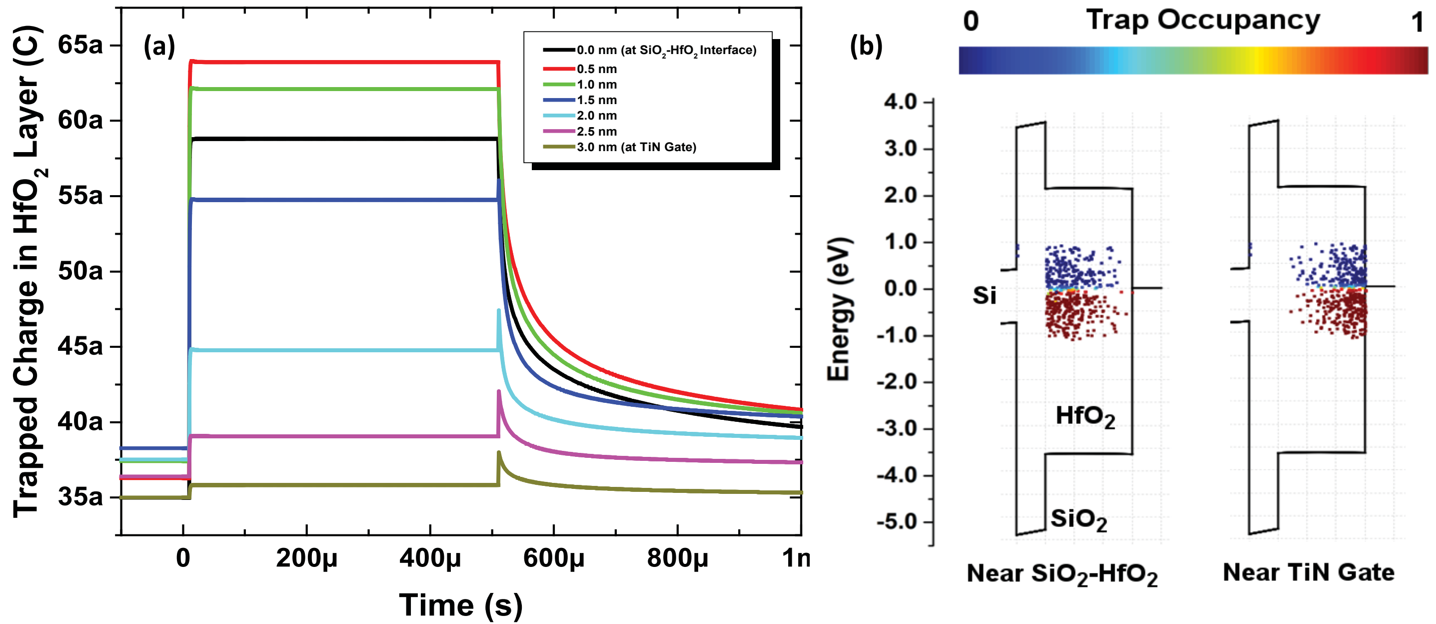
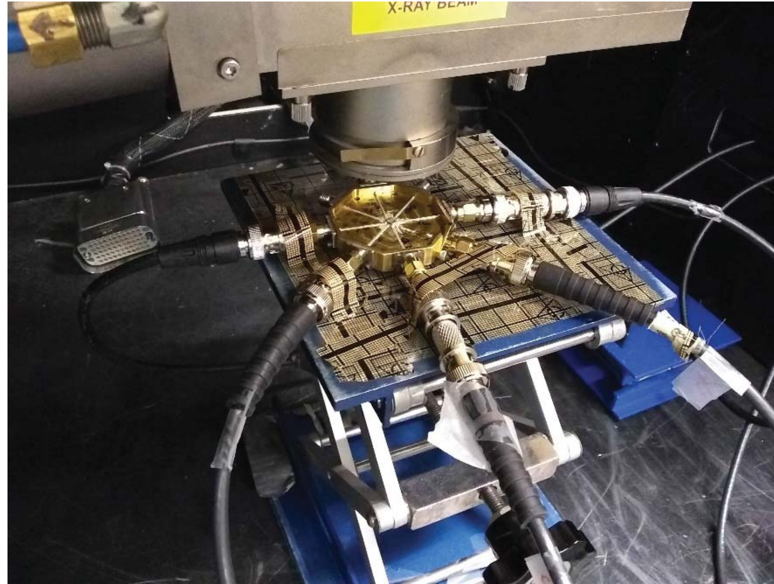


Figure 6: ***(a)*** *Total trapped charge within the HfO_2 gate dielectric during and after the ideal applied gate pulse of 500 µs (shown in Fig. 4(a));* ***(b)*** *Associated energy band diagram cross-section of the gate stack, displaying two different Gaussian-distributed trap distributions simulated in* ***(a)***

To investigate the effect of spatial trap location on the amount of trapped charge and relative retention, the trap distribution was modified to be Gaussian-distributed in the Z-direction, towards the gate. As the mean position of the distribution shifted from the SiO_2-HfO_2 interface towards the TiN gate, the amount of trapped charge during and after the pulse significantly decreased, as seen above in Fig. 6(a). The amount of trapped charge during and after the pulse was maximized at a mean distance of 0.5 nm from the interface. Again, the associated energy band diagram cross-section of the gate is also shown in Fig. 6(b), in order to visualize the different trap distributions. Understanding the impact of spatial trap variation on the trapping and retention of the CTT is important for exploring the impacts of radiation-induced damage in the oxide.

- **Initial CTT radiation utilizing ARACOR X-ray irradiator**
  To understand the impact of x-ray radiation on as-fabricated CTT's, various test macros previously described were bonded out and irradiated in the ARACOR X-ray irradiator system at Vanderbilt. The experimental package setup is shown below in Fig. 7, with a schematic of the induced trapped charge from the impinging photon. The trapped charge within the buried oxide (BOX) has a much large volume to trap within, compared to the gate stack. Furthermore, in FDSOI nodes the effect of trapped charge within the BOX is more pronounced due to the thinner silicon region, leading to stronger capacitive coupling to the channel potential. The devices were grounded and irradiated as-fabricated to establish a baseline of radiation tolerance. The dose rate was 30.26 krad (SiO_2) per minute, achieved by applying -35 kV and 30 mA to the X-ray tube. A total dose of 500 krad was supplied to each test device, with

device I_D-V_G characteristics taken at 50 krad, 100 krad, 200 krad, 300 krad, and 500 krad increments.



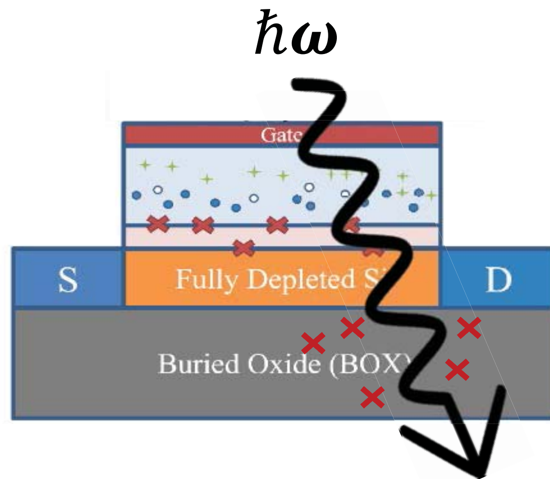
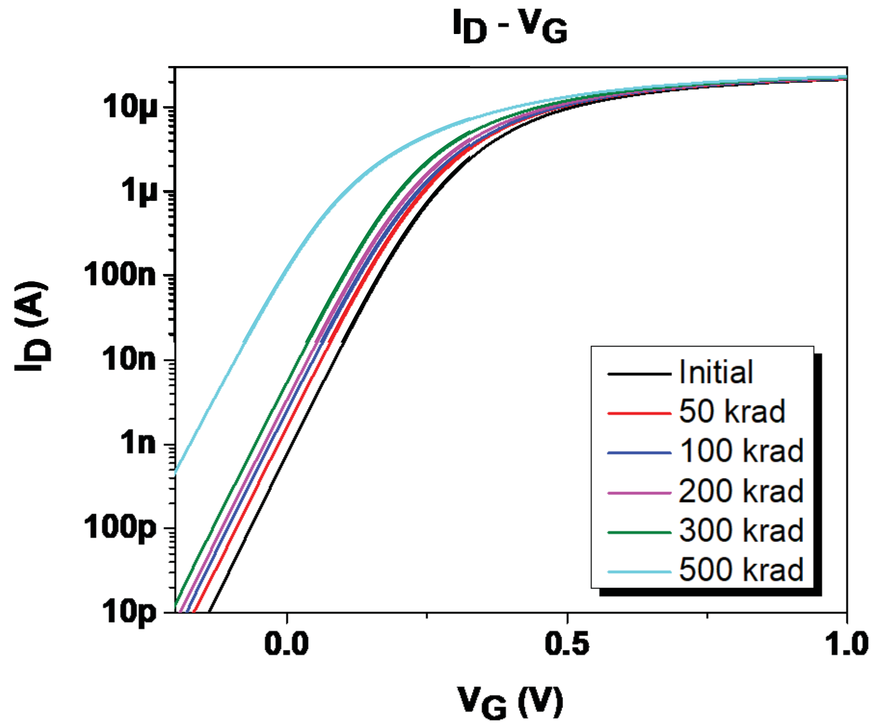


Figure 7: *Experimental test package and diced chip containing CTT test macros inside the ARACOR X-ray irradiator, with a schematic image of the track of trapped charge from the impinging photon of energy* $\hbar\omega$

Various devices were tested and have all shown similar responses under a grounded bias condition. All devices had a nominal electrical gate length of 22 nm and equivalent oxide thickness of 1.25 nm; one device was a "super-low V_{th}" variant with a gate width of W_G = 120 nm, while the other two were "high V_{th}" variant with gate widths of 120 nm and 170 nm. The I_D-V_G characteristics of the devices are shown below in Figs. 8-10, with the associated I_{DS} shift measured at V_{GS} = 300 mV and V_{DS} = 50 mV as a function of total dose, which varies linearly in the linear biasing regime.



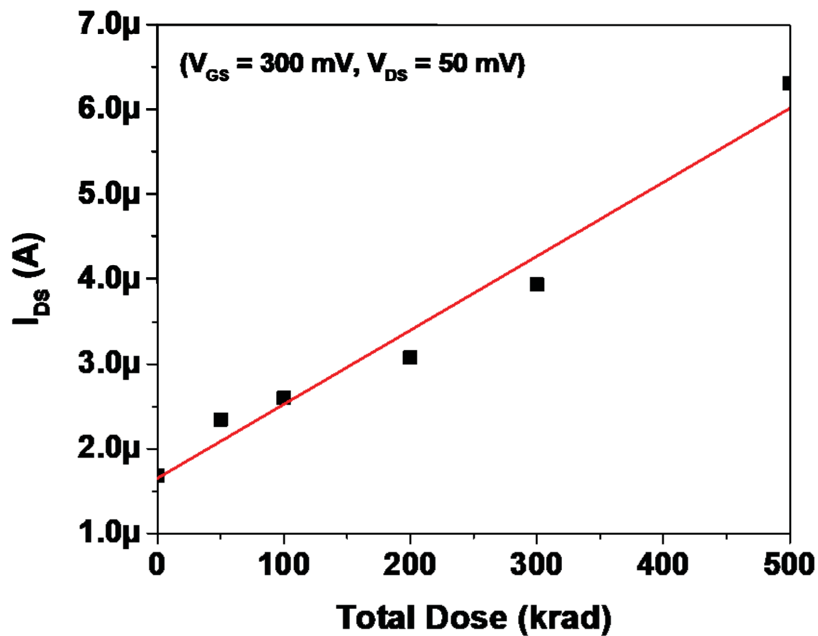
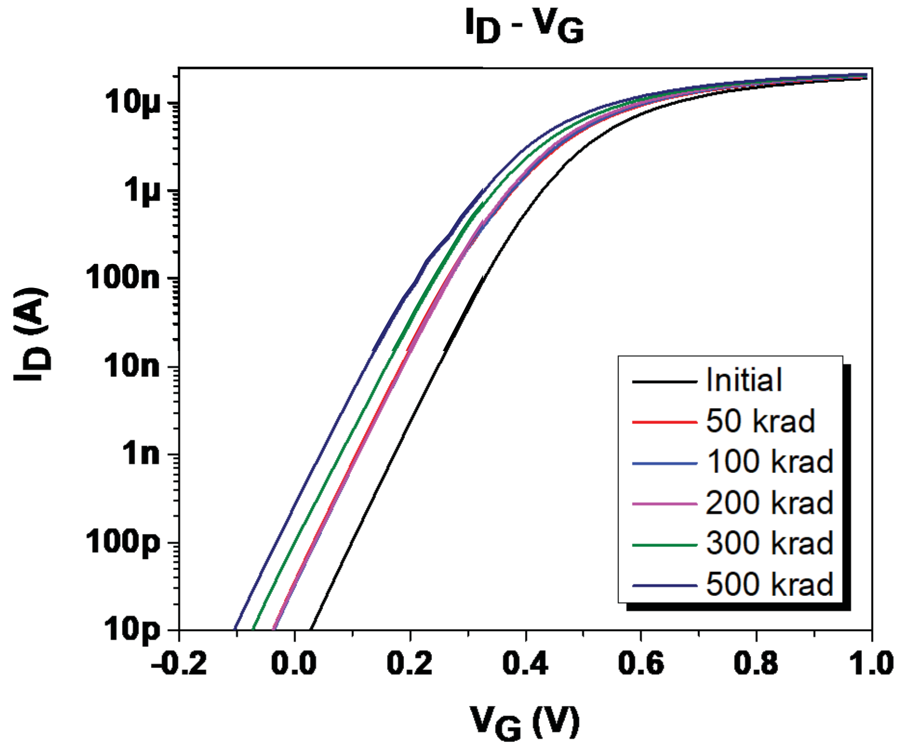


Figure 8: ***(top)*** *I_D-V_G characteristics of a CTT as a function of total dose up to 500 krad. Device characteristics: L_G = 22 nm, W_G = 120 nm, EOT = 1.25 nm, "super-low V_{th}" variant;* ***(bottom)*** *I_{DS} measured at V_{GS} = 300 mV, V_{DS} = 50 mV as a function of total dose*



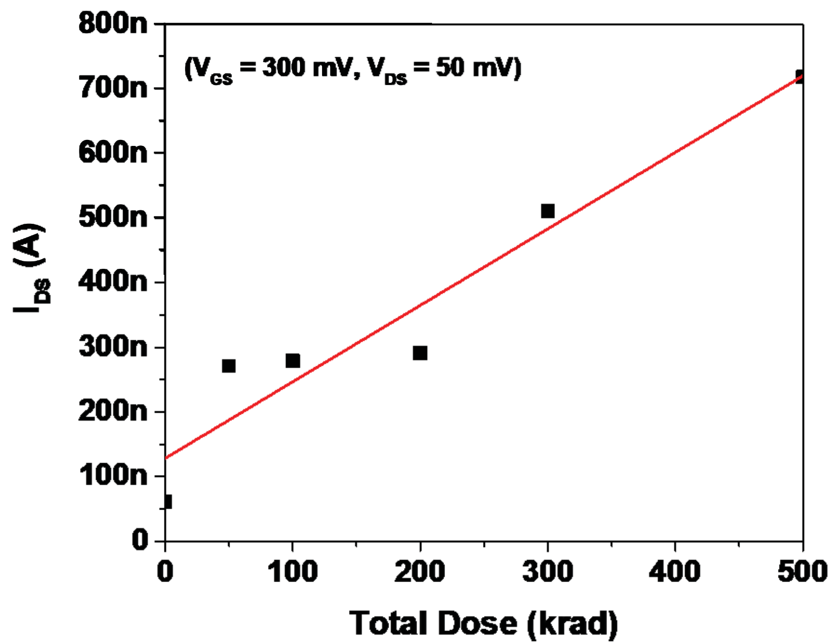
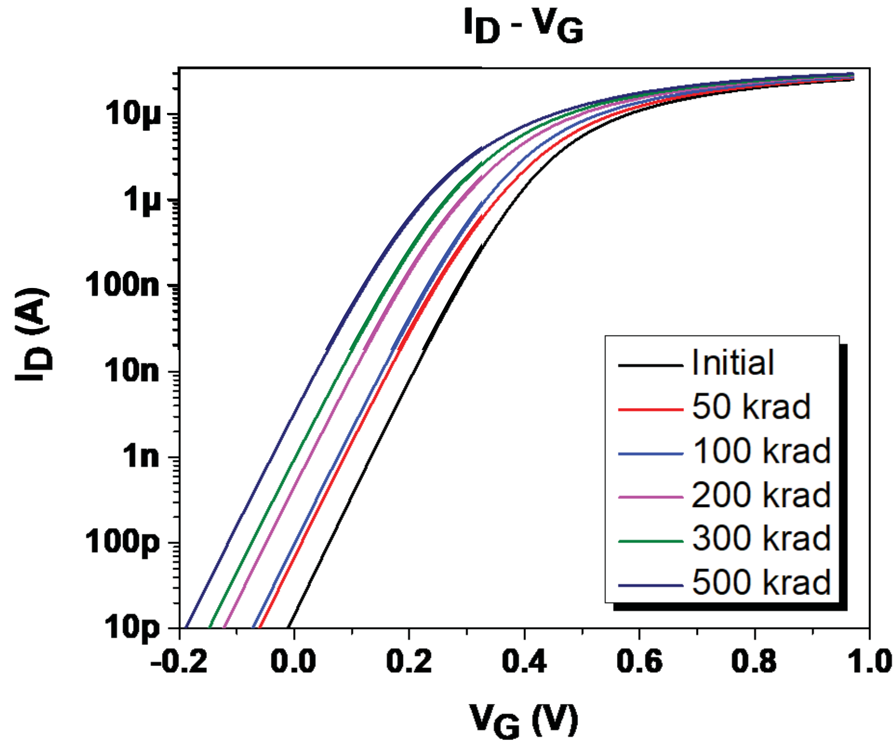


Figure 9: ***(top)*** *I_D-V_G characteristics of a CTT as a function of total dose up to 500 krad. Device characteristics: L_G = 22 nm, W_G = 120 nm, EOT = 1.25 nm, "high V_{th}" variant;* ***(bottom)*** *I_{DS} measured at V_{GS} = 300 mV, V_{DS} = 50 mV as a function of total dose*



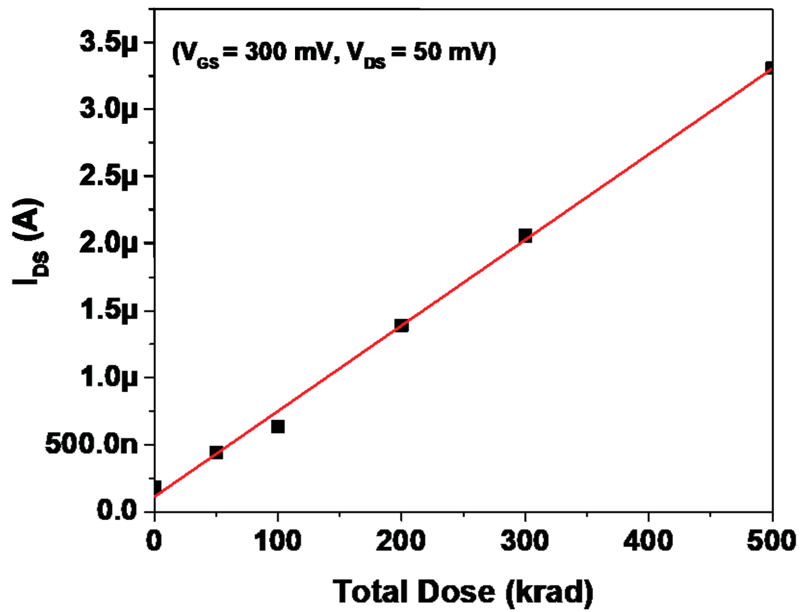


Figure 10: ***(top)*** *I_D-V_G characteristics of a CTT as a function of total dose up to 500 krad. Device characteristics: L_G = 22 nm, W_G = 170 nm, EOT = 1.25 nm, "high V_{th}" variant;* ***(bottom)*** *I_{DS} measured at V_{GS} = 300 mV, V_{DS} = 50 mV as a function of total dose*

In order for an analog memory to be resilient in an analog network, gate integrity must remain consistent. As expected in the unbiased case, there was no apparent TID damage to the front gate dielectric stack, as indicated by no apparent change in the I_G-V_G characteristics with total dose, shown below in Fig. 11(a). For comparison, the gate leakage I_G of a device discussed previously is plotted as a function of the program and erase characteristics in Fig. 11(b). It is apparent that the amount of gate leakage increase induced in this case is a direct function of the trapped charge (V_{th} shift) in the front gate oxide, and increases cycle-to-cycle for equivalent V_{th} values, indicated trap generation. This is not observed in the irradiated sample with similar magnitude V_{th} shifts, indicating that the trapped charge is located primarily in the BOX rather than the front gate dielectrics.

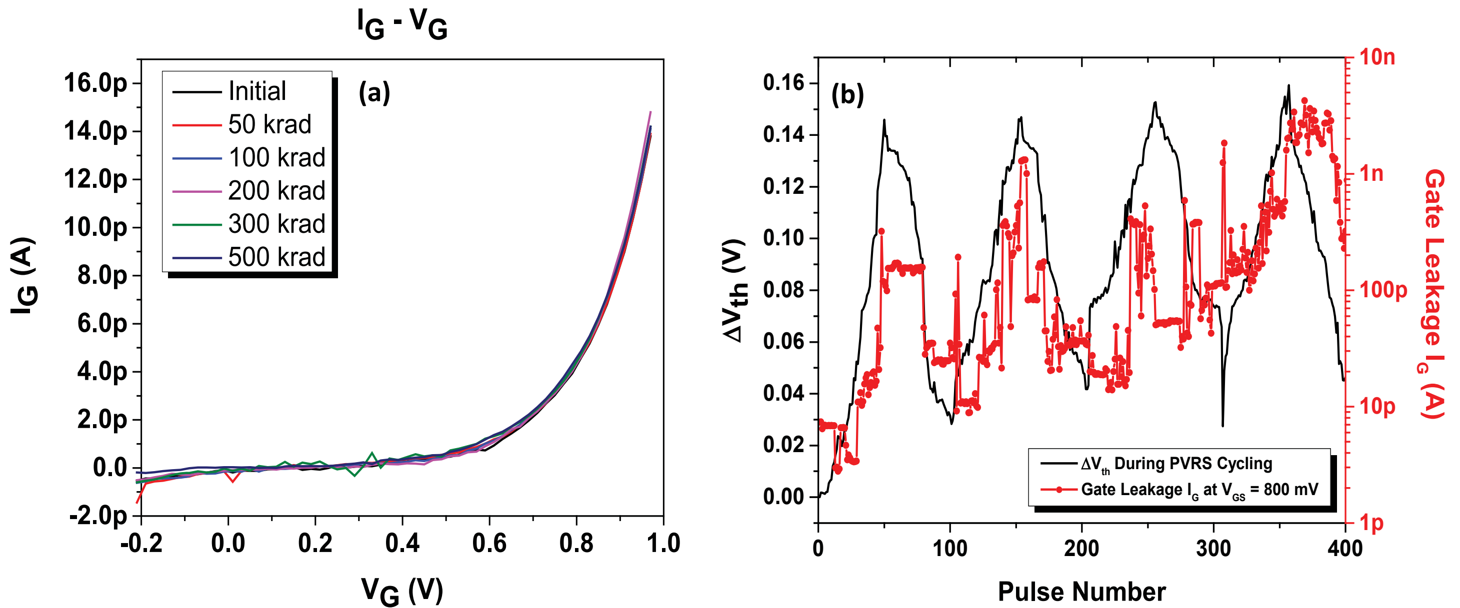


Figure 11: ***(a)*** *I_G-V_G characteristics as a function of total dose in the "super-low V_{th}" device – these characteristics are consistent across all tested devices, with no noticeable increase in I_G with total dose.* ***(b)*** *For comparison, the measured gate leakage I_G plotted during the program and erase cycling of a device discussed previously in the report. Note the correlation of the gate leakage magnitude with threshold voltage shift, which increases for the same V_{th} cycle-to-cycle, indicating trap generation*

The change in the I_{DS} at a particular bias as a function of the total dose across all devices is shown below in Fig. 12. The device with a lower initial V_{th} has the highest rate of change with the total dose, and as expected the larger device of the same threshold type has a higher rate of change than its counterpart that starts at nearly the same I_{DS}. Devices that are fabricated with "super low V_{th}" provide an analog neural network with a wider dynamic weight range, which may be useful if the sensing circuitry is weak or leakage current is of concern to the designer. The absolute magnitude of change in the device weight (conductance) will be larger than the "high V_{th}" counterpart, but the relative change compared to the dynamic range of used weights may likely be similar. Further investigation is ongoing to establish a more statistically-significant study on the resiliency differences in device type and geometry.

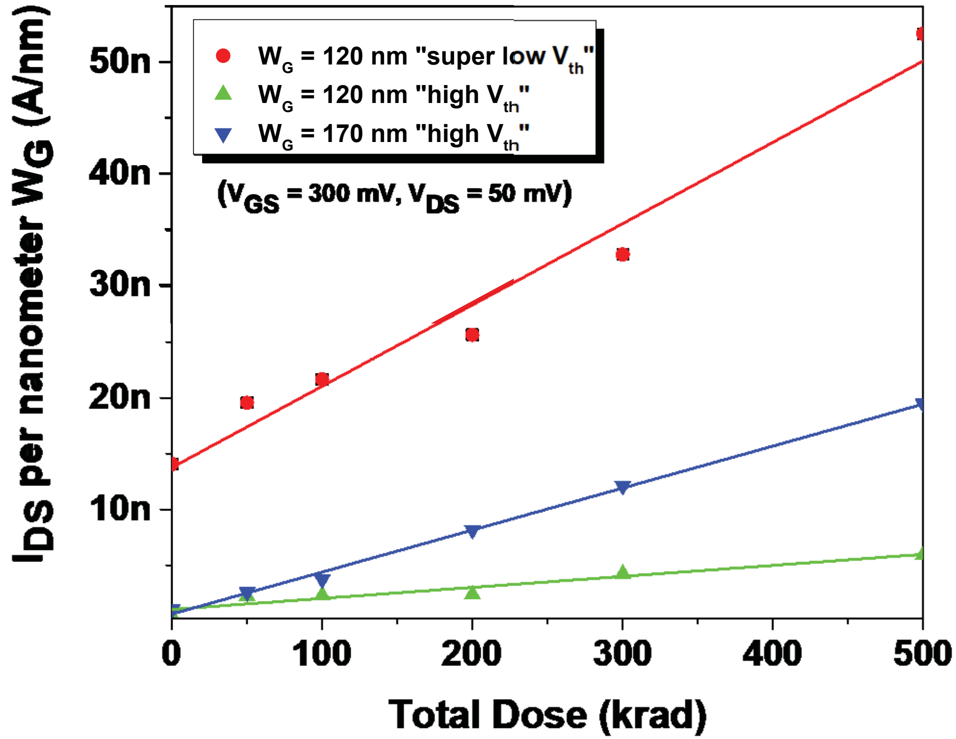


Figure 12: *I_{DS} (normalized to W_G) of the various tested devices as a function of total dose*

The amount of trapping in the BOX can estimated by calculated an equivalent series capacitance to the channel, including the front gate capacitance, the FD silicon capacitance, and the BOX capacitance. And equivalent expression for this is:

$$C_{tot} = \left(C_{ox}^{-1} + C_{si}^{-1} + 2C_{box}^{-1}\right)^{-1}$$

The factor of 2 in the C_{box} term in this case comes from the fact that the BOX thickness was assumed to be half and the equivalent sheet charge density would be deposited at $\frac{t_{BOX}}{2}$ in the center of the BOX. The magnitude of trapped charge between the three devices appears to increase linearly with total dose as expected in Fig. 13 above with the largest device containing the highest charge density. More data is being collected to explore the statistical variation as well.

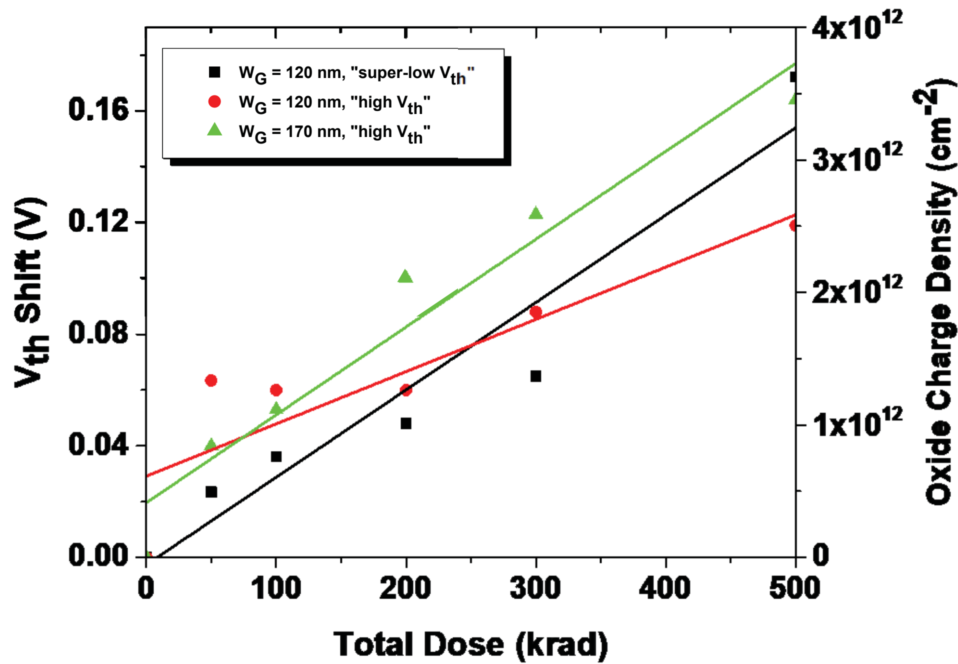


Figure 13: *Threshold shift V_{th} in the devices as a function of total dose, and the corresponding amount of trapped charge density distributed in the BOX*

- **Proposed mitigation techniques to enhance radiation resiliency**
  While the result of such large threshold voltage shifts may be unappealing, this effect should only appear at this magnitude in the FDSOI nodes; if, say, partially-depleted SOI (PDSOI) or bulk technologies are employed for analog CTT devices, it is expected that their radiation resiliency will dramatically increase. Alternatively, analog circuit tricks can be employed to compensate for the shift in the I_D-V_G characteristics – 22FDX has an attractive design feature known as back-biasing, altering the channel potential and thus the device V_{th}. If a back-bias is applied to the device, the characteristics can be shifted to undo the alteration. The amount of shift achievable is shown below in Fig. 14, both by simulation and experiment. In Fig. 14(a,b), the I_D-V_G characteristics of a "super-low V_{th}" device are altered by applying positive gate bias up to 1.5 V. The rate of change of the V_{th} with applied back-bias is roughly -88 mV/V. The simulated structure in Fig 14(c) demonstrates similar magnitude of shift, although the predicted rate of change is slightly higher than seen in experiment.

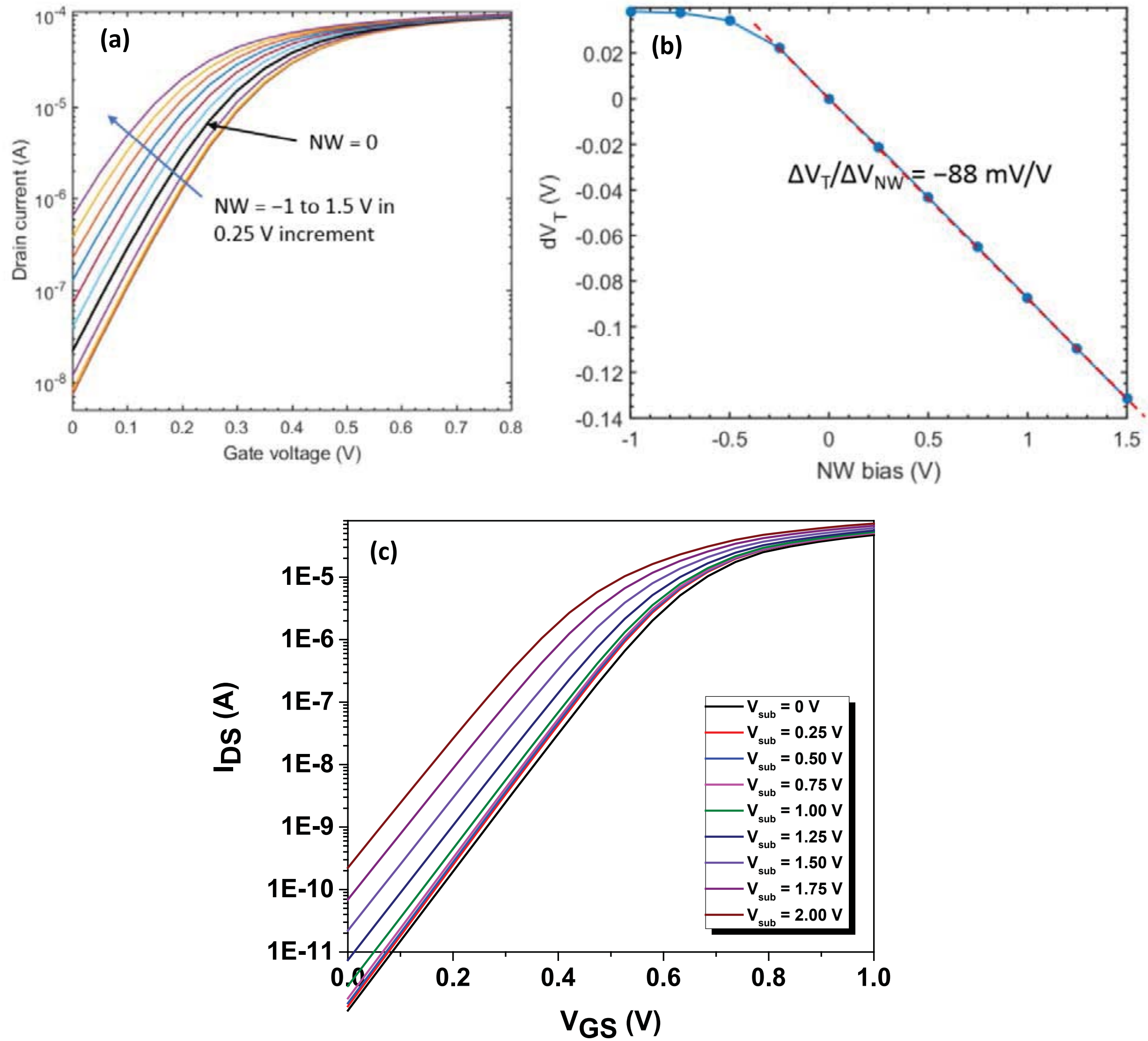


Figure 14: ***(a)*** *Measured characteristics of a "super-low V_{th}" device as a function of back-bias;* ***(b)*** *Rate of change in V_{th} with applied back-bias potential;* ***(c)*** *Simulated characteristics of an equivalent device as a function of back-bias*

Another method of mitigation being investigated by our colleague Zhe Wan and Professor Vwani Chowdhury is the modification of the stochastic gradient descent (SGD) algorithm for training neural networks to incorporate the anticipated device (weight) variance into the learning. By doing so, the algorithm will not assume the weights to be perfect 32-bit floating point digits, but rather to have some variation. This effectively causes the algorithm not to search for the global maximum steepest descent, but to look for a shallower descent. If the weight value is perturbed then, the impact on the system loss function and ultimately the system accuracy or performance will be mitigated. Using this methodology, they have shown that the network resiliency can be dramatically increased for a very small penalty on the maximum achievable network accuracy, shown below in Fig. 15. A typical analog device can have accuracy variation from 5 to 20%, which would typically destroy the

network accuracy, but can provide only marginal loss of accuracy using this novel training algorithm. This can be extended to any analog memory device as well, making it a powerful methodology.

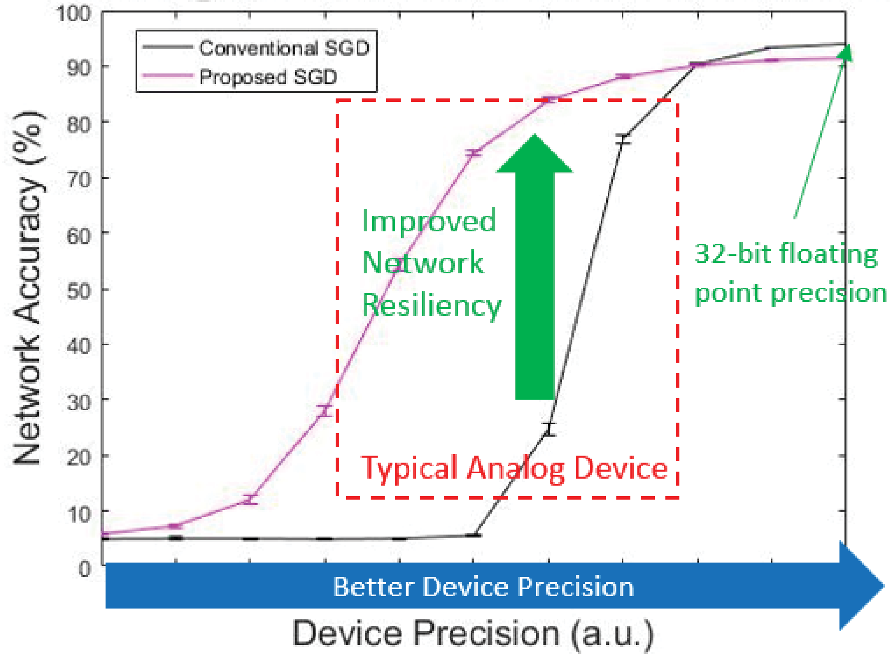


Figure 15: *Modified stochastic gradient descent (SGD) algorithm for network training is able to achieve greatly improved network resiliency to the device variation. Radiation-induced changes in the network will effectively appear as a loss of programmed device accuracy, indicating this method can be used for increasing radiation resiliency in analog neural networks as well*

In order to estimate the impact of the induced device variation on the network accuracy, various networks were trained assuming some level of weight variation. Networks were trained to classify handwritten digits from the MNIST database (Fig. 16(a)), which showed little degradation in network accuracy (2% worst case) even for up to a 50% randomness of weights in both layers of the network. On a more difficult problem, such as the CIFAR-100 dataset, the Wide ResNet network accuracy was analyzed (Fig. 16(b)) as a function of device variation for network depths of 16 and 28 layers. The accuracy of this network falls off more rapidly, and network depth only does so much to mitigate against the variation; at a typical 5% analog weight variation, network accuracy can degrade up to 20%. Finally, the last fully connected layer of GoogLeNet was analyzed in Fig. 16(c), with the Top1 and Top5 accuracies plotted as a function of device variation. While this network is perturbed less by the effect of device variation, it is not nearly as insensitive to weight variation as the example in Fig. 16(a). In any of these cases, it is important to develop methods for mitigation against as much as-fabricated and induced device variation as possible in order to maintain accurate and reliable networks.

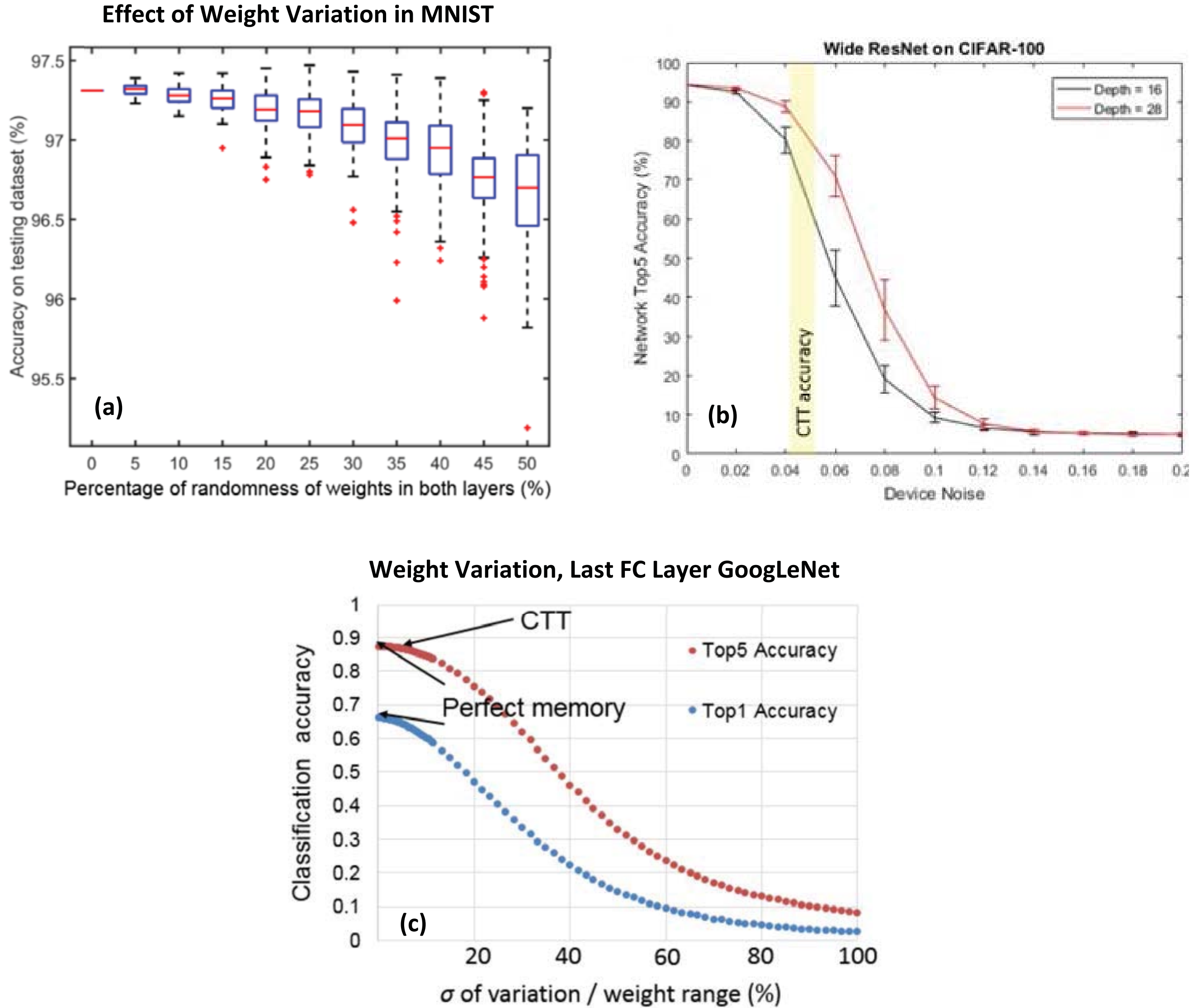


Figure 16: *(**a**) Effect of stochastic weight variation on a network trained to recognize handwritten digits from the MNIST dataset; (**b**) Effect of variation of Wide ResNet on the CIFAR-100 classification dataset; (**c**) Effect of variation on the last fully-connected layer of GoogLeNet on Top1 and Top5 accuracies*

- **ReRAM Simulation Setup**
  The behavior of RRAM and the operation under the effect of radiation has been simulated using the Ginestra package. Ginestra can perform atomic level simulation that describe the atomistic forming-set-reset cycle of ReRAM operation by solving charge transport, thermal dissipation, and oxygen vacancy generation / recombination / diffusion self-consistently. Individual oxygen vacancies or oxygen ions can be randomly generated, and the evolution of vacancies and/or ions can be monitored throughout the simulation. The generation of oxygen vacancy can be described with the thermochemical model:

$$P = f \cdot exp\left(-\frac{E_a - \Delta\varphi}{kT}\right)$$

Where P is the probability of V_O generation and hopping, f is the frequency prefactor which corresponds to the oxygen ion vibration frequency, E_a is the energy barrier for generation processes and $\Delta\varphi$ is the barrier lowering due to local electric field. The drift-diffusion model is used to simulate carriers moving in the conduction and valance bands, and the trap-assisted-tunneling model is used to simulate the carrier conduction from trap to trap. The self-heating model using Fourier's equation is also enabled to capture the local heating effect which is a crucial factor the filament formation process of ReRAM.

A HfO_x-based ReRAM cell was modeled by the stacked structure shown in Fig. 17. The switching layer is a 5 nm thick slab of HfO_x followed by a 1 nm TiO_x oxygen reservoir; the switching layers are contacted by TiN layers as inert electrodes. The area of the cell is 10 nm x 10 nm. In the initial state, the oxygen vacancies are considered uniformly distributed in the HfO_x layer with a density of 10^{19} cm^{-3}. The generation activation energy of the oxygen vacancies is set to be 2 eV. Extra oxygen vacancy-ion pairs are randomly introduced in the HfO_x layer to mimic the effect of radiation induced damage. As shown in Fig. 17 below, 40 O^- / V_o^+ pairs are generated in one region of the device. This is used to mimic the displacement damage (DD) after the device is irradiated by neutrons or heavy-ions. The set and reset processes have been simulated on both damaged and undamaged ReRAM cells to investigate the effect of DD on the performance of ReRAM.

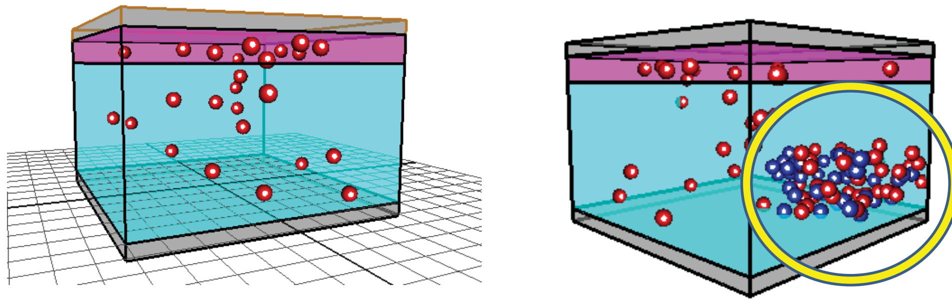


Figure 17: ***(left)*** *Undamaged ReRAM cell;* ***(right)*** *Radiation-damaged ReRAM cell with 40 additional* V_O^+ / O^- *pairs localized to the corner of the cell*

In this analysis, the initial state of the ReRAM will be denoted as State 1, the state after the set process will be State 2, and after reset process denoted as State 3. The set and reset biasing conditions are shown in Fig. 18. For the undamaged device, O^- / V_o^+ pairs are generated almost randomly throughout the whole device area, and the conductive filament is eventually formed in the area with highest Vo^+ density. For the damaged device, the conductive filament is formed at the damaged position after the set process, and the generation of O^- / V_o^+ in the undamaged regions is very limited. This is because the region damaged by radiation creates a "weak spot" in the HfO_x layer, like that described in time dependent dielectric breakdown theory. The filament formation in the damaged device is highly localized to the damaged region due to the

positive feedback of increased local electric field and local self-heating. The undamaged device ends up having more defects in State 2 than the damaged device due to the highly localized filament formation in the damaged device.

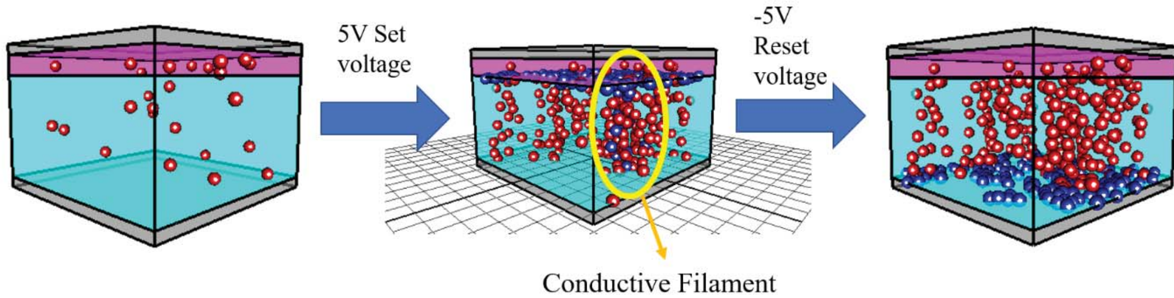


***Figure 18:* (left-to-right)** *Undamaged device, SET operation, and final device after RESET*

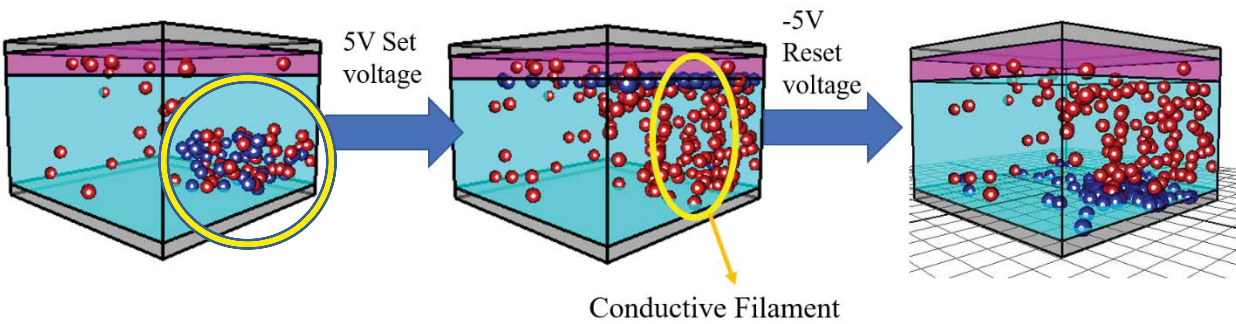


***Figure 19:* (left-to-right)** *Initially damaged device, SET operation, and final device after RESET*

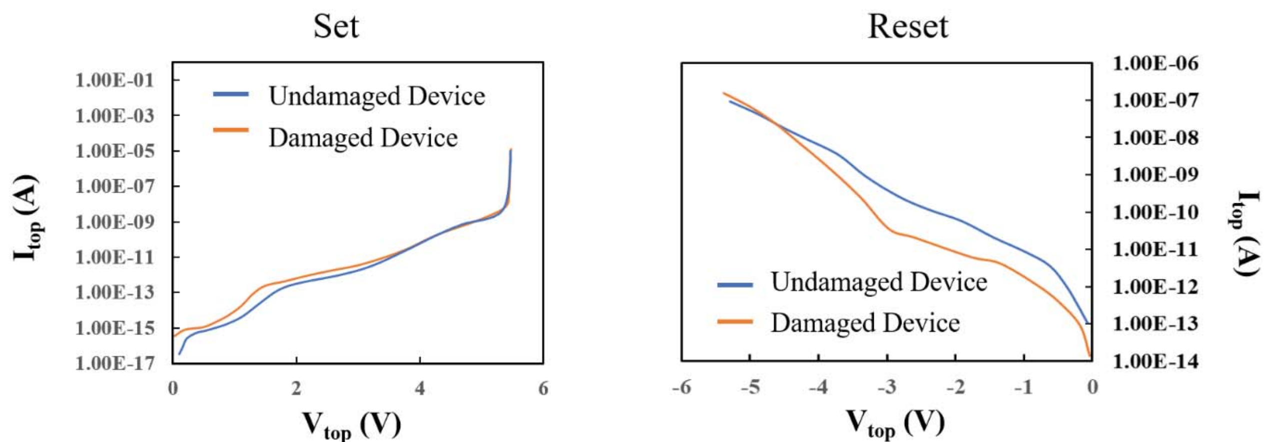


***Figure 20:* (left)** *I_{top} – V_{top} characteristics during the SET operation;* **(*right*)** *characteristics during the RESRT operation. Note the dramatically different transition curve during RESET for the damaged device compared to the undamaged device*

The I-V characteristics of the devices are shown in Fig. 20. During the set process, the current at low bias of the damaged device is larger than that of the undamaged device. This difference is caused by the introduction of extra defect at the beginning of the set process. As the voltage increases, the current of the damaged and undamaged devices becomes almost identical. This is because the current conduction is localized once a filament is formed. Therefore, although damaged ReRAM has more localized filament formation and less defects, this will have little effect on the characteristics of an ReRAM in low resistance state.

During the reset process, the voltage on the top electrode ramps from 0 V to -5 V. The O^- ions located in the TiO_x reservoir region will be driven out due to the electric field and recombine with the V_o^+ states in the filament. This will cause a decrease in the number of defects and reset the device into a high resistance state. The number of defects drops by 34 from State 2 to State 3 in the undamaged device, and 28 for the damaged device. Only a small part of the conductive filament is dissolved for both cases.

Compared to the set process, the damaged and undamaged devices show more differences in their reset behavior. The irradiated device has lower current at small reset voltages (0 V to -4.5 V) and reaches a current level comparable to the undamaged device at -5 V. This can be explained by the self-limited reset process of ReRAM. When the reset starts, the undamaged device has more defects and is more conductive. The damaged device has most of the defects located in the filament region, so the filament dissolves faster than the undamaged device. As the reset process continues, the filament dissolution becomes self-limited, as the increase of voltage will cause a decrease in the conductivity of the devices. Therefore, both devices reached similar current level when the reset voltage is roughly -5 V.

The results of the simulation study suggest that while the radiation damage can affect defect density, filament position and set-reset process of a ReRAM device, it has a relatively small effect on the $\frac{R_{on}}{R_{off}}$ characteristic. This is a result of the highly-localized set process and self-limited reset process. On the other hand, I-V characteristics during state transitions are affected by the radiation damage. Therefore, even if ReRAM is relatively robust against radiation for digital applications, its performance as an analog memory can be significantly affected by radiation.

- **Preliminary ReRAM Radiation Experiment**

  Radiation experiments on commercial ReRAM chips have been set up to investigate the radiation susceptibility of the integrated ReRAM-MOFSET (1T1R) system, which will be the fundamental building block of an ReRAM-based neuromorphic system. The experiment will use Fujitsu MB85AS4MT ReRAM chips as test vehicles. The MB85AS4MT is a 4MB ReRAM chip based on 40nm technology with CMOS and TaOx based ReRAM cells as described in [1]. Because the chips use Serial Peripheral Interface (SPI), an Arduino UNO microcontroller board is used for programming,

reading and writing the chips. The connection schematics and a picture are shown below in Fig. 21.

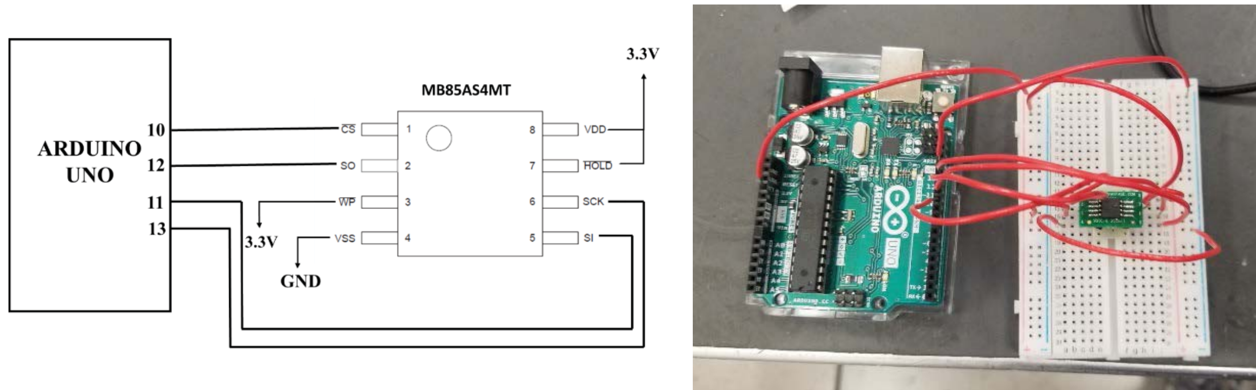


Figure 21: ***(left)*** *Schematic of microcontroller and ReRAM chip interface* ***(right)*** *image of the chip on a breakout board, integrated with an Arduino MCU via breadboard and serial peripheral interface*

The experiment will follow a write-irradiate-read-compare procedure. The whole ReRAM array will be programmed with test patterns, such as checker board or random data. The chip will then be put into standby mode and irradiated with varying dose in the ARACOR X-ray irradiator and Pelletron linear particle accelerator. After the radiation, the MCU will run an automatic readout-compare program to output the number of errors within the memory array. The radiation dose will be incrementally stepped until upsets are detected, either in the ReRAM, peripheral circuitry, or access transistors. Using heavy ions and protons as sources, the effect of various sources and particle fluences can be investigated Another experiment is irradiate the device while it is in read/write mode. All the hardware and software part of the experiments has been setup, and the experiments are currently in progress at Vanderbilt University.

[1] Y. Hayakawa *et al.*, "Highly reliable TaOx ReRAM with centralized filament for 28-nm embedded application," *2015 Symposium on VLSI Circuits (VLSI Circuits)*, Kyoto, 2015, pp. T14-T15.

4) Key Outcomes

As previously stated in last year's report, the overarching objective of this work is to determine the robustness and resiliency of brain-inspired computing platforms implemented in non-von Neumann architectures, as compared to traditional von Neumann architectures. In this past year, we have established a robust electrical characterization environment for probing single devices and advanced test macros. Using a Keysight B1500A analyzer, the device characteristics of the CTT were measured and the device's operation as an analog memory was explored. By utilizing careful program and erase methodology (PVRS), the threshold voltage of a CTT has been shown to be controlled in a linear, reversible, and repeatable fashion. The retention properties of this stored charge was also investigated, and a programming methodology to counteract the nuisance of short-term charge loss due to energetically shallow traps was elaborated. Utilizing the hardware obtained from the team's chip tapeout last year, arrays of devices were programmed to

benchmark the accuracy of programming across a wide, continuous dynamic range of analog weights (device conductances).

Utilizing the characterization data obtained by experiment, the TCAD simulation platform Ginestra was employed to model the charge trapping kinetics of the CTT. Preliminary models capture the important characteristics of the device programming and erasing, and are currently being further developed quantitatively to predict device behavior in experiment. The ARACOR X-ray irradiator at Vanderbilt was used to study the effect of total dose on the CTT characteristics as an analog memory. It was discovered that the strong electrostatic coupling of the BOX to the channel potential caused the device to be sensitive to 10-keV X-rays, as trapped charge significantly shifted the I_D-V_G characteristics. This shift in V_{th} characteristics is critical for an analog memory, as this can lead to large shift in the subthreshold current at a particular bias condition, which is the equivalent of the analog weight (device conductance). While this is a detrimental effect, several methods of mitigation were proposed, from using a different base technology (PDSOI or bulk), to utilizing the back-bias effect in FDSOI, to modifying the training algorithm to be more agnostic to weight variation.

Preliminary simulations of radiation-induced damage effects in ReRAM were also conducted, and found that the digital operation of the device may be rather unperturbed, however the analog characteristics were changed significantly. The additional device damage causes localized “hot spots” for electric fields during set and reset operations, providing a positive current and temperature feedback for filamentary formation. The transition between the high resistance and low resistance states changes significantly with radiation-induced damage, making the operation as an analog memory unpredictable and uncontrollable. Initial experiments utilizing a Fujitsu commercial ReRAM chip are underway to investigate the resiliency of a product with CMOS overhead and CMOS access transistors.

What do you plan to do during the next reporting period to accomplish the goals?

We will proceed with the current plan with more radiation studies, simulations, and modeling to be conducted during the next reporting period. Specifically, we will investigate the dynamic behavior of X-ray radiation on single CTT devices under various bias conditions, after programming, after erasing, after multiple program & erase cycles, and measure breakdown characteristics after radiation (TDDB, etc.) The tapeout from last year provided arrays of standalone devices that can be irradiated as-is or after programming in order to determine the effect on arrays of CTT devices and possibility for mitigation using back-bias. Another version of the chip was also just taped out, featuring two large layers (1024 x 512 cells) and a more complex overhead for multi-chip integration. These chips are flip chip and we plan to irradiate standalone arrays and devices on these chips, as well as the neuromorphic inference engine under operation and pre-programmed. Utilizing these methods, the direct effect of various radiation sources on a live analog neuromorphic can be measured and determined. The preliminary models discussed in this report will be elaborated upon and calibrated to experimental measurements as well, with emphasis on predicting the charge trapping behavior in various device states. Studies on ReRAM will be further developed following the observations from the upcoming first experiments, along with the simulated mechanisms underlying the physical operation of the device. Investigation into the effects of radiation upsets in GPU-based implementations will also be of prime importance and compared to CPU-based implementations. Finally, the IBM TrueNorth will be utilized to understand how these errors propagate in these neuromorphic architectures, irrespective of the microarchitecture used.

Year 3

Major Activities

i. Establish reliability characteristics of CTT for use as analog synapse

Elaborating on prior work regarding the programming characteristics of the CTT, we have further explored the reliability components which set limits on maximum voltage conditions and programming pulse time durations that can be reliably applied to the device during programming. In addition, we have studied both immediate and long-term relaxation of the device which allows us to model how we expect each respective device's Vth (or σ_{ch}) to change as a function of time. We are able to purposefully over-program devices to counteract shallow de-trapping which occurs almost immediately after a programming cycle. Similarly, we have explored long-term charge de-trapping as a function of both time and temperature which allows us to determine how frequently a network may require weight updates. Another important component relating to the reliability of the CTT device is understanding the subthreshold degradation which occurs in some devices due to the considerable stress occurred across the device during programming.

ii. Irradiate CTT Devices and compare results across different candidate technology platforms including FDSOI and bulk technologies

GlobalFoundries 22FDX (FDSOI) and 14LP (bulk) wafers were quartered and diced to isolate several device test macros onto singulated dies which were bonded to hybrid ceramic packages with Au wirebonds. For 22FDX, approximately 50 macros each of various types were provided (single devices with width scaling, devices with length scaling, metal-oxide-metal capacitors, and ganged-gate devices). For 14LP, several ganged-gated configurations were provided with varying Vth-implant (slvt, lvt, rvt, & hvt), gate-pitch (78nm, 84nm, 190nm, & 310nm), and number of fins (2, 4, & 40 fins). All irradiation studies were performed at Vanderbilt University using their ARACOR 10-keV X-ray irradiator system. Results have shown that for FDSOI (specifically 22FDX), significant trapping occurred in the buried oxide (BOX) layer of the devices, which is highly coupled to the channel in FDSOI technology nodes. The induced trapping in the BOX caused a significant Vth-shift across all devices. This effect could be potentially mitigated by employing circuitry to leverage the back-gate available in 22FDX effectively depending on the systematic nature of the radiation-induced Vth-shift across all devices.

iii. Develop physics-based models for confirming observed radiation effects on CTT

A physics-based model was developed using TCAD software to explore the effects of radiation on FDSOI devices. The results obtained in this simulation are provided in detail under Section 3: Significant Results. The underlying result supporting hole trapping in the buried oxide (BOX) layer as the dominate effect during TID. This effect is especially pronounced in FDSOI technology nodes due to the strong coupling between the BOX layer and the device channel.

iv. Explore RRAM Physics & Radiation Effect Modeling

A HfOx-based RRAM cell was modeled by the stacked structure shown in the figure below with

a 5nm HfOx switching layer followed by a 1nm TiOx oxygen reservoir. The switching layers were contacted by TiN layers as inert electrodes. The area of the cell is 10nm × 10nm. The oxygen vacancies are considered uniformly distributed in the HfOx layer with a density of 1×10^{19} cm^{-3} in the initial state. The initial position of each defect is randomly generated throughout the HfOx layer. Extra oxygen vacancy/ion pairs are randomly introduced throughout the HfOx layer to mimic the effect of TID. As shown in **Fig. 1**, 40 oxygen ions/vacancies (O-/Vo+) pairs are generated in the one region of the device. The set/reset process has been simulated on both damaged and undamaged RRAM cells to investigate the effect of DD on the RRAM performance.

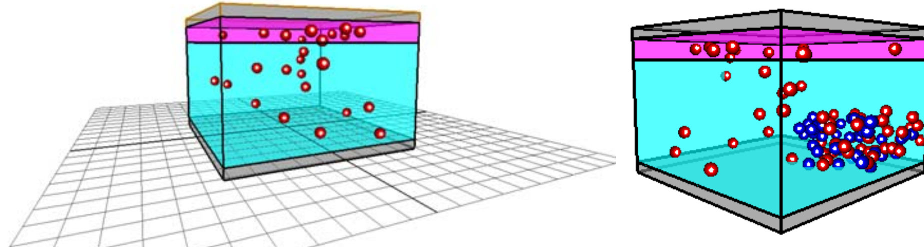


Fig. 1. Simulated RRAM structure.

v. Irradiate Commercial RRAM chip

We irradiated a commercial RRAM chip from Fujitsu and studied how the device characteristics of a large RRAM array shifted with radiation dose. It is important to note here that we were particularly interested in observing how the integrated CMOS overhead influences the performance of the RRAM array given that the CMOS logic access circuitry is required in order to utilize the RRAM technology.

vi. Investigate mitigation techniques for general NVM-based analog networks

Mitigation techniques for general NVM-based analog networks have been explored as shown in **Fig. 2**. Simply by considering an additional ΔW term when training the network weights—given that exact weights cannot be programmed into a CTT array due to intrinsic device variation in an analog network—we can train the network to be more resilient to both programmed device variation as well as device relaxation over time.

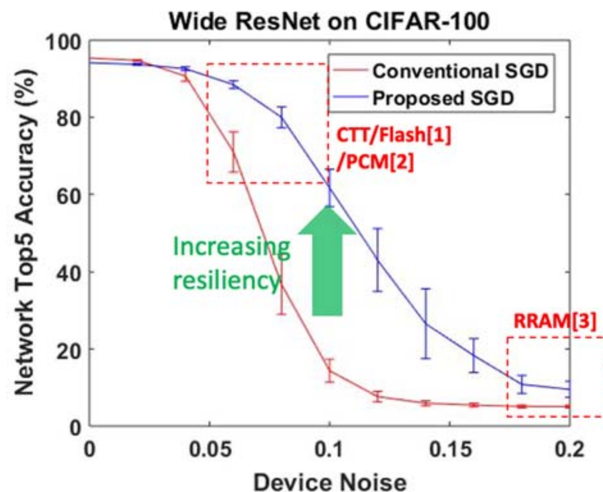


Fig. 2. Network Top5 Accuracy vs. Device Noise utilizing hardware-aware stochastic gradient descent training algorithm.

vii. Develop software-level training & inference simulator

In order to understand the effect of CTT variation, device relaxation, etc., we developed a software-level Matlab-based simulator which allows us to (1) train a network to map onto our twin-cell CTT architecture and (2) evaluate inference performance considering various device variations which generally lower system inference accuracy. This allows us to evaluate specifically how much of an effect device variation has on individual network architectures and target problem sets (e.g. MNIST). This also allows us to perform Monte Carlo simulations to see how a particularly trained network may be expected to perform across several different chips, all with different device variations. This work also includes some modifications that have been made to the training algorithm itself to minimize the effect of device variation on the expected output. Given that each individual CTT must be programmed to a roughly precise state, it should not necessarily be trained in the same way a digital network may be trained. We have modified the training loss function to account for the fact that we cannot exactly program or encode the weights obtained during the training process into an array of CTT devices—there will always be some variation between the trained set of weights and actual programmed weights within the CTT array.

viii. Develop CTT-hardware-based Inference Realistic Circuit Simulator (CIRCS)

We have also begun work towards developing a circuit-level simulator (CIRCS) which allows us to not only consider the effects of device variation but also circuit-level variations including post-layout parasitics, process corners, temperature, leakage and other non-idealities including bondwire inductance. The circuit-level simulator allows us to more accurately demonstrate the inference process by deploying circuit-level models for all peripheral circuitry including the Word Line Driver (WLD) and Neuron. It allows us to quickly iterate over the entire network test dataset and evaluate the performance of the network given all of the known circuit non-idealities (specifically in regard to the peripheral circuitry) and the CTT device non-idealities. It also allows us to perform Monte Carlo simulation to assess how a specific network may be expected to perform across a distribution of chips and determine the effect of device relaxation to give us an idea of when a chip may require weight updates or a refresh cycle.

2) Specific Objectives

- Study technology-dependent effects of TID irradiation on CTT devices (by comparing irradiation experiments leveraging GlobalFoundries 14LP and 22FDX technologies)
- Develop subsequent charge-trapping models utilizing available experimental data to incorporate mature device relaxation models, TID-induced ΔVth and other long-term device effects, and endurance
- Deploy charge-trapping models in analog neural network simulations & more complex circuit-level simulations to study effect
- Study effects of device error on different types of deep learning network architectures & classification problems (e.g. image classification, object detection, segmentation, audio/keyword recognition, etc.)
- Determine what types of problems and which network architectures the CTT is most suitable for
- Develop Circuit-Level simulator to simulate the effects of device non-idealities such as programming error (Vth variation), device Vth relaxation and TID-induced hole-trapping & ΔVth as well as circuit/implementation non-idealities such as parasitics, process corners, BL-biasing instability, common-mode feedback, integrator design and gain-bandwidth product, etc.
- Irradiate commercially available RRAM chips and determine their resiliency
- Explore physical models governing the operation of RRAM cells using TCAD simulation and determine whether logic access circuitry is the bottleneck of RRAM chip performance under irradiation

3) Significant Results

i. Establish Programming & Reliability Characteristics of CTT

Shallow de-trapping is observed to occur immediately after programming. **Figure 3** shows that this de-trapping or relaxation effect increases the more a device is programmed. A relaxation compensation method has been developed to 'over-program' devices to compensate for the large relaxation of the device which occurs within ~1hr of programming each device. This relaxation compensation method leads to a stable & systematic correction in achieved programmed/target weights after measuring 200 weights approximately 200 hours after the initial programming step.

Fig 3. (a) CTT relaxation 1 hour after programming as a function of target program current. (b) Demonstrating CTT Relaxation compensation where 200 CTT devices are purposefully over-programmed to account for any immediate loss of charge due to shallow de-trapping. (c) End result of CTT Relaxation Compensation after 200 hours of relaxation. Charge de-trapping is exponential in-time leading to large initial de-trapping after programming each device.

Subthreshold degradation is also an issue worth considering given both its effect on increased BL leakage as well its reliability implications. **Figure 4** depicts the ideal I_{DS}-V_G curve after a programming step followed by an erase step which returns the Vth closer to the virgin or baseline state. **Figure 5** then elaborates on this by demonstrating simply shifting each I_{DS}-V_G curve by the programmed ΔVth may not be sufficient given that some devices may experience subthreshold degradation. The effect of subthreshold degradation on network performance cannot be assumed to be negligible and must be studied.

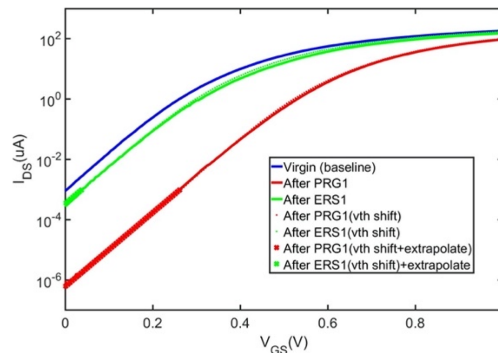


Fig. 4. Example 22nm CTT device where a virgin device I_{DS}-V_G curve is shifted by the corresponding programmed ΔVth after a program (PRG1) followed by an erase (ERS1) step. This plot does not accurately depict the subthreshold degradation which occurs across some devices after stressing the device under large bias conditions during programming and erase conditions (see **Fig. 5**).

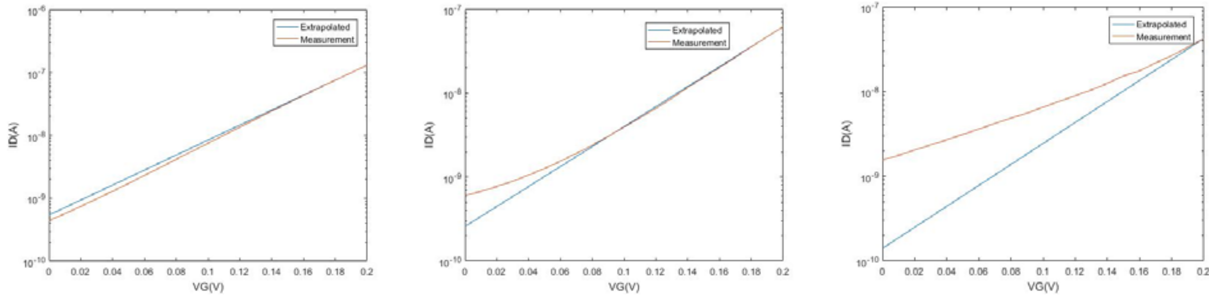


Fig. 5. Three sample CTT devices with varying subthreshold degradation leading to a difference in subthreshold current between the actual subthreshold current (measurement) and the expected subthreshold current obtained by simply shifting the entire Ids-Vg curve by the programmed Vth.

It turns out that given that our architecture reverse biases off devices during inference by providing a gate voltage of -0.3V, the effect of subthreshold degradation is largely negligible on network performance. We need only consider the effect of increased subthreshold current when analyzing the rise & fall times associated with the constant-amplitude pulse-width modulated (PWM) input waveforms as shown in **Fig. 6**.

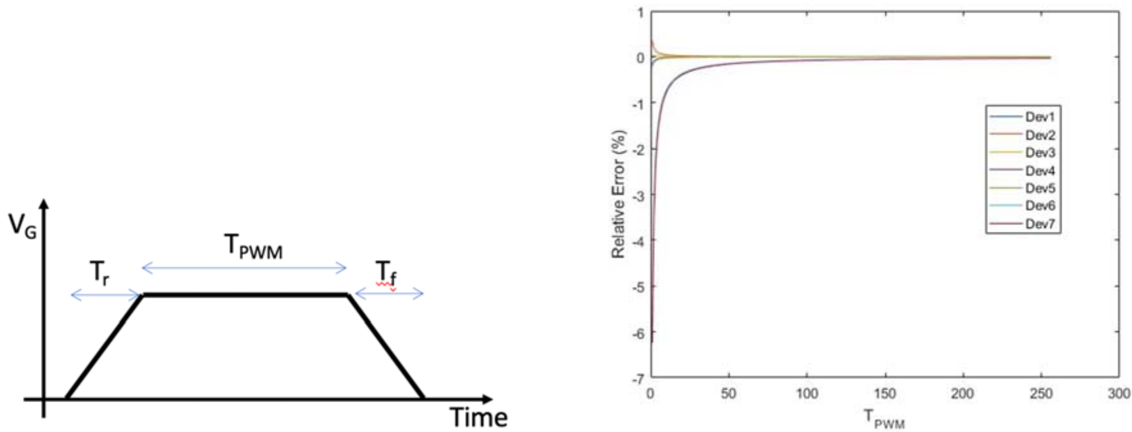


Fig. 6. (a) Constant-amplitude pulse-width modulated (PWM) input waveforms applied to each WL in the CTT array. Each input waveform is a function of the input ($T_{PWM}=N_{INPUT}$ x T_{cycle}) and the finite rise and fall times associated with that waveform. The middle part of the waveform is characterized at a constant V_G bias while the rise and fall portions of the waveform have varying VG and hence varying σ_{ch}—which is a function of V_G. The longer the rise & fall times are, the larger the error due to subthreshold degradation (increased leakage current) is. (b) Relative error due to subthreshold degradation across 7 different devices, given the rise & fall time of the PWM input waveforms. Larger inputs (e.g. longer T_{PWM}) reduces the effect of increased subthreshold degradation on the overall output. This error in most cases (for most inputs and devices) is <1% with one device showing up to 6% area for a PWM input of 1 and reducing for larger inputs. The input space supports inputs between 0 and 255.

ii. Develop model for assessing the effect of radiation on CTT

TCAD simulations were used to investigate the charge-trapping mechanisms responsible for the programming and radiation responses. Programming was simulated through the addition of the

density of trapped electrons at the upper transistor interface with the gate dielectric until bias-induced shifts were reproduced. For the data of Fig. 10, for example, the inferred trapped-electron density projected to the gate oxide/Si interface is ~6×10^{12} cm^{-2}. Similarly, radiation-induced shifts were simulated through the addition of trapped holes at the lower interface of the transistor with the buried oxide until the TID-induced shifts were reproduced. For the data of Fig. 10, the inferred trapped-hole density projected to the buried oxide/Si interface is ~ 1.5×10^{12} cm^{-2}.

Figure 7 plots the simulated charge densities in the transistor channel for four simulations conducted on a representative 22 nm FDSOI TCAD model based on the process details of [17] and the experimental data of Fig. 10. The simulated ID-VG curves that match the experimental results of Fig. 10 are shown in Fig. 8 as a function of gate voltage. The simulated cases include a “pre” case where no additional carriers were introduced into FDSOI transistor model, a “programming” case with electrons in the gate oxide, an “irradiation” case with holes in the buried oxide, and a “combined” case with electrons in the gate oxide and holes in the buried oxide. As expected, the “programming” case results in a lower channel electron density (transistor turned more strongly off), the “irradiation” case shows a higher electron density in the channel (transistor turned strongly on), and the “combined” case is turned on more strongly than the “pre” case but not as strongly as the irradiation-only case. These results show that a self-consistent model of the programming and irradiation responses of these devices can be built, with plausible charge densities in the gate dielectric and buried oxide, under the simple assumptions stated above. In the full paper, the evolution of charge densities during more complex programming and irradiation sequences will also be illustrated.

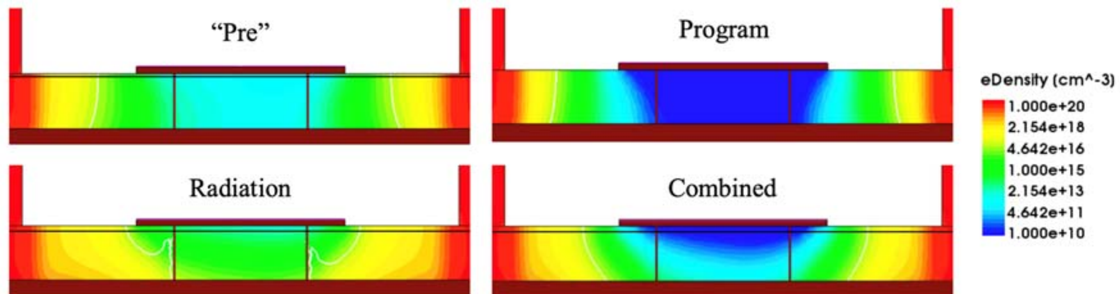


Fig. 7. Representative 22nm FDSOI TCAD model based on [17] showing simulated electron density for an un-irradiated & un-programmed CTT (top left), programmed CTT (top right), irradiated CTT (bottom left), and a programmed CTT that has been irradiated (bottom right).

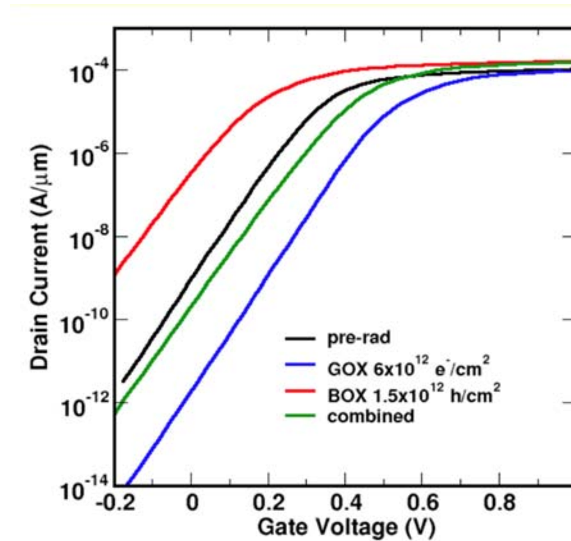


Fig. 8. TCAD model showing simulated results of “pre” transistor responses with respect to a simulated pre-irradiation, pre-programmed CTT (black). The blue curve simulates programming, and red curve simulates radiation. The combination of simulating both programming and radiation results in the green curve.

iii. Irradiate FDSOI CTT devices (GlobalFoundries 22FDX)

Weak programming before irradiation: **Figure 9(a)** shows I_D-V_G curves for Device A, programmed before irradiation. Programming the device resulted in a positive V_{th} shift of 50 mV, as shown by the solid black (pre-programming) and dashed-black (post-programming) curves. The TID response of Device A after programming is shown in the remaining curves. After 50 krad(SiO_2) irradiation (blue curve), V_{th} shifted negatively by ~40mV, which nearly offsets the initial programming. Irradiating the devices up to 500 krad(SiO_2) resulted in V_{th} shifts of ~ -140mV.

Weak programming after irradiation: **Figure 9(b)** shows I_D-V_G curves for Device B, programmed after irradiation. The solid black curve again shows the pre-irradiation, pre-programming response. The remaining solid curves show the TID response of the device. V_{th} again shifted negatively by ~40mV after the device was irradiated to 50 krad(SiO_2), and by ~ -155mV after irradiation to 500 krad(SiO_2). The post-irradiation programming resulted in a positive Vth shift of 30 mV. The similarity of the TID response for programmed and unprogrammed devices in **Figures 9(a) & 9(b)**, and the relative similarity of the V_{th} shifts during programming for the unirradiated and irradiated devices, suggest that the charge trapping responsible for the TID response and programming may not be interacting strongly, at least to first order. If so, this would contrast with the responses of Flash memory devices, for example, where radiation-induced charge neutralization **[5]-[7]** can lead to rapid loss of the programmed state. We now explore these mechanisms in more detail for devices programmed to more positive V_{th} values.

Strong programming before and after irradiation: **Fig. 10** shows V_{th} shifts for a device programmed before irradiation, irradiated to 500 krad(SiO_2), annealed at room temperature for 52 h, and re- programmed. The initial programming in this case was stronger (higher voltage, longer time) than for the devices of **Fig. 9(a)**, and resulted in a positive 205 mV V_{th} shift. After this programming, irradiation to 500 krad(SiO_2) led to a -145mV shift, similar to the post-programming TID response in Fig. 1. Annealing led to a positive 20 mV V_{th} shift. The device was

then programmed again in the same way as before irradiation. This programming sequence resulted in only a positive 55 mV shift (135 mV with respect to baseline). These results indicate that the post-irradiation and programming response of the device can be quite sensitive to device history.

For all cases in **Figures 9 and 10**, the TID response of these fully-depleted devices is expected to be dominated by hole trapping in the buried oxide **[14]-[16]**. The presence or absence of trapped charge in the gate dielectric does not significantly affect the electric field in the BOX, leading to similar amounts of charge trapping when devices are irradiated before or after programming. However, the presence of pre-existing programming and/or radiation-induced charge in the gate dielectric can limit the ability of dielectric layers to capture additional charge due to trap filling, charge neutralization, and pulse modification effects **[8], [13]**.

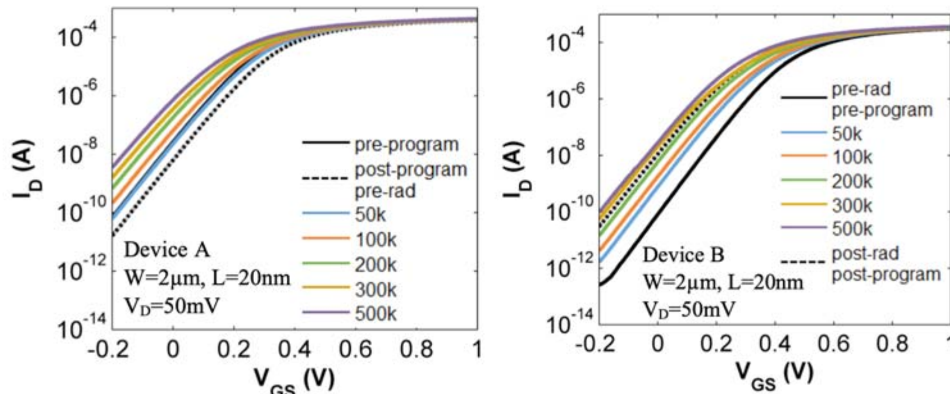


Fig. 9. *(a)* Drain current as a function of gate voltage for Device A programmed *before* irradiation. The solid black curve (partially hidden by the blue curve) shows the pre-irradiation, pre-programmed DC sweep. The dashed black curve shows the response of the device after programming. The other five curves show the post-programming radiation response of the device. *(b)* Drain current as a function of gate voltage for Device B *after* irradiation. The solid black curve shows the pre-irradiation, pre-programmed, radiation-induced shifts. The dashed blacak curve is the post-programming response of the device after irradiation to 500 krad(SiO_2) (purple curve) and subsequent programming.

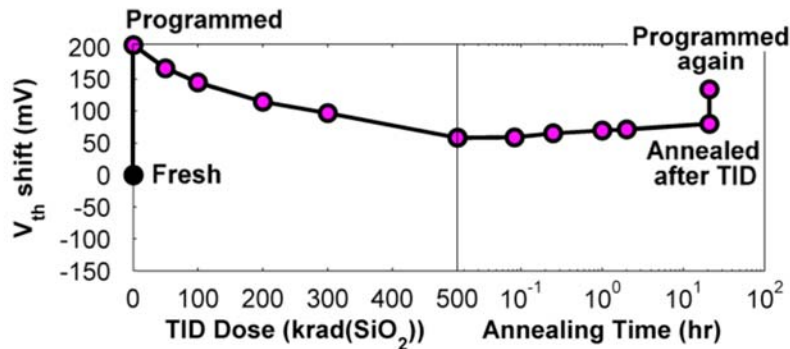


Fig. 10. Threshold voltage shifts of a device as it is programmed, irradiated, annealed, and programmed again.

Irradiation of 22nm CTTs fabricated in an FDSOI technology results in threshold voltage shifts between 40mV and155 mV for doses between 50 and 500 krad(SiO_2). These shifts are due most likely to charge trapping in the buried oxide **[15], [18]**. The programming threshold voltage shifts are due primarily to electron trapping in the gate oxide. To first order, the charge

distributions in the gate dielectric and buried oxide do not interact strongly, enabling the construction of TCAD models with plausible charge densities. In this summary, inferred densities of trapped electrons are derived for a strong programming pulse, and inferred densities of trapped holes in the buried oxide are derived for devices irradiated to 500 krad(SiO_2).

v. Irradiate bulk FinFET CTT devices (GlobalFoundries 14LP)

We also irradiated unprogrammed GlobalFoundries 14LP devices and studied the effects of irradiation on a non-FDSOI technology for comparison. **Figure 11** shows a substantially smaller Vth shift for the 14nm bulk technology compared to previously shown 22FDX (FDSOI) results.

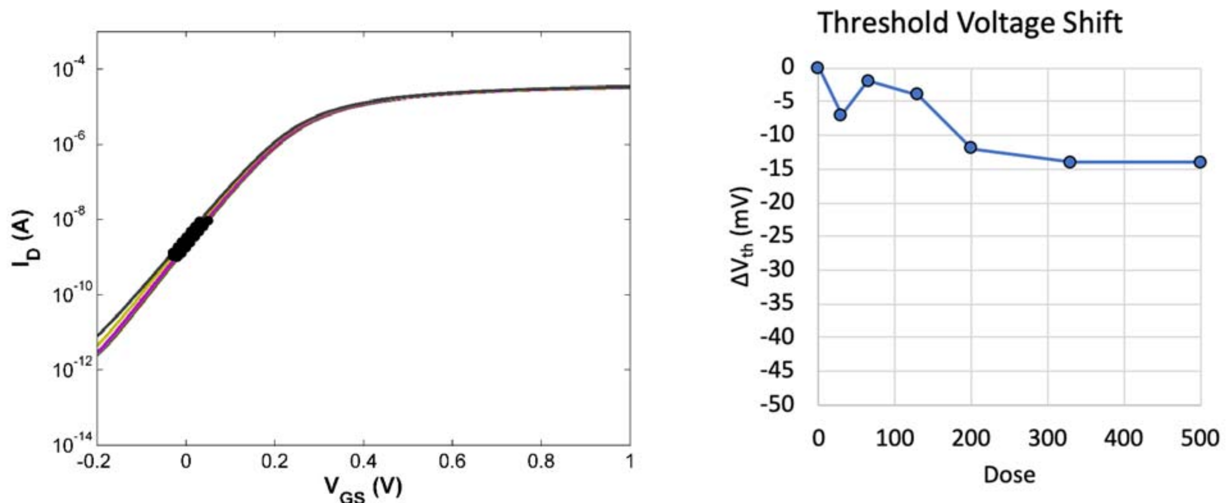


Fig. 11. 14nm bulk FinFET Semilog IV and V_{th} for SLVT device with L=16nm and CPP=84nm, and N_{fins}=2. Device exposed to up 500 krad(SiO_2).

v. Negligible effects on RRAM cell observed in simulation

The results of the simulation study suggest that while the radiation damage can affect defect density, filament position and the set-reset process of an RRAM device, it has relatively small effect on the value of R_{ON}/R_{OFF}. Therefore, RRAM is relatively robust against radiation for digital applications.

Figure 12 displays the Set-Reset cycle for an undamaged device. O^-/Vo^+ pairs are generated almost randomly throughout the whole device area, and the conductive filament is eventually formed in the area with the highest Vo^+ density. For the damaged device—shown in **Fig. 13**—the conductive filament is formed at the damaged position after the Set process, and the generation of O^-/Vo^+ pairs in the undamaged regions is very limited. This is because the region damage dby radiation creates a 'weak spot' in the HfOx layer, similar to time-dependent dielectric breakdown (TDDB). The filament formation in the damaged device is highly localized to the damaged region due to the positive feedback of increased local electric field and self-heating. The undamaged device ends up having more defects after the Set process (state 2) than what is observed in a damaged device due to the highly localized filament formation in the damaged device.

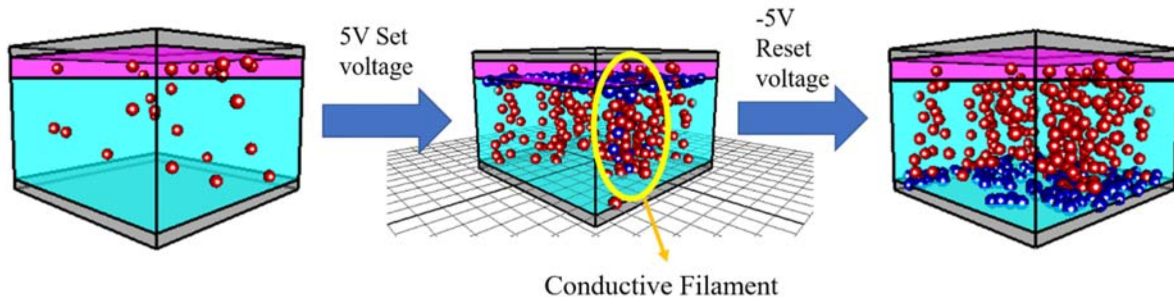


Fig. 12. Set-Reset cycle for undamaged RRAM cell.

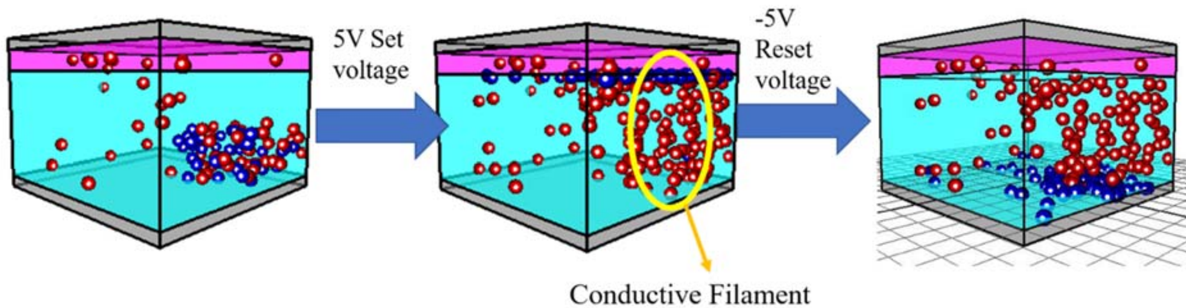


Fig. 13. Set-Reset cycle for damaged RRAM cell.

The I-V characteristics of the devices are shown in **Fig. 14**. During the Set process, the current at low bias is larger for the damaged bias compared to the undamaged device. This difference is caused by the introduction of extra defects at the beginning of the Set process. As the voltage increases, the currents of the damaged and undamaged devices become almost identical because the current conduction is localized once a filament is formed. Therefore, although a damaged RRAM device has more localized filament formation and fewer defects, this will have little effect on the characteristics of an RRAM device in a Low-Resistance (R_{ON}) state. During the Reset process, the O^-'s located in the TiOx reservoir region will be driven out due to the electric field and recombine with the Vo^+'s in the filament; however, only a small part of the conductive filament is dissolved for both cases. The I-V characteristics are shown in **Fig. 14**. Compared to the Set process, a larger difference between damaged and undamaged devices is observed in the Reset process. When the reset starts, the undamaged device has more defects and is more conductive. The damaged device has most of the defects located in the filament region, so the filament dissolves faster than in the undamaged device. As the Reset process continues, the filament dissolution becomes self-limited because the increase of the voltage leads to a decrease in the conductivity of the devices. Therefore, both devices reached a similar current level when the reset voltage is about -5V.

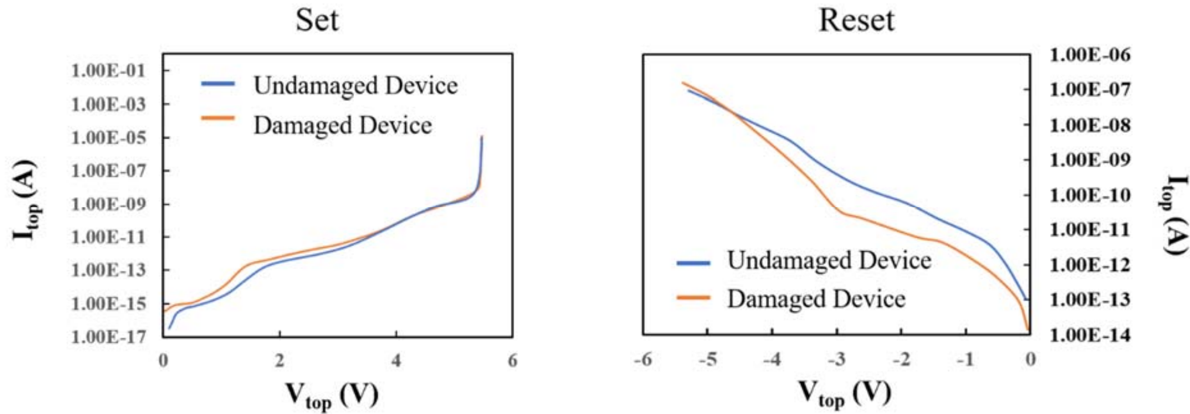


Fig. 14. I-V characteristics (Set/Reset) for both undamaged and damaged RRAM cells.

vi. RRAM performance under irradiation is limited by the CMOS logic access circuitry

The CMOS control circuits were found to be more vulnerable to TID than the RRAM cells themselves. As shown in **Table 1**, the DUT (4Mb RRAM, Fujitsu MB85AS4MT) showed a 100% success rate at 50 krad(Si); however, after further increasing the radiation dose above 50 krad(Si), a zero success rate was observed—indicating a functional failure. The functional failure is characterized by loss of communication with the DUT to program and read data. There were no errors observed within the RRAM memory array prior to observing the functional failure. The behavior of the RRAM and the operation of RRAM under the effect of radiation was also simulated using the Ginestra software package. The RRAM cells were determined to be fairly robust to radiation in simulation, and the irradiation test results suggest that the CMOS control circuitry is more vulnerable to failure than the RRAM cells themselves at these particular TID levels.

Table 1 shows the success rate of programming and reading data from the test sample. At 0 and 50 krad(Si), the success rates for all four procedures were 100%, while they were 0% at 100 krad(Si). This result clearly indicates that the test sample was functional up to 50 krad(Si), and its functionality started to degrade beyond 50 krad(Si) TID.

The simulation results provide further evidence that the degradation of the device is due to the sensitivity of the CMOS control circuitry under the effect of radiation. As shown in **Fig. 8**, the damaged device has lower current at low bias, but eventually both devices reach similar current levels. Therefore, we conclude that there is little effect on the characteristics of an RRAM cell in both the Low-Resistance (R_{ON}) and High-Resistance (R_{OFF}) states. The results of the simulation indicate that RRAM is relatively robust to radiation—especially for digital applications. Meanwhile, it is shown in **[6]** that CMOS will have aa >0.5V V_{th} shift after being exposed to 100 krad proton dose. It is possible that this V_{th} shift will lead to failure in the CMOS control circuitry. Therefore, it is more likely that the chip failed because of failures within the CMOS control circuitry rather than failures within the RRAM memory array itself. This also explains the immediate transition between 100% success rate to 0% success rate as the data in the RRAM memory array is no longer accessible.

TID level	Procedure	Success rate
0 krad(Si)	Initial read after irradiation	100%
	Re-read for verification	100%
	Re-program and then read	100%
	Erase followed by program and then read	100%
50 krad(Si)	Initial read after irradiation	100%
	Re-read for verification	100%
	Re-program and then read	100%
	Erase followed by program and then read	100%
100 krad(Si)	Initial read after irradiation	0%
	Re-read for verification	0%
	Re-program and then read	0%
	Erase followed by program and then read	0%

Table 1. Irradiation testing results at various TID levels (10-keV X-ray exposure).

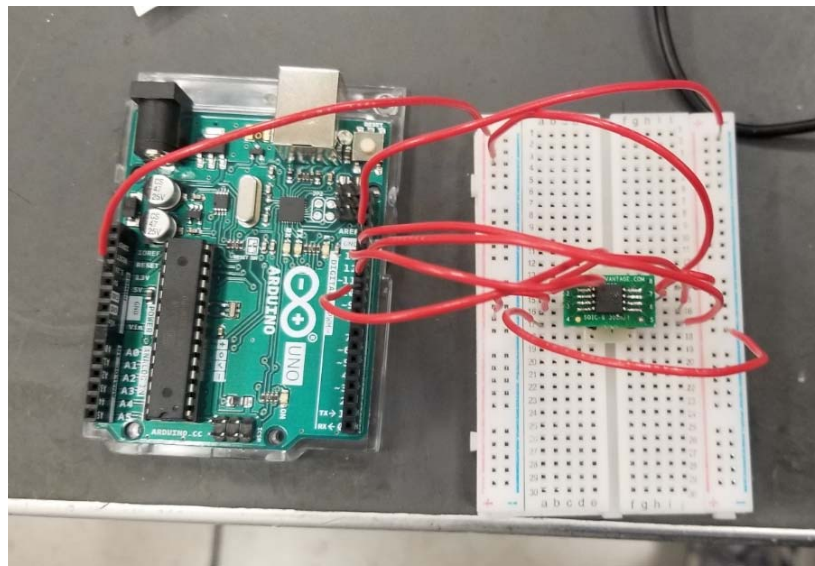


Fig. 15. Photograph of the measurement setup showing Arduino Uno on the left and the MB85AS4MT on the right

vii. Develop CTT-hardware-based Inference Realistic Circuit Simulator (CIRCS)
The CIRCS simulator allows us to provide realistic system performance information including both circuit & device non-idealities. In addition, it allows us to evaluate the accuracy of a network across several chips using a Monte Carlo simulation approach. CIRCS is in the process of being developed by includes all crucial circuit non-idealities including parasitics, process corners, temperature variation, etc. as well as random CTT variation. It also allows us to study the effect of CTT device relaxation and evaluate how frequently a system may require a lightweight refresh to reinforce the target network weights.

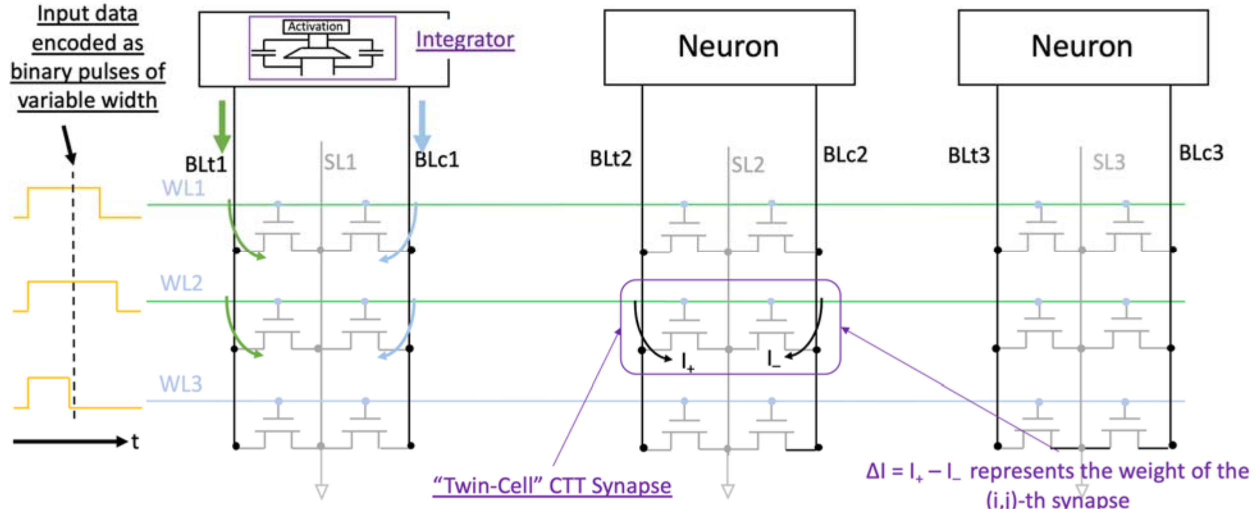


Fig. 16. Proposed Architecture for CTT Inference Engine. Inputs to the network are applied as constant-amplitude pulse-width modulated (PWM) signals to the WLs. Weights are differential weights using two pre-programmed CTTs (programmed to specific states prior to performing inference, determined during training phase). Each neuron output corresponds to a column of devices. The neuron output is itself a constant-amplitude PWM signal that is proportional to the integrated differential current ($\int (I_{BLt} - I_{BLc})dt$).

Figure 17 shows the simulated BLt and BLc currents for a random set of 256 inputs for a particular column or neuron output. The neuron output would then be a constant-amplitude pulse-width modulated (PWM) output that is proportional to the integrated differential current.

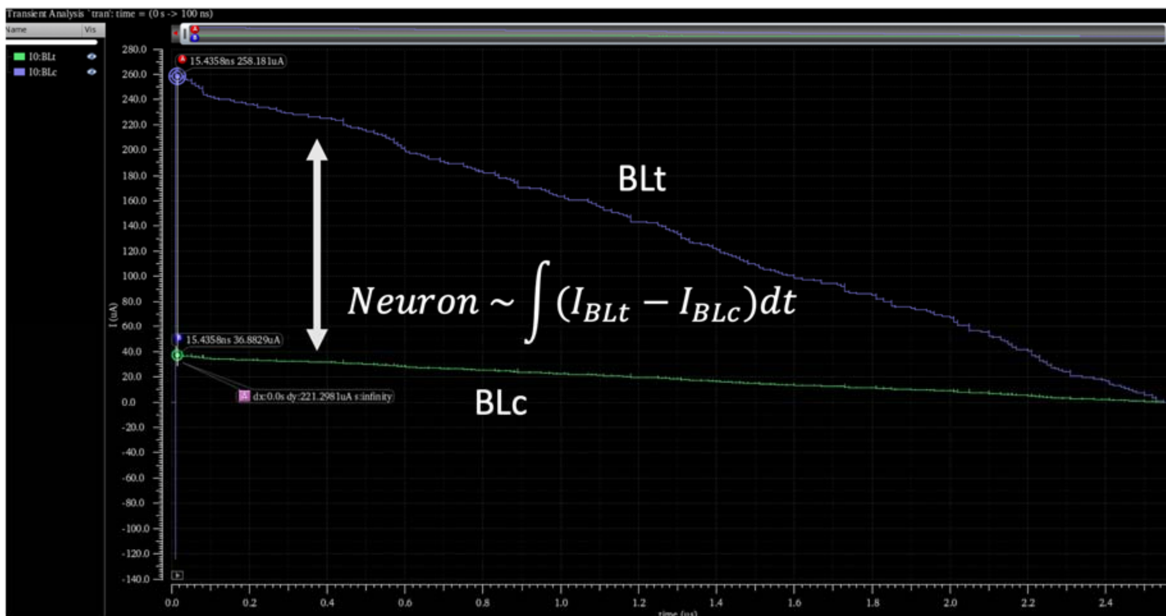


Fig. 17. Single Neuron Output given 256 inputs. The neuron output corresponds to the integrated differential current ($\int (I_{BLt} - I_{BLc})dt$). Both I_{BLt} and I_{BLc} are shown as monotonically decreasing current waveforms given that all 256 pulse-width modulated inputs begin at t=0 and end depending on their respective input values. Input values in this example are assigned randomly. The neuron output is itself a constant amplitude PWM signal that is proportional to the integrated differential current.

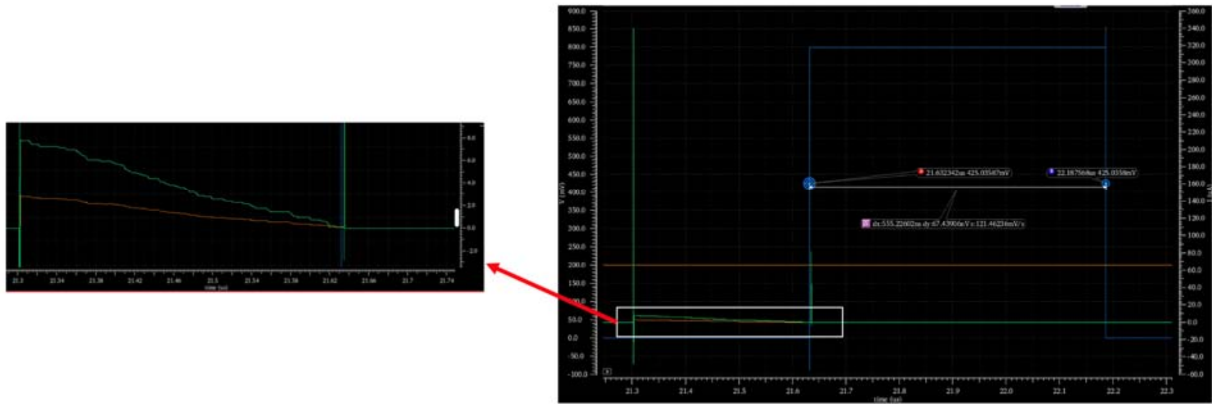


Fig. 18. CIRCS Simulator neuron output for a randomly initialized CTT array with random inputs. The zoomed in view on the left shows the I_{BLt} and I_{BLc} curves for a given input sample to the network and the graph on the right shows the pulse-width modulated (PWM) output of the neuron in blue which is directly proportional to the integrated differential current ($\sim \int (I_{BLt} - I_{BLc})dt$). In this example, Tcycle = 1.333ns and the neuron output is ~555ns.

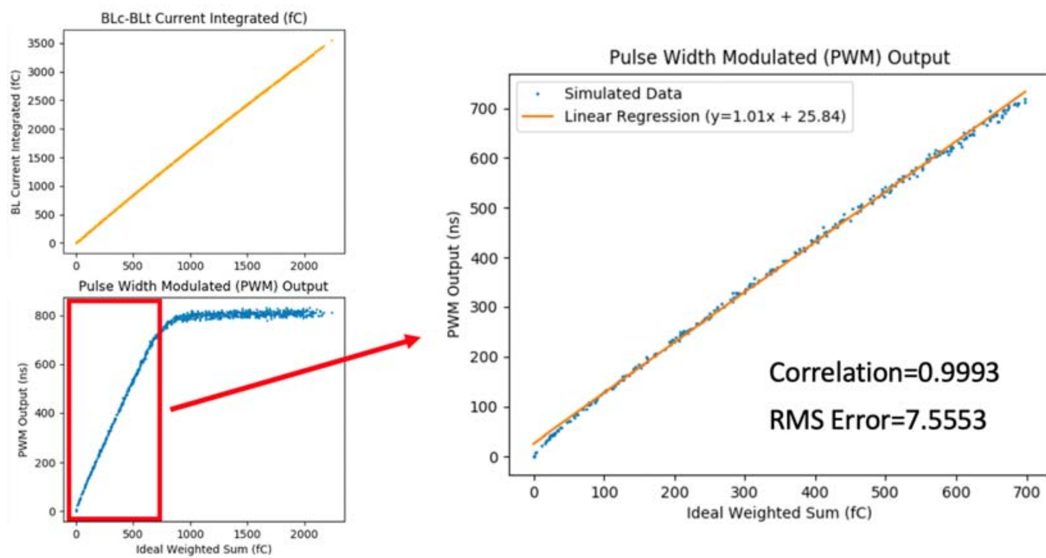


Fig. 19. The accuracy of the simulator was evaluated by applying 1000 different input vectors (samples) to network and evaluating the neuron output's accuracy. The image on the top left shows the 100 different inputs applied showing the measured integrated differential BL current. The bottom left plot shows the neuron's PWM output for those 1000 different input samples. A linear relationship between the ideal weighted sum and the actual neuron PWM output is shown for a subset of the input space while the neuron does saturate at some point above a certain threshold. The right-hand plot zooms in on the linear region and displays the correlation and RMS error statistics between the ideal weighted sum and neuron PWM output.

Figure 19 shows a clear linear relationship between the ideal weighted sum for random inputs and the neuron's pulse-width modulated (PWM) output duration. The neuron was designed to support a subset of the input space required for all desired applications. The neuron could be re-designed to support a larger input space before it saturates by increasing the neuron's capacitor amongst several other tradeoffs which we will not discuss here.

4) Key Outcomes

Throughout the course of this project, the objective has been to determine the robustness and resiliency of brain-inspired computing platforms relying on non-von Neumann architectures. In the past year, we have managed to study the effects of TID irradiation on both FDSOI and bulk CTT devices as well as RRAM devices. We have observed substantial effects in FDSOI due to hole trapping in the buried oxide (BOX) layer which is strongly coupled to the device channel. In contrast, we have observed less pronounced TID effects in bulk technologies such as GlobalFoundries 14LP where the dominate TID effect is trapping within the gate dielectric itself.

We have utilized various TCAD software including Ginestra to model the effects of irradiation on both CTT & RRAM devices which support all of our conclusions obtained from experimental data. We have observed that TID effects on the CTT device are more so technology-node dependent and not necessarily intrinsic to the CTT device itself. RRAM devices, while vary resilient to irradiation, require CMOS logic access circuitry which limits the overall performance of RRAM memory technologies under irradiation.

We have concluded our work by studying the effects of irradiation and other device non-idealities including device relaxation & subthreshold degradation on analog network inference accuracy utilizing the CTT. Our final effort has been to develop a CTT-hardware-based Inference Realistic Circuit Simulator (CIRCS) which allows us to consider the effects of both device non-idealities as well as circuit non-idealities when evaluating the performance of the CTT-based analog network. By providing a circuit-level simulation, we can consider all peripheral circuitry, device layouts, and parasitics to gather a more accurate understanding regarding how each of these device errors including V_{th} shifts due to irradiation may degrade network performance and accuracy.

Student Manuscript published in IEEE Transactions on Nuclear Science (TNS):

- “The impact of proton-induced single events on image classification in a neuromorphic architecture”, Rachel Brewer (vol. 67, no. 1, pp. 108-115, Dec. 2019)

References:

[1] X. Guo, *et al.*, “Fast, Energy-Efficient, Robust, and Reproducible Mixed-Signal Neuromorphic Classifier Based on Embedded NOR Flash Memory Technology,” IEDM 2017.

[2] G. Burr, *et al.*, “Recent Progress in Phase-Change Memory Technology,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 2016.

[3] X. Zheng, et al., “Error-Resilient Analog Image Storage and Compression with Analog-Valued RRAM Arrays: An Adaptive Joint Source-Channel Coding Approach,” IEDM 2018.

[4] Francesco Maria Puglisi, et al., “Bipolar Resistive RAM Based on HfO_2: Physics, Compact Modeling, and Variability Control,” IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 6, no. 2, pp. 171-184, Apr. 2016.

[5] Fujitsu Semiconductor, “Memory ReRAM: 4M (512K × 8) Bit SPI MB85AS4MT,” MB85AS4MT datasheet, Dec. 2016.

[6] Arshiya Anjum et al., “Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms,” vol. 379, pp. 265-271, 15 July 2016.

[7] F. Khan, E. Cartier, J. C. S. Woo and S. S. Iyer, "Charge trap transistor (CTT): An embedded fully logic- compatible multiple-time programmable non-volatile memory element for high-k-metal-gate CMOS technologies," IEEE Electron Device Letters, vol. 38, no. 1, pp. 44-47, Jan. 2017.

[8] A. Kerber, S. A. Krishnan and E. A. Cartier, "Voltage ramp stress for bias-temperature instability testing of metal-gate/high-k stacks," IEEE Electron Device Letters, vol. 30, no. 12, pp. 1347-1349, Dec. 2009.

[9] F. Khan, E. Cartier, C. Kothandaraman, J. C. Scott, J. C. S. Woo and S. S. Iyer, “The impact of self-heating on charge trapping in high-k-metal-gate nFETs,” IEEE Electron Device Letters, vol. 37, no. 1, pp. 88-91, Jan. 2016.

[10] P. J. McWhorter, S. L. Miller, and T. A. Dellin, “Radiation response of SNOS nonvolatile memory transistors,” IEEE Trans. Nucl. Sci., vol. 33, no. 6, pp. 1414-1419, Dec. 1986.

[11] E. S. Snyder, P. J. McWhorter, T. A. Dellin, and J. Sweetman, “Radiation response of floating gate EEPROM memory cells,” IEEE Trans. Nucl. Sci., vol. 36, no. 6, pp. 2131–2139, Dec. 1989.

[12] S. Gerardin, M. Bagatin, A. Paccagnella, K. Grürmann, F. Gliem, T. R. Oldham, F. Irom, and D. N. Nguyen, “Radiation effects in Flash memories,” IEEE Trans. Nucl. Sci., vol. 60, no. 3, pp. 1953-1969, Jun. 2013.

[13] D. M. Fleetwood, “Radiation-induced charge neutralization and interface-trap buildup in metal-oxide-semiconductor devices,” J. Appl. Phys., vol. 67, no. 1, pp. 580-583, Jan. 1990.

[14] I. S. Esqueda et al. “Modeling the radiation response of fully-depleted SOI n-Channel MOSFETs,” IEEE Trans. Nucl. Sci., vol. 56, no. 4, pp. 2247-2250, Aug. 2009.

[15] N. N. Mahatme et al., “Impact of back-gate bias and device geometry on the total ionizing dose response of 1-transistor floating body RAMs,” IEEE Trans. Nucl. Sci., vol. 59, no. 6, pp. 2966-2973, Dec. 2012.

[16] E. Simoen et al., “Radiation effects in advanced multiple gate and silicon-on-insulator transistors,” IEEE Trans. Nucl. Sci., vol. 60, no. 3, pp. 1970-1991, Dec. 2013.

[17] R. Carteret al., "22nm FDSOI technology for emerging mobile, Internet-of-Things, and RF applications," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 2.2.1-2.2.4.

[18] N. Rezzak et al., "Total-ionizing-dose radiation response of 32 nm partially and 45 nm fully-depleted SOI devices," 2012 IEEE International SOI Conference (SOI), NAPA, CA, 2012.

What do you plan to do during the next reporting period to accomplish the goals?

We will proceed with the current plan with finalizing efforts on the project. We plan on performing more exhaustive technology comparisons to study how irradiation interacts with various high-k technologies including both bulk and FDSOI technologies to fully understand the technology-node-dependent implications for the CTT device. We will further examine solutions such as back-gate biasing on 22FDX to counteract the effect of hole trapping in the buried oxide (BOX) layer on each device's corresponding V_{th}.

We will also prioritize work on the CTT-hardware-based Inference Realistic Circuit Simulator (CIRCS) so we can fully-evaluate the implications of irradiation on device and circuit performance as well as incorporate all other device and circuit non-idealities to study the overall system performance and accuracy under irradiation. Utilizing these methods, the direct effect of various radiation sources on a live analog neuromorphic can be measured and determined.

No Cost Extension Period

Major Activities

i. Establish reliability characteristics of CTT for use as analog synapse

Elaborating on prior work regarding the programming characteristics of the CTT, we have further explored the reliability components which set limits on maximum voltage conditions and programming pulse time durations that can be reliably applied to the device during programming. In addition, we have studied both immediate and long-term relaxation of the device which allows us to model how we expect each respective device's Vth (or σ_{ch}) to change as a function of time. We are able to purposefully over-program devices to counteract shallow de-trapping which occurs almost immediately after a programming cycle. Similarly, we have explored long-term charge de-trapping as a function of both time and temperature which allows us to determine how frequently a network may require weight updates. Another important component relating to the reliability of the CTT device is understanding the subthreshold degradation which occurs in some devices due to the considerable stress occurred across the device during programming.

ii. Irradiate Candidate CTT (22nm FDSOI) Devices and study effects of TID as well as mitigation techniques for reducing the impact of TID on device operation

GlobalFoundries 22FDX (FDSOI) wafers were quartered and diced to isolate several device test macros onto singulated dies which were bonded to hybrid ceramic packages with Au wire bonds. For 22FDX, approximately 50 macros each of various types were provided (single devices with width scaling, devices with length scaling, metal-oxide-metal capacitors, and ganged-gate devices). Results have shown that for FDSOI (specifically 22FDX), significant trapping occurred in the buried oxide (BOX) layer of the devices, which is highly coupled to the channel in FDSOI technology nodes. The induced trapping in the BOX caused a significant Vth-shift across all devices. This effect could be potentially mitigated by employing circuitry to leverage the back-gate available in 22FDX effectively depending on the systematic nature of the radiation-induced Vth-shift across all devices.

iii. Irradiate 14LP Bulk FinFET Devices for comparison with 22FDX TID Study Results

GlobalFoundries 14LP (bulk) wafers were quartered and diced to isolate several device test macros onto singulated dies which were bonded to hybrid ceramic packages with Au wire bonds. For 14LP, several ganged-gated configurations were provided with varying Vth-implant (slvt, lvt, rvt, & hvt), gate-pitch (78nm, 84nm, 190nm, & 310nm), and number of fins (2, 4, & 40 fins). All irradiation studies were performed at Vanderbilt University using their ARACOR 10-keV

X-ray irradiator system. Results show that bulk FinFET devices are substantially less susceptible to TID than in FDSOI technologies; however, multi-fin devices show increased subthreshold leakage with increasing TID exposure.

iv. Develop physics-based models for confirming observed radiation effects on CTT

A physics-based model was developed using TCAD software to explore the effects of radiation on FDSOI devices. The results obtained in this simulation are provided in detail under Section 3: Significant Results. The underlying result supporting hole trapping in the buried oxide (BOX) layer as the dominate effect during TID. This effect is especially pronounced in FDSOI

technology nodes due to the strong coupling between the BOX layer and the device channel.

v. Explore RRAM Physics & Radiation Effect Modeling

A HfOx-based RRAM cell was modeled by the stacked structure shown in the figure below with a 5nm HfOx switching layer followed by a 1nm TiOx oxygen reservoir. The switching layers were contacted by TiN layers as inert electrodes. The area of the cell is 10nm × 10nm. The oxygen vacancies are considered uniformly distributed in the HfOx layer with a density of 1×10^{19} cm^{-3} in the initial state. The initial position of each defect is randomly generated throughout the HfOx layer. Extra oxygen vacancy/ion pairs are randomly introduced throughout the HfOx layer to mimic the effect of TID. As shown in **Fig. 1**, 40 oxygen ions/vacancies (O-/Vo+) pairs are generated in the one region of the device. The set/reset process has been simulated on both damaged and undamaged RRAM cells to investigate the effect of DD on the RRAM performance.

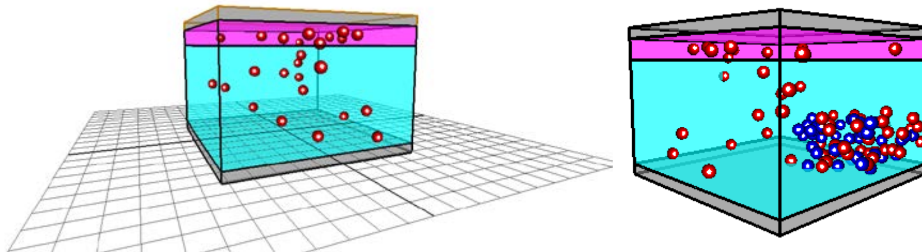


Fig. 1. *Simulated RRAM structure.*

vi. Irradiate Commercial RRAM chip

We irradiated a commercial RRAM chip from Fujitsu and studied how the device characteristics of a large RRAM array shifted with radiation dose. It is important to note here that we were particularly interested in observing how the integrated CMOS overhead influences the performance of the RRAM array given that the CMOS logic access circuitry is required in order to utilize the RRAM technology.

vii. Investigate mitigation techniques for general NVM-based analog networks

Mitigation techniques for general NVM-based analog networks have been explored as shown in **Fig. 2**. Simply by considering an additional ΔW term when training the network weights—given that exact weights cannot be programmed into a CTT array due to intrinsic device variation in an analog network—we can train the network to be more resilient to both programmed device variation as well as device relaxation over time.

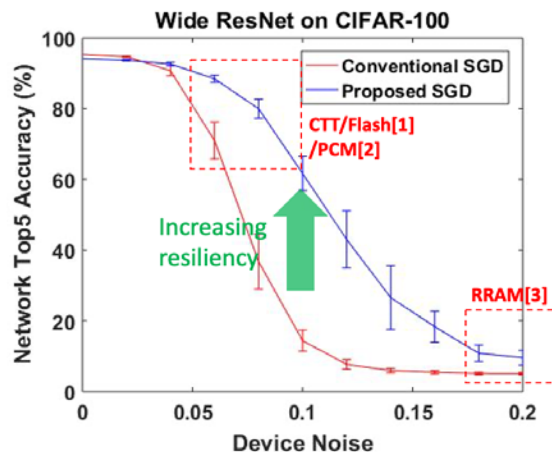


Fig. 2. Network Top5 Accuracy vs. Device Noise utilizing hardware-aware stochastic gradient descent training algorithm.

viii. Develop software-level training & inference simulator

In order to understand the effect of CTT variation, device relaxation, etc., we developed a software-level Matlab-based simulator which allows us to (1) train a network to map onto our twin-cell CTT architecture and (2) evaluate inference performance considering various device variations which generally lower system inference accuracy. This allows us to evaluate specifically how much of an effect device variation has on individual network architectures and target problem sets (e.g. MNIST). This also allows us to perform Monte Carlo simulations to see how a particularly trained network may be expected to perform across several different chips, all with different device variations. This work also includes some modifications that have been made to the training algorithm itself to minimize the effect of device variation on the expected output. Given that each individual CTT must be programmed to a roughly precise state, it should not necessarily be trained in the same way a digital network may be trained. We have modified the training loss function to account for the fact that we cannot exactly program or encode the weights obtained during the training process into an array of CTT devices—there will always be some variation between the trained set of weights and actual programmed weights within the CTT array.

ix. Develop CTT-hardware-based Inference Realistic Circuit Universal Simulator (CIRCUS)

We have also begun work towards developing a circuit-level simulator (CIRCUS) which allows us to not only consider the effects of device variation but also circuit-level variations including post-layout parasitics, process corners, temperature, leakage and other non-idealities including bondwire inductance. The circuit-level simulator allows us to more accurately demonstrate the inference process by deploying circuit-level models for all peripheral circuitry including the Word Line Driver (WLD) and Neuron. It allows us to quickly iterate over the entire network test dataset and evaluate the performance of the network given all of the known circuit non-idealities (specifically in regard to the peripheral circuitry) and the CTT device non-idealities. It also allows us to perform Monte Carlo simulation to assess how a specific network may be expected to perform across a distribution of chips and determine the effect of device relaxation to give us an idea of when a chip may require weight updates or a refresh cycle. Radiation effects are modelled as additional variation in the threshold voltage of the device. As noted elsewhere the source of radiation induced variation is regional and due to trapped charge in the buried oxide – not the gate oxide.

2) Specific Objectives

- Study technology-dependent effects of TID irradiation on CTT devices (by comparing irradiation experiments leveraging GlobalFoundries 14LP and 22FDX technologies)
- Develop subsequent charge-trapping models utilizing available experimental data to incorporate mature device relaxation models, TID-induced ΔVth and other long-term device effects, and endurance
- Deploy charge-trapping models in analog neural network simulations & more complex circuit-level simulations to study effect
- Study effects of device error on different types of deep learning network architectures & classification problems (e.g. image classification, object detection, segmentation, audio/keyword recognition, etc.)
- Determine what types of problems and which network architectures the CTT is most suitable for
- Develop Circuit-Level simulator to simulate the effects of device non-idealities such as programming error (Vth variation), device Vth relaxation and TID-induced hole-trapping & ΔVth as well as circuit/implementation non-idealities such as parasitics, process corners, BL-biasing instability, common-mode feedback, integrator design and gain-bandwidth product, etc.
- Irradiate commercially available RRAM chips and determine their resiliency
- Explore physical models governing the operation of RRAM cells using TCAD simulation and determine whether logic access circuitry is the bottleneck of RRAM chip performance under irradiation

3) Significant Results

i. Establish Programming & Reliability Characteristics of CTT

Shallow de-trapping is observed to occur immediately after programming. **Figure *3*** shows that this de-trapping or relaxation effect increases the more a device is programmed. A relaxation compensation method has been developed to 'over-program' devices to compensate for the large relaxation of the device which occurs within ~1hr of programming each device. This relaxation compensation method leads to a stable & systematic correction in achieved programmed/target weights after measuring 200 weights approximately 200 hours after the initial programming step.

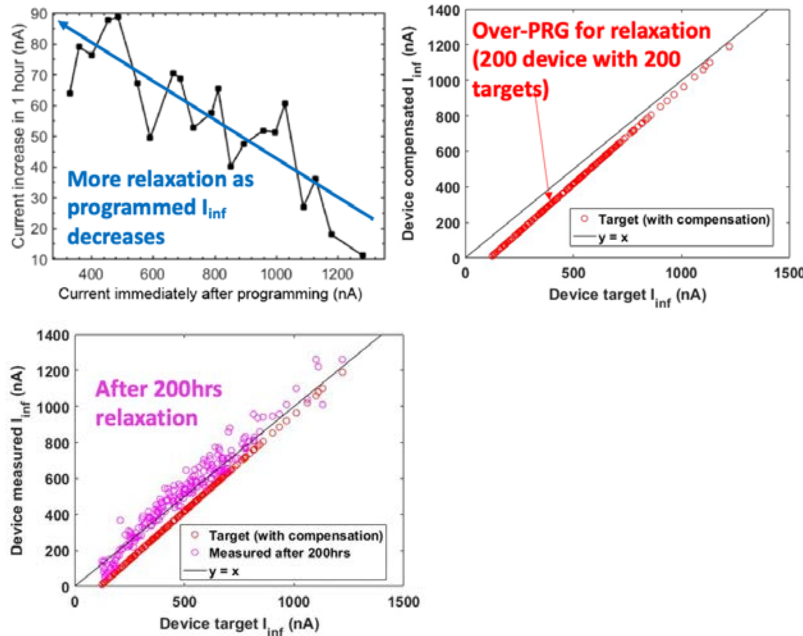


Fig. 3. *(a) CTT relaxation 1 hour after programming as a function of target program current. (b) Demonstrating CTT Relaxation compensation where 200 CTT devices are purposefully over-programmed to account for any immediate loss of charge due to shallow de-trapping. (c) End result of CTT Relaxation Compensation after 200 hours of relaxation. Charge de-trapping is exponential in-time leading to large initial de-trapping after programming each device.*

Subthreshold degradation is also an issue worth considering given both its effect on increased BL leakage as well its reliability implications. **Figure 4** depicts the ideal I_{DS}-V_G curve after a programming step followed by an erase step which returns the Vth closer to the virgin or baseline state. **Figure 5** then elaborates on this by demonstrating simply shifting each I_{DS}-V_G curve by the programmed ΔVth may not be sufficient given that some devices may experience subthreshold degradation. The effect of subthreshold degradation on network performance cannot be assumed to be negligible and must be studied.

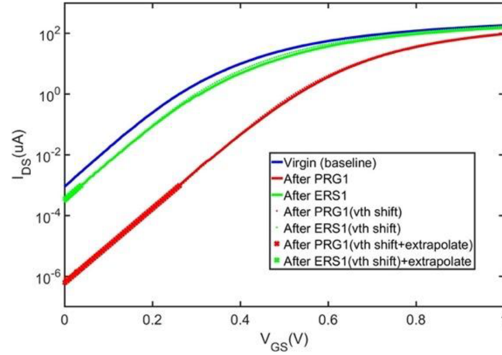


Fig. 4. *Example 22nm CTT device where a virgin device IDS-VG curve is shifted by the corresponding programmed ΔVth after a program (PRG1) followed by an erase (ERS1) step. This plot does not accurately depict the subthreshold degradation which occurs across some devices after stressing the device under large bias conditions during programming and erase conditions (see* ***Fig. 5****).*

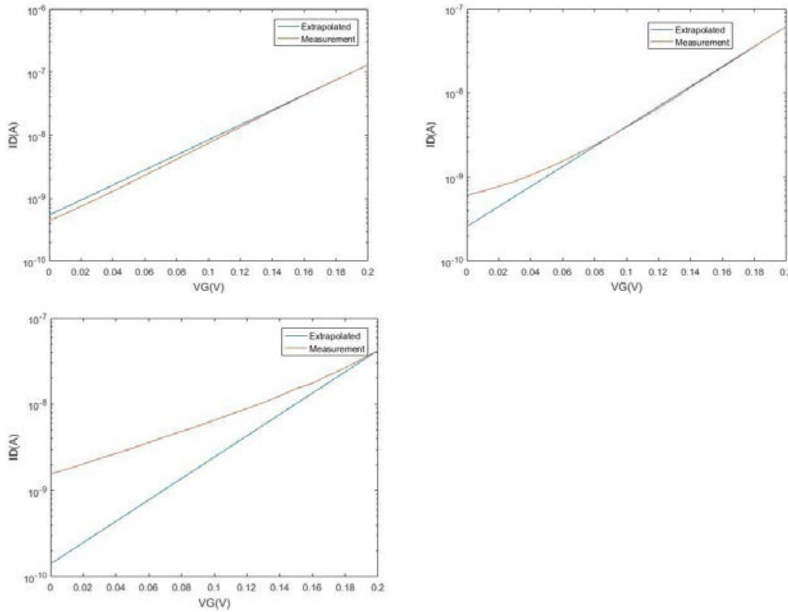


Fig. 5. *Three sample CTT devices with varying subthreshold degradation leading to a difference in subthreshold current between the actual subthreshold current (measurement) and the expected subthreshold current obtained by simply shifting the entire Ids-Vg curve by the programmed Vth.*

It turns out that given that our architecture reverse biases off devices during inference by providing a gate voltage of -0.3V, the effect of subthreshold degradation is largely negligible on network performance. We need only consider the effect of increased subthreshold current when analyzing the rise & fall times associated with the constant-amplitude pulse-width modulated (PWM) input waveforms as shown in **Fig. 6**.

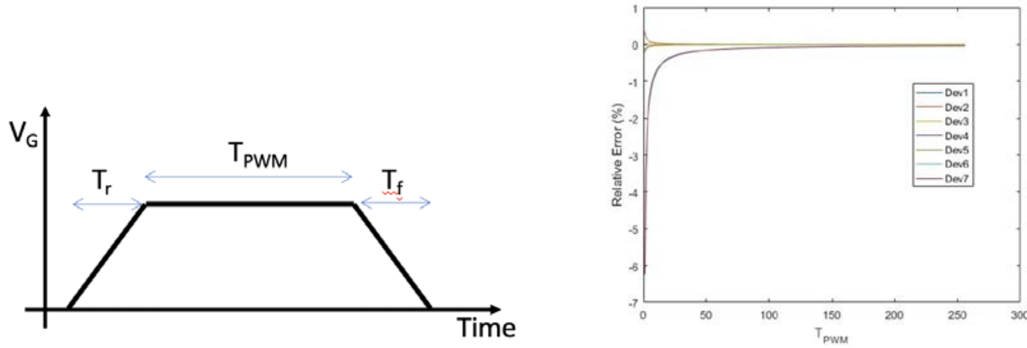


Fig. 6. (a) Constant-amplitude pulse-width modulated (PWM) input waveforms applied to each WL in the CTT array. Each input waveform is a function of the input (TPWM=NINPUT x Tcycle) and the finite rise and fall times associated with that waveform. The middle part of the waveform is characterized at a constant VG bias while the rise and fall portions of the waveform have varying VG and hence varying σch—which is a function of VG. The longer the rise & fall times are, the larger the error due to subthreshold degradation (increased leakage current) is. (b) Relative error due to subthreshold degradation across 7 different devices, given the rise & fall time of the PWM input waveforms. Larger inputs (e.g. longer TPWM) reduces the effect of increased subthreshold degradation on the overall output. This error in most cases (for most inputs and devices) is <1% with one device showing up to 6% area for a PWM input of 1 and reducing for larger inputs. The input space supports inputs between 0 and 255.

ii. Develop model for assessing the effect of radiation on CTT

TCAD simulations were used to investigate the charge-trapping mechanisms responsible for the programming and radiation responses. Programming was simulated through the addition of the density of trapped electrons at the upper transistor interface with the gate dielectric until bias-induced shifts were reproduced. For the data of **Fig. 10**, for example, the inferred trapped-electron density projected to the gate oxide/Si interface is ~6×10^{12} cm^{-2}. Similarly, radiation-induced shifts were simulated through the addition of trapped holes at the lower interface of the transistor with the buried oxide until the TID-induced shifts were reproduced. For the data of **Fig. 10**, the inferred trapped-hole density projected to the buried oxide/Si interface is ~ 1.5×10^{12} cm^{-2}.

Figure 7 plots the simulated charge densities in the transistor channel for four simulations conducted on a representative 22 nm FDSOI TCAD model based on the process details of **[17]** and the experimental data of **Fig. 10**. The simulated ID-VG curves that match the experimental results of **Fig. 10** are shown in **Fig. 8** as a function of gate voltage. The simulated cases include a “pre” case where no additional carriers were introduced into FDSOI transistor model, a “programming” case with electrons in the gate oxide, an “irradiation” case with holes in the buried oxide, and a “combined” case with electrons in the gate oxide and holes in the buried oxide. As expected, the “programming” case results in a lower channel electron density (transistor turned more strongly off), the “irradiation” case shows a higher electron density in the channel (transistor turned strongly on), and the “combined” case is turned on more strongly than the “pre” case but not as strongly as the irradiation-only case. These results show that a self-consistent model of the programming and irradiation responses of these devices can be built, with plausible charge densities in the gate dielectric and buried oxide, under the simple assumptions stated above. In the full paper, the evolution of charge densities during more complex programming and irradiation sequences will also be illustrated.

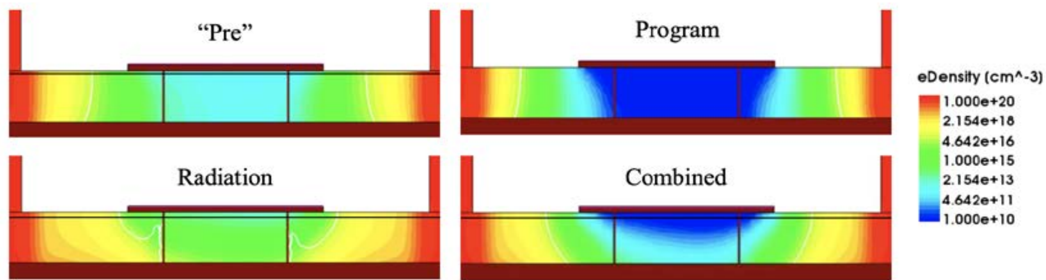


Fig. 7. *Representative 22nm FDSOI TCAD model based on [17] showing simulated electron density for an un-irradiated & un-programmed CTT (top left), programmed CTT (top right), irradiated CTT (bottom left), and a programmed CTT that has been irradiated (bottom right).*

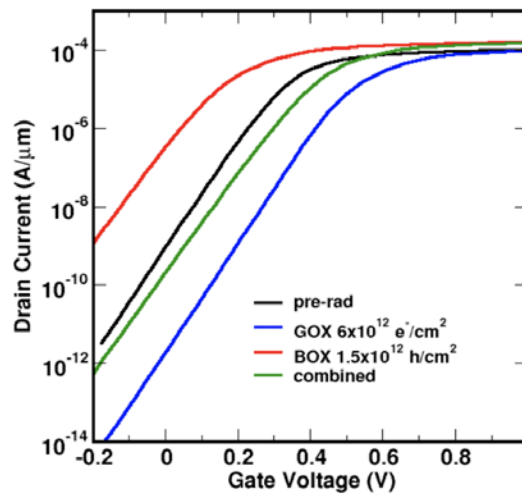


Fig. 8. *TCAD model showing simulated results of "pre" transistor responses with respect to a simulated pre-irradiation, pre-programmed CTT (black). The blue curve simulates programming, and red curve simulates radiation. The combination of simulating both programming and radiation results in the green curve.*

iii. Irradiate FDSOI CTT devices (GlobalFoundries 22FDX)

Weak programming before irradiation: **Fig. 9(a)** shows I_D-V_G curves for Device A, programmed before irradiation. Programming the device resulted in a positive V_{th} shift of 50 mV, as shown by the solid black (pre-programming) and dashed-black (post-programming) curves. The TID response of Device A after programming is shown in the remaining curves. After 50 krad(SiO_2) irradiation (blue curve), V_{th} shifted negatively by ~40mV, which nearly offsets the initial programming. Irradiating the devices up to 500 krad(SiO_2) resulted in V_{th} shifts of ~ -140mV.

Weak programming after irradiation: **Fig. 9(b)** shows I_D-V_G curves for Device B, programmed after irradiation. The solid black curve again shows the pre-irradiation, pre-programming response. The remaining solid curves show the TID response of the device. V_{th} again shifted negatively by ~40mV after the device was irradiated to 50 krad(SiO_2), and by ~ -155mV after irradiation to 500 krad(SiO_2). The post-irradiation programming resulted in a positive Vth shift of 30 mV. The similarity of the TID response for programmed and unprogrammed devices in **Fig. 9 (a)(b)**, and the relative similarity of the V_{th} shifts during programming for the unirradiated and irradiated devices, suggest that the charge trapping responsible for the TID response and programming may not be interacting strongly, at least to first order. If so, this would contrast with the responses of Flash memory devices, for example, where radiation- induced charge

neutralization [5]-[7] can lead to rapid loss of the programmed state. We now explore these mechanisms in more detail for devices programmed to more positive V_{th} values.

Strong programming before and after irradiation: **Fig. 10** shows V_{th} shifts for a device programmed before irradiation, irradiated to 500 krad(SiO_2), annealed at room temperature for 52 h, and re- programmed. The initial programming in this case was stronger (higher voltage, longer time) than for the devices of **Fig. 9(a)**, and resulted in a positive 205 mV V_{th} shift. After this programming, irradiation to 500 krad(SiO_2) led to a -145mV shift, similar to the post-programming TID response in Fig. 1. Annealing led to a positive 20 mV V_{th} shift. The device was then programmed again in the same way as before irradiation. This programming sequence resulted in only a positive 55 mV shift (135 mV with respect to baseline). These results indicate that the post-irradiation and programming response of the device can be quite sensitive to device history.

For all cases in **Figures 9** and **10**, the TID response of these fully-depleted devices is expected to be dominated by hole trapping in the buried oxide [14]-[16]. The presence or absence of trapped charge in the gate dielectric does not significantly affect the electric field in the BOX, leading to similar amounts of charge trapping when devices are irradiated before or after programming. However, the presence of pre-existing programming and/or radiation-induced charge in the gate dielectric can limit the ability of dielectric layers to capture additional charge due to trap filling, charge neutralization, and pulse modification effects [8], [13].

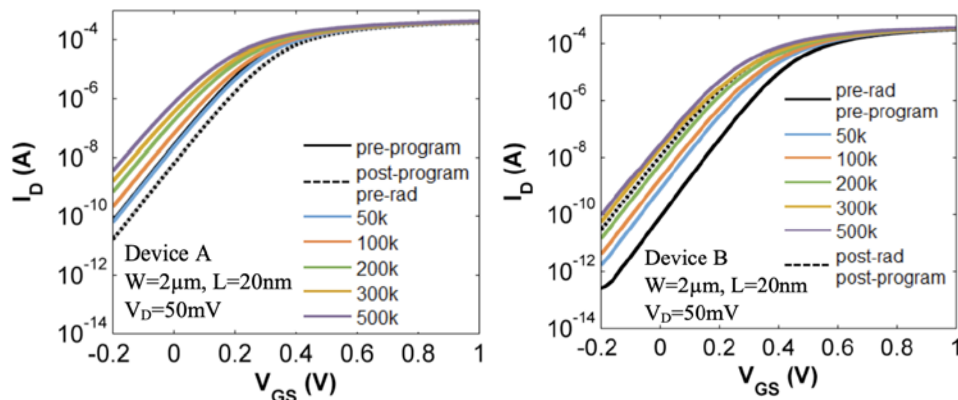


Fig. 9. *(a) Drain current as a function of gate voltage for Device A programmed before irradiation. The solid black curve (partially hidden by the blue curve) shows the pre-irradiation, pre-programmed DC sweep. The dashed black curve shows the response of the device after programming. The other five curves show the post-programming radiation response of the device. (b) Drain current as a function of gate voltage for Device B after irradiation. The solid black curve shows the pre-irradiation, pre-programmed, radiation-induced shifts. The dashed blacak curve is the post-programming response of the device after irradiation to 500 krad(SiO_2) (purple curve) and subsequent programming.*

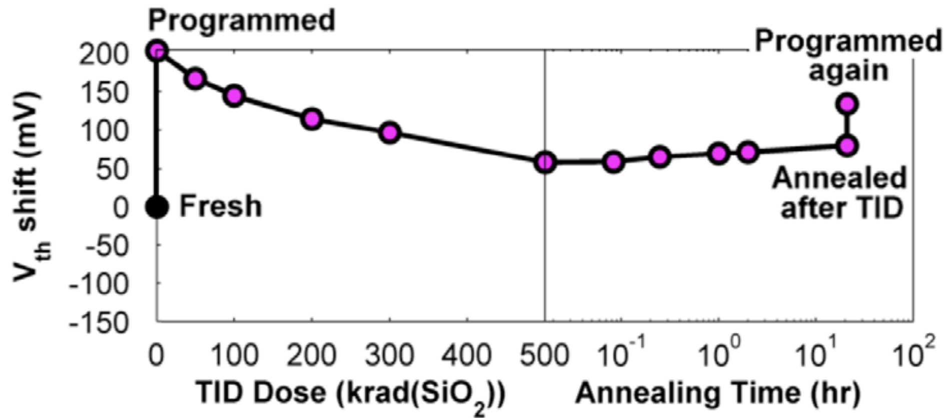


***Fig. 10.** Threshold voltage shifts of a device as it is programmed, irradiated, annealed, and programmed again.*

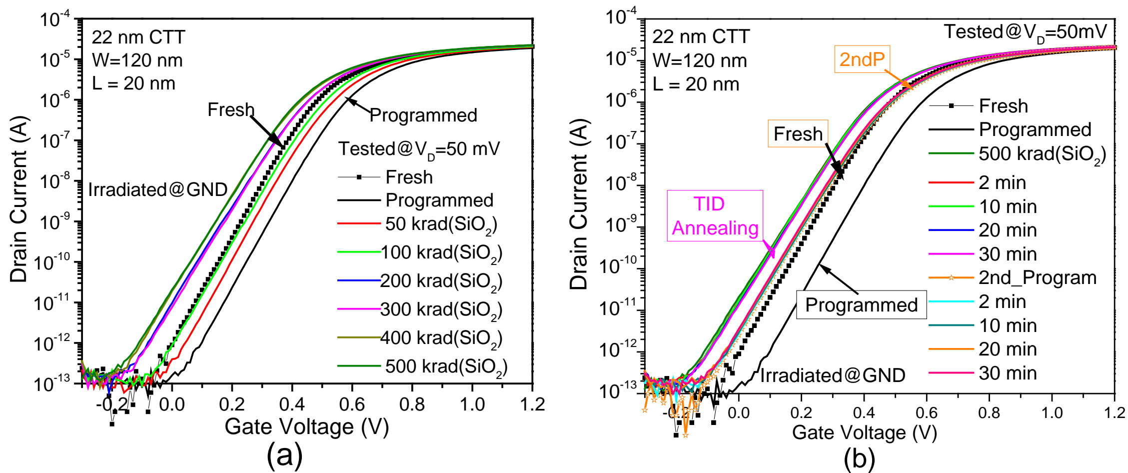


***Fig. 11.** <u>Programmed First then Irradiated:</u> Typical drain current as a function of gate voltage for 22 nm FD SOI CTT devices after programming, (a) irradiation to 500 krad(SiO_2), and (b) 30 minutes of annealing, second programming, and an additional annealing.*

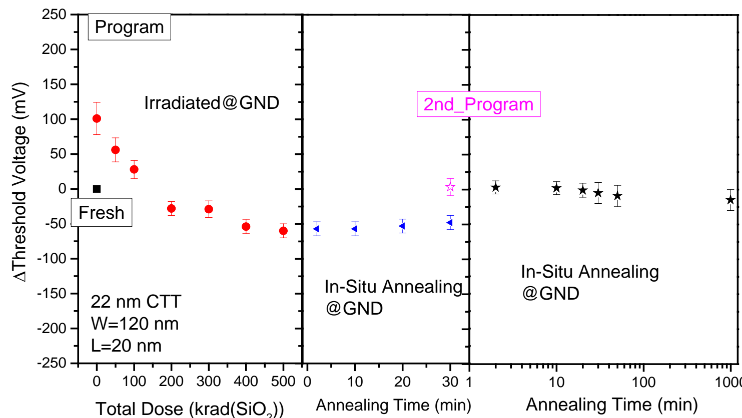


***Fig. 12.** Extracted threshold voltage shifts with error bars denoting standard deviation for three 22 nm FDSOI CTT devices programmed an irradiated using sequences similar to those shown in **Fig. 11**. Fresh devices were (1) initially programmed, (2) irradiated up to 500 krad(SiO_2), (3)*

annealed (in-situ) for 30 minutes, (4) programmed for a second time, and (5) followed by additional (in-situ) annealing.

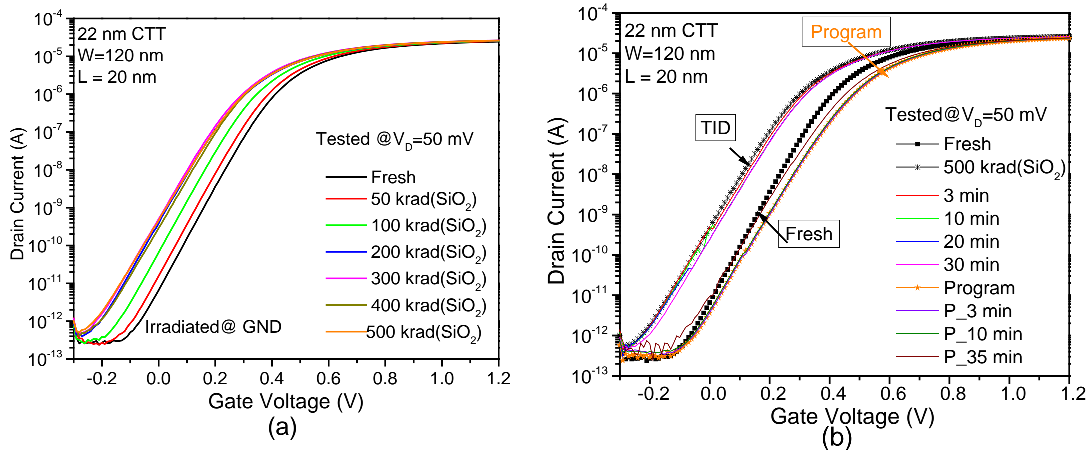


Fig. 13. *Irradiated First: Typical drain current as a function of gate voltage for 22 nm FD SOI CTT devices after (a) irradiation to 500 krad(SiO_2), and (b) 30 minutes of annealing and programming followed by additional annealing.*

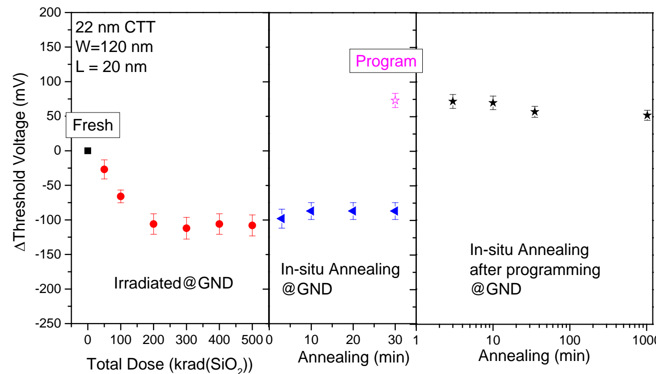


Fig. 14. *Extracted threshold voltage shifts for three 22 nm FDSOI CTT devices programmed and irradiated using sequences similar to those shown in **Fig. 13**. Fresh devices were (1) initially programmed, (1) irradiated up to 500 krad(SiO_2), (2) annealed (in-situ) for 30 minutes, (3) programmed, and (4) followed by additional (in-situ) annealing.*

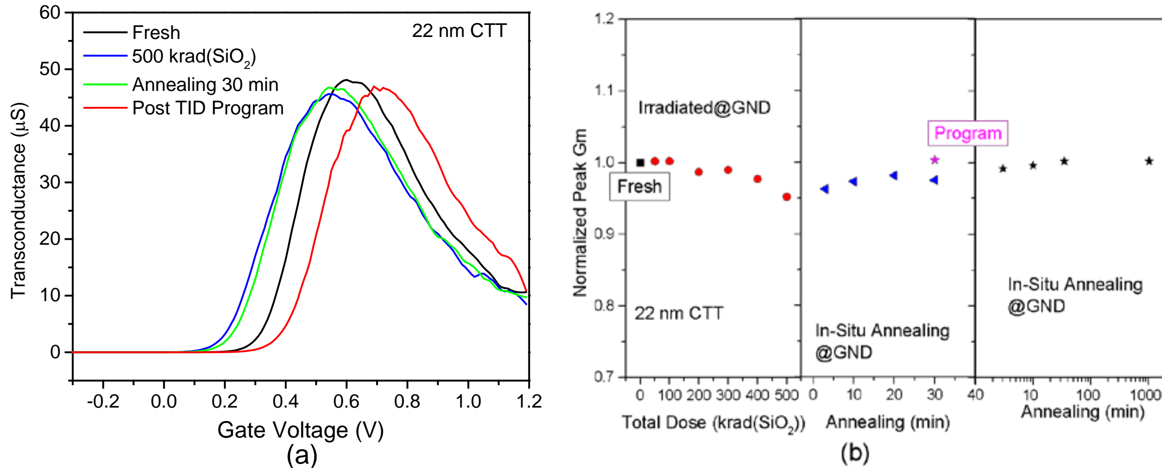


Fig. 15. *(a) selected G_m-V_G curves and (b) peak transconductance normalized to the initial ('fresh') state for the 22 nm FD SOI CTT devices.*

Irradiation of 22nm CTTs fabricated in an FDSOI technology results in threshold voltage shifts between 40mV and 155 mV for doses between 50 and 500 krad(SiO_2). These shifts are due most likely to charge trapping in the buried oxide **[15], [18]**. The programming threshold voltage shifts are due primarily to electron trapping in the gate oxide. To first order, the charge distributions in the gate dielectric and buried oxide do not interact strongly, enabling the construction of TCAD models with plausible charge densities. In this summary, inferred densities of trapped electrons are derived for a strong programming pulse, and inferred densities of trapped holes in the buried oxide are derived for devices irradiated to 500 krad(SiO_2). Additional 22nm device results are shown in ***Figures 11-15***.

v. Irradiate bulk FinFET CTT devices (GlobalFoundries 14LP)

We also irradiated unprogrammed GlobalFoundries 14LP devices and studied the effects of irradiation on a non-FDSOI technology for comparison. ***Figure 16*** shows a substantially smaller Vth shift for the 14nm bulk technology compared to previously shown 22FDX (FDSOI) results.

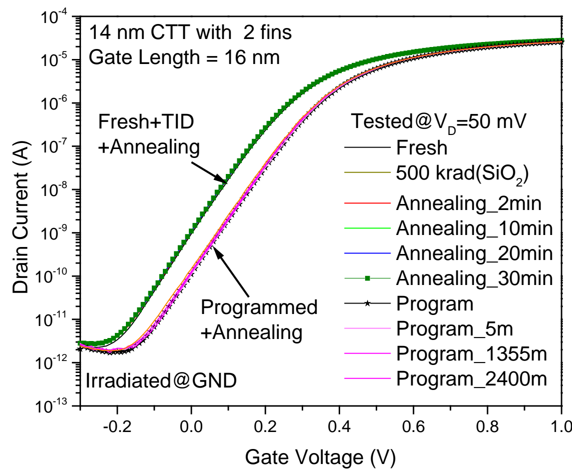


Fig. 16. *Typical drain current as a function of gate voltage for 14 nm CTT devices with 2 fins after irradiation to 500 krad(SiO_2), 30 minutes of annealing, and programming.*

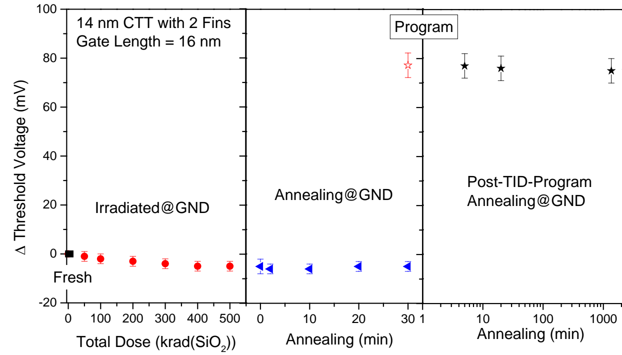


***Fig. 17.** Extracted Vth shifts for three 14 nm CTT devices with 2 fins. Fresh devices were (1) irradiated up to 500 krad(SiO_2), (2) annealed (in-situ) for 30 minutes, (3) programmed by applying PVRS (V_g, V_d) programming pulses to the devices, and (4) annealed (in-situ) for an additional 2400 minutes.*

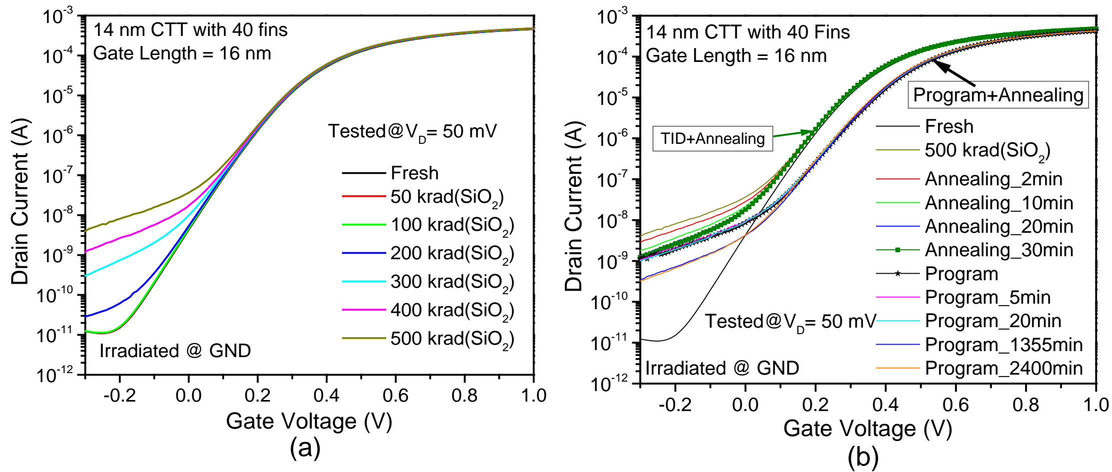


***Fig. 18.** Typical drain current as a function of gate voltage for 14 nm CTT devices with 40 fins after irradiation to 500 krad(SiO_2), 30 minutes of annealing, and programming.*

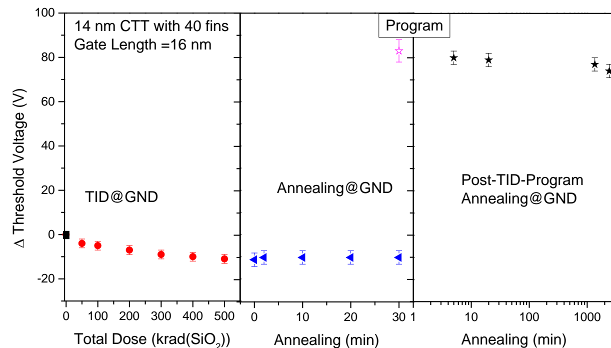


***Fig. 19.** Extracted Vth shifts for three 14 nm CTT devices with 40 fins. Fresh devices were (1) irradiated up to 500 krad(SiO_2), (2) annealed (in-situ) for 30 minutes, (3) programmed by applying PVRS (V_g, V_d) programming pulses to the devices, and (4) annealed (in-situ) for an additional 2400 minutes.*

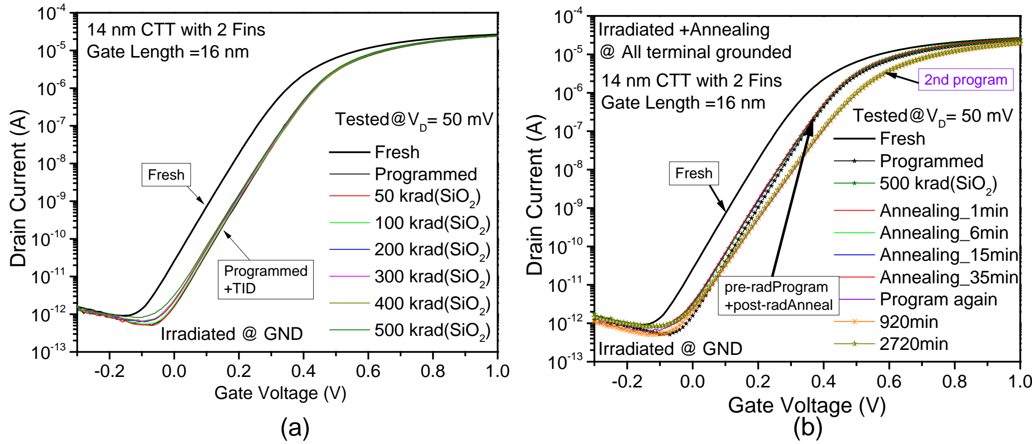


Fig. 20. Typical drain current as a function of gate voltage for 14 nm CTT devices with 2 fins (programmed before irradiation). (a) Device is first programmed then irradiated to 500 krad(SiO_2). (b) After irradiation, device is annealed for 35 minutes, programmed again, and in-situ annealed for an additional 2,720 minutes.

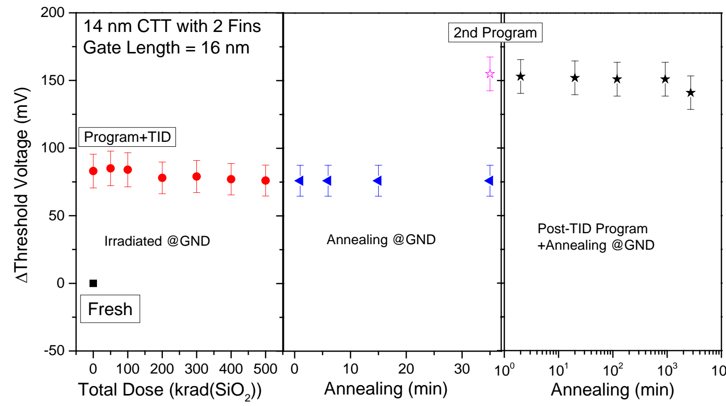


Fig. 21. Extracted V_{th} shifts for 14nm devices with 2 fins (from Fig. 15) programmed first then irradiated.

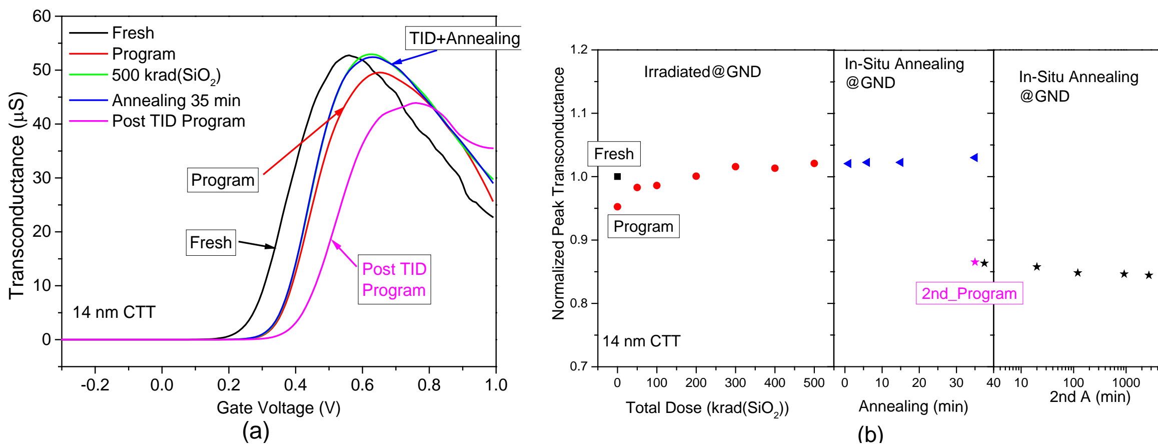


Fig. 22. Peak transconductance plots for 14nm devices with 2 fins from Fig. 15. (a) selected Gm-VG curves and (b) peak transconductance normalized to the initial ('fresh') state.

These results have significance. FDSOI in general is designed to have greater back gate sensitivity. This back-gate sensitivity is accentuated by improving the electrostatics of the back-

gate bias by thinning the box to tens of nm. Unfortunately, this also increases the radiation sensitivity of the device. A significant conclusion of this work is that while FDSOI has many advantages in low power devices especially for the IOT application space, and also exhibits a robust CTT effect (primarily because of the thermal isolation afforded by the BOX), It is not a preferred technology in high radiation environments. In those cases where the possibility of high radiation exists, it would be preferable to use partially depleted SOI (PDSOI). We have shown earlier **[19][20]**, that PDSOI has a comparable CTT effect to FDSOI, and independent studies on PDSOI, show that the radiation tolerance is, in fact acceptable. Clearly, CTT devices in bulk technologies have higher radiation tolerance as shown here, but concomitantly show smaller CTT effects. Hence our recommendation for CTT technology would be to use FDSOI in low radiation or well shielded environments and PDSOI for higher radiation environments. From a design perspective, our FDSOI based CTT neuro-engines are easily ported to PDSOI, with little performance degradation. Our reason for using FDSOI is primarily due to easier availability through GlobalFoundries. Porting to PDSOI, would require a thicker box substrate, and elimination of any back-bias control. For this and other reasons, we have not employed back-bias control in our chips.

It is also important to point out that larger (e.g. nFins=40) bulk FinFET devices suffer from increased subthreshold leakage for doses above ~200 kRad(SiO_2) as shown in **Fig. 18**. This is likely due to charge-trapping in the shallow trench isolation (STI) in proximity to the sub-fin regions of the devices. Increased subthreshold leakage during TID irradiation is not observed for devices with 2 fins, shown in ***Fig. 16*** and ***Fig. 20***.

In ***Fig. 22***, it is demonstrated that the increase in subthreshold slope is accompanied by a significant reduction in peak G_M. Hence, the increased subthreshold slope for the 14 nm FinFETs is due most likely to the generation of interface traps.

v. Negligible effects on RRAM cell observed in simulation

The results of the simulation study suggest that while the radiation damage can affect defect density, filament position and the set-reset process of an RRAM device, it has relatively small effect on the value of R_{ON}/R_{OFF}. Therefore, RRAM is relatively robust against radiation for digital applications.

Figure *23* displays the Set-Reset cycle for an undamaged device. O^-/Vo^+ pairs are generated almost randomly throughout the whole device area, and the conductive filament is eventually formed in the area with the highest Vo^+ density. For the damaged device—shown in **Fig. *24***—the conductive filament is formed at the damaged position after the Set process, and the generation of O^-/Vo^+ pairs in the undamaged regions is very limited. This is because the region damage by radiation creates a 'weak spot' in the HfOx layer, similar to time-dependent dielectric breakdown (TDDB). The filament formation in the damaged device is highly localized to the damaged region due to the positive feedback of increased local electric field and self-heating. The undamaged device ends up having more defects after the Set process (state 2) than what is observed in a damaged device due to the highly localized filament formation in the damaged

device.

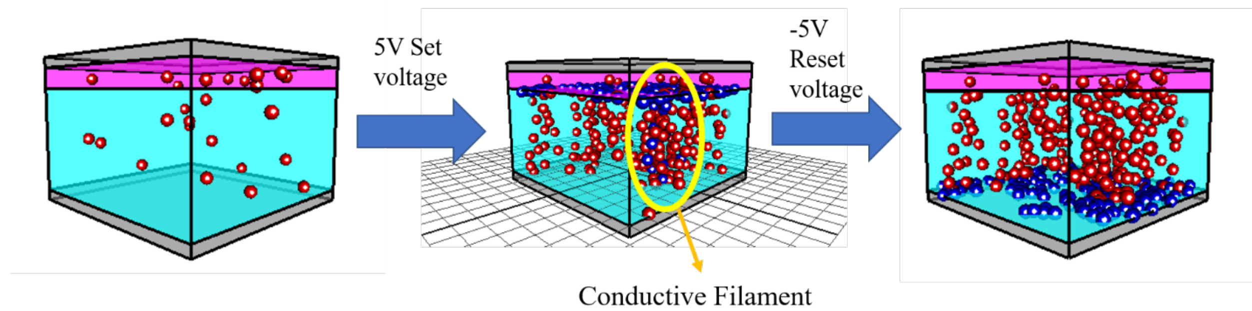


Fig. 23. *Set-Reset cycle for undamaged RRAM cell.*

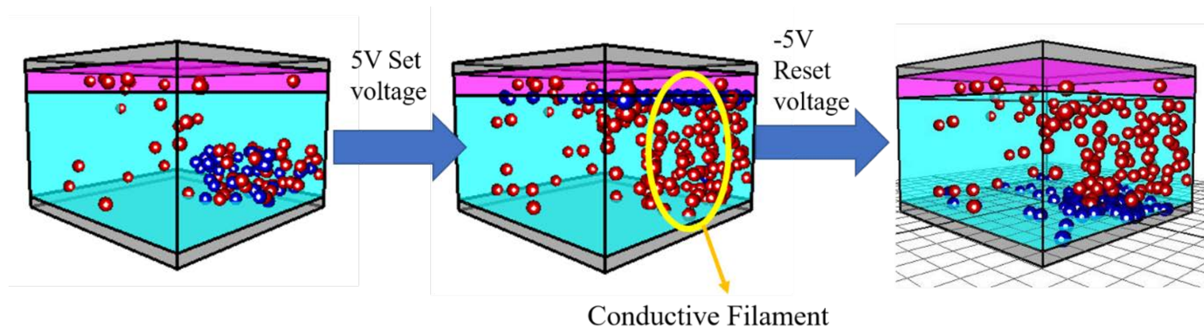


Fig. 24. *Set-Reset cycle for damaged RRAM cell.*

The I-V characteristics of the devices are shown in **Fig. 25**. During the Set process, the current at low bias is larger for the damaged bias compared to the undamaged device. This difference is caused by the introduction of extra defects at the beginning of the Set process. As the voltage increases, the currents of the damaged and undamaged devices become almost identical because the current conduction is localized once a filament is formed. Therefore, although a damaged RRAM device has more localized filament formation and fewer defects, this will have little effect on the characteristics of an RRAM device in a Low-Resistance (R_{ON}) state. During the Reset process, the O^-'s located in the TiOx reservoir region will be driven out due to the electric field and recombine with the Vo^+'s in the filament; however, only a small part of the conductive filament is dissolved for both cases. The I-V characteristics are shown in **Fig. 25**. Compared to the Set process, a larger difference between damaged and undamaged devices is observed in the Reset process. When the reset starts, the undamaged device has more defects and is more conductive. The damaged device has most of the defects located in the filament region, so the filament dissolves faster than in the undamaged device. As the Reset process continues, the filament dissolution becomes self-limited because the increase of the voltage leads to a decrease in the conductivity of the devices. Therefore, both devices reached a similar current level when the reset voltage is about -5V.

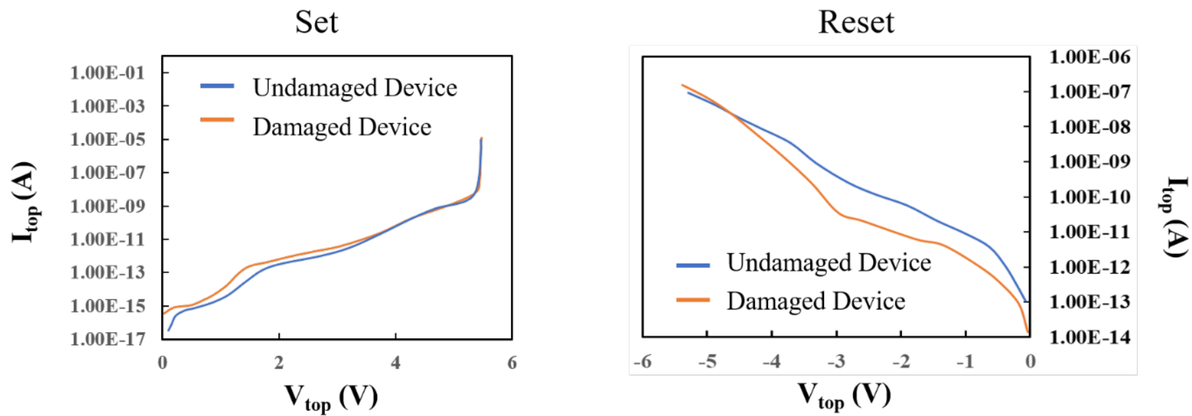


Fig. 25. I-V characteristics (Set/Reset) for both undamaged and damaged RRAM cells.

vi. RRAM performance under irradiation is limited by the CMOS logic access circuitry
The CMOS control circuits were found to be more vulnerable to TID than the RRAM cells themselves. As shown in **Table 1**, the DUT (4Mb RRAM, Fujitsu MB85AS4MT) showed a 100% success rate at 50 krad(Si); however, after further increasing the radiation dose above 50 krad(Si), a zero success rate was observed—indicating a functional failure. The functional failure is characterized by loss of communication with the DUT to program and read data. There were no errors observed within the RRAM memory array prior to observing the functional failure. The behavior of the RRAM and the operation of RRAM under the effect of radiation was also simulated using the Ginestra software package. The RRAM cells were determined to be fairly robust to radiation in simulation, and the irradiation test results suggest that the CMOS control circuitry is more vulnerable to failure than the RRAM cells themselves at these particular TID levels.

Table 1 shows the success rate of programming and reading data from the test sample. At 0 and 50 krad(Si), the success rates for all four procedures were 100%, while they were 0% at 100 krad(Si). This result clearly indicates that the test sample was functional up to 50 krad(Si), and its functionality started to degrade beyond 50 krad(Si) TID.
The simulation results provide further evidence that the degradation of the device is due to the sensitivity of the CMOS control circuitry under the effect of radiation. As shown in **Fig. 8**, the damaged device has lower current at low bias, but eventually both devices reach similar current levels. Therefore, we conclude that there is little effect on the characteristics of an RRAM cell in both the Low-Resistance (R_{ON}) and High-Resistance (R_{OFF}) states. The results of the simulation indicate that RRAM is relatively robust to radiation—especially for digital applications. Meanwhile, it is shown in **[6]** that CMOS will have a >0.5V V_{th} shift after being exposed to 100 krad proton dose. It is possible that this V_{th} shift will lead to failure in the CMOS control circuitry. Therefore, it is more likely that the chip failed because of failures within the CMOS control circuitry rather than failures within the RRAM memory array itself. This also explains the immediate transition between 100% success rate to 0% success rate as the data in the RRAM memory array is no longer accessible.

TID level	Procedure	Success rate
0 krad(Si)	Initial read after irradiation	100%
	Re-read for verification	100%
	Re-program and then read	100%

	Erase followed by program and then read	100%
50 krad(Si)	Initial read after irradiation	100%
	Re-read for verification	100%
	Re-program and then read	100%
	Erase followed by program and then read	100%
100 krad(Si)	Initial read after irradiation	0%
	Re-read for verification	0%
	Re-program and then read	0%
	Erase followed by program and then read	0%

Table 1. *Irradiation testing results at various TID levels (10-keV X-ray exposure).*

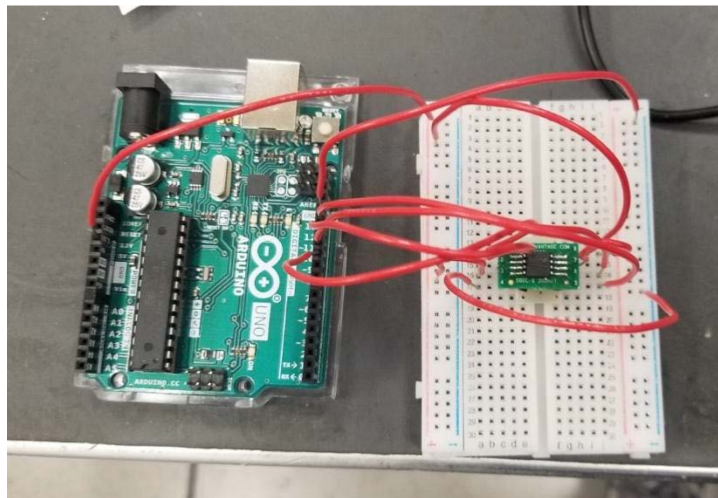


Fig. 26. *Photograph of the measurement setup showing Arduino Uno on the left and the MB85AS4MT on the right.*

vii. Develop CTT-hardware-based Inference Realistic Circuit Universal Simulator (CIRCUS)

The CIRCUS simulator allows us to provide realistic system performance information including both circuit & device non-idealities. In addition, it allows us to evaluate the accuracy of a network across several chips using a Monte Carlo simulation approach. CIRCUS is in the process of being developed by includes all crucial circuit non-idealities including parasitics, process corners, temperature variation, etc. as well as random CTT variation. It also allows us to study the effect of CTT device relaxation and evaluate how frequently a system may require a lightweight refresh to reinforce the target network weights.

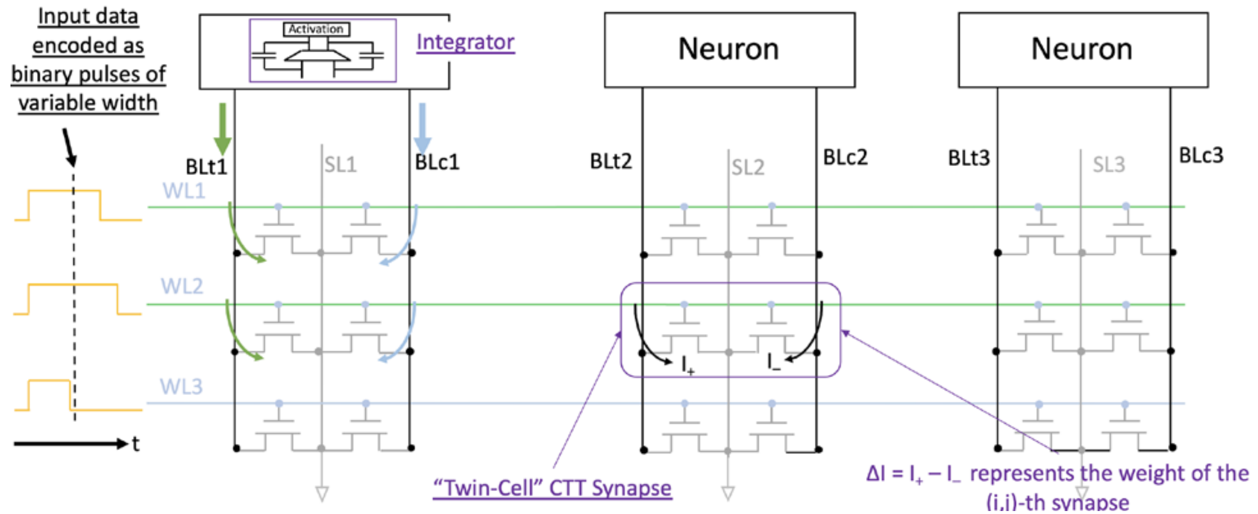


Fig. 27. *Proposed Architecture for CTT Inference Engine. Inputs to the network are applied as constant-amplitude pulse-width modulated (PWM) signals to the WLs. Weights are differential weights using two pre-programmed CTTs (programmed to specific states prior to performing inference, determined during training phase). Each neuron output corresponds to a column of devices. The neuron output is itself a constant-amplitude PWM signal that is proportional to the integrated differential current (*$\int (I_{BLt} - I_{BLc})dt$*).*

Figure *28* shows the simulated BLt and BLc currents for a random set of 256 inputs for a particular column or neuron output. The neuron output would then be a constant-amplitude pulse-width modulated (PWM) output that is proportional to the integrated differential current.

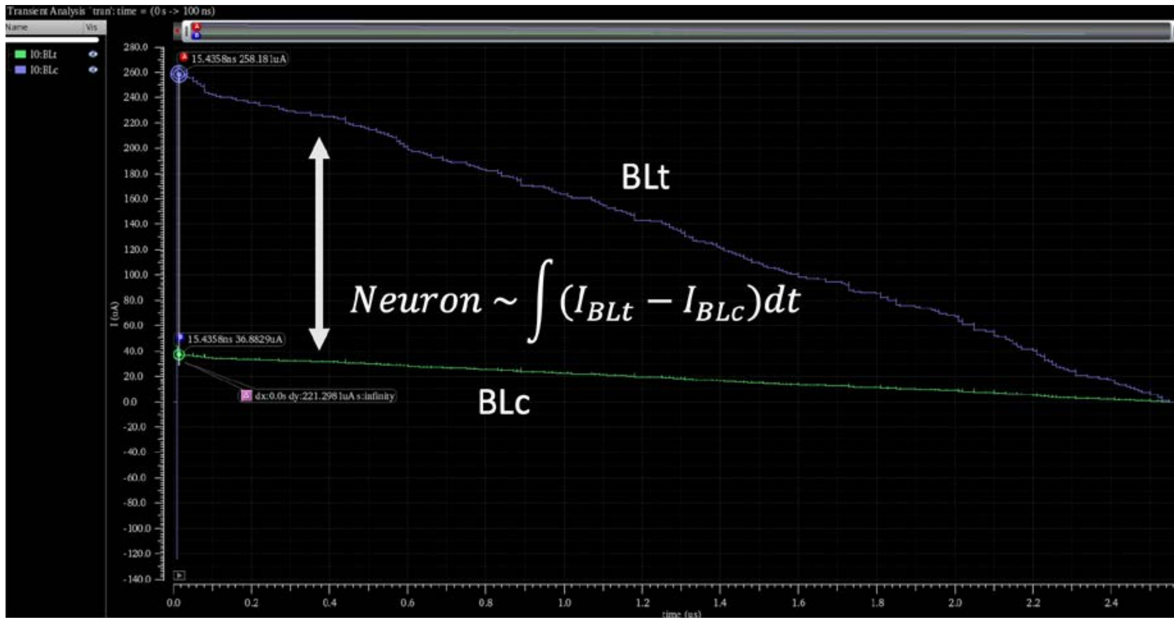


Fig. 28. *Single Neuron Output given 256 inputs. The neuron output corresponds to the integrated differential current ($\int (I_{BLt} - I_{BLc})dt$). Both I_{BLt} and I_{BLc} are shown as monotonically decreasing current waveforms given that all 256 pulse-width modulated inputs begin at t=0 and end depending on their respective input values. Input values in this example are assigned randomly. The neuron output is itself a constant amplitude PWM signal that is proportional to the integrated differential current.*

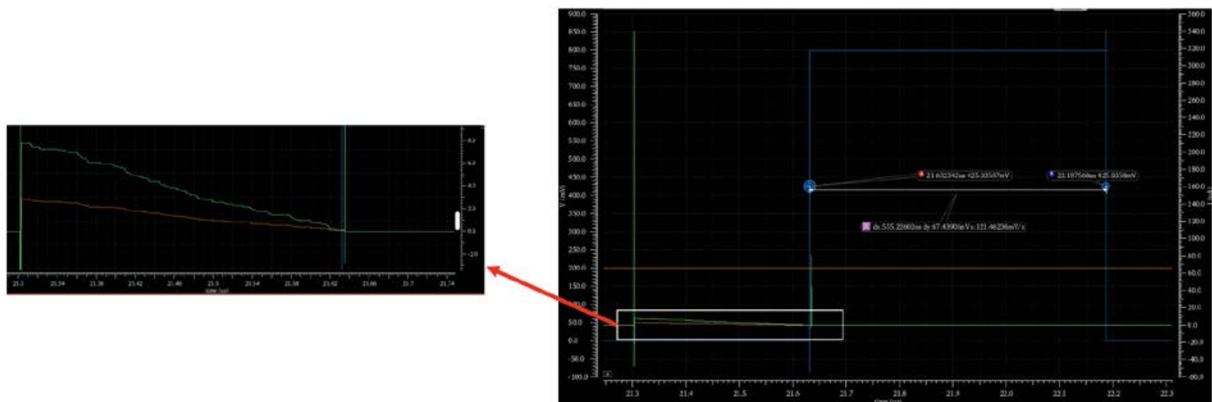


Fig. 29. *CIRCUS Simulator neuron output for a randomly initialized CTT array with random inputs. The zoomed in view on the left shows the I_{BLt} and I_{BLc} curves for a given input sample to the network and the graph on the right shows the pulse-width modulated (PWM) output of the neuron in blue which is directly proportional to the integrated differential current ($\sim \int (I_{BLt} - I_{BLc})dt$). In this example, Tcycle = 1.333ns and the neuron output is ~555ns.*

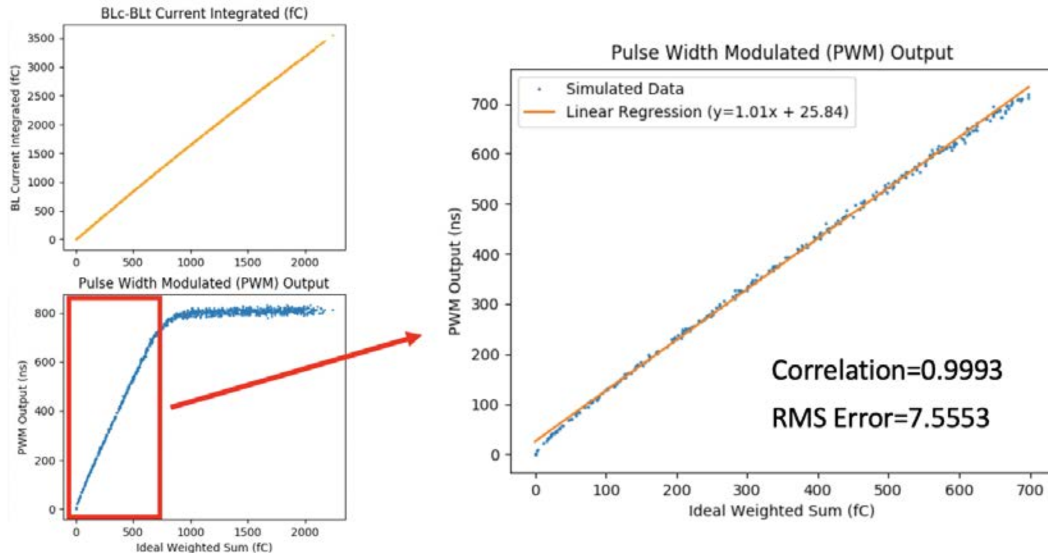


Fig. 30. *The accuracy of the simulator was evaluated by applying 1000 different input vectors (samples) to network and evaluating the neuron output's accuracy. The image on the top left shows the 100 different inputs applied showing the measured integrated differential BL current. The bottom left plot shows the neuron's PWM output for those 1000 different input samples. A linear relationship between the ideal weighted sum and the actual neuron PWM output is shown for a subset of the input space while the neuron does saturate at some point above a certain threshold. The right-hand plot zooms in on the linear region and displays the correlation and RMS error statistics between the ideal weighted sum and neuron PWM output.*

Figure *30* shows a clear linear relationship between the ideal weighted sum for random inputs and the neuron's pulse-width modulated (PWM) output duration. The neuron was designed to support a subset of the input space required for all desired applications. The neuron could be re-designed to support a larger input space before it saturates by increasing the neuron's capacitor amongst several other tradeoffs which we will not discuss here.

4) Key Outcomes

Throughout the course of this project, the objective has been to determine the robustness and resiliency of brain-inspired computing platforms relying on non-von Neumann architectures. In the past year, we have managed to study the effects of TID irradiation on both FDSOI and bulk CTT devices as well as RRAM devices. We have observed substantial effects in FDSOI due to hole trapping in the buried oxide (BOX) layer which is strongly coupled to the device channel. In contrast, we have observed less pronounced TID effects in bulk technologies such as GlobalFoundries 14LP where the dominate TID effect is trapping within the gate dielectric itself.

We have utilized various TCAD software including Ginestra to model the effects of irradiation on both CTT & RRAM devices which support all of our conclusions obtained from experimental data. We have observed that TID effects on the CTT device are more so technology-node dependent and not necessarily intrinsic to the CTT device itself. RRAM devices, while vary resilient to irradiation, require CMOS logic access circuitry which limits the overall performance of RRAM memory technologies under irradiation.

We have concluded our work by studying the effects of irradiation and other device non-idealities including device relaxation & subthreshold degradation on analog network inference accuracy utilizing the CTT. Our final effort has been to develop a CTT-hardware-based Inference Realistic Circuit Universal Simulator (CIRCUS) which allows us to consider the effects of both device non-idealities as well as circuit non-idealities when evaluating the performance of the CTT-based analog network. By providing a circuit-level simulation, we can consider all peripheral circuitry, device layouts, and parasitics to gather a more accurate understanding regarding how each of these device errors including V_{th} shifts due to irradiation may degrade network performance and accuracy.

Student Manuscript published in IEEE Transactions on Nuclear Science (TNS):

- "Total Ionizing Dose Responses of 22 nm FDSOI and 14 nm Bulk FinFET Charge-Trap Transistors", Rachel Brewer (Mar. 2021, *submitted*)

References:

[1] X. Guo, *et al.*, "Fast, Energy-Efficient, Robust, and Reproducible Mixed-Signal Neuromorphic Classifier Based on Embedded NOR Flash Memory Technology," IEDM 2017.
[2] G. Burr, *et al.*, "Recent Progress in Phase-Change Memory Technology," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 2016.
[3] X. Zheng, et al., "Error-Resilient Analog Image Storage and Compression with Analog-Valued RRAM Arrays: An Adaptive Joint Source-Channel Coding Approach," IEDM 2018.
[4] Francesco Maria Puglisi, et al., "Bipolar Resistive RAM Based on HfO_2: Physics, Compact Modeling, and Variability Control," IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 6, no. 2, pp. 171-184, Apr. 2016.
[5] Fujitsu Semiconductor, "Memory ReRAM: 4M (512K × 8) Bit SPI MB85AS4MT," MB85AS4MT datasheet, Dec. 2016.
[6] Arshiya Anjum et al., "Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms," vol. 379, pp. 265-271, 15 July 2016.
[7] F. Khan, E. Cartier, J. C. S. Woo and S. S. Iyer, "Charge trap transistor (CTT): An embedded fully logic- compatible multiple-time programmable non-volatile memory element for high-k-metal-gate CMOS technologies," IEEE Electron Device Letters, vol. 38, no. 1, pp. 44-47, Jan. 2017.
[8] A. Kerber, S. A. Krishnan and E. A. Cartier, "Voltage ramp stress for bias-temperature instability testing of metal-gate/high-k stacks," IEEE Electron Device Letters, vol. 30, no. 12, pp. 1347-1349, Dec. 2009.
[9] F. Khan, E. Cartier, C. Kothandaraman, J. C. Scott, J. C. S. Woo and S. S. Iyer, "The impact of self-heating on charge trapping in high-k-metal-gate nFETs," IEEE Electron Device Letters, vol. 37, no. 1, pp. 88-91, Jan. 2016.
[10] P. J. McWhorter, S. L. Miller, and T. A. Dellin, "Radiation response of SNOS nonvolatile memory transistors," IEEE Trans. Nucl. Sci., vol. 33, no. 6, pp. 1414-1419, Dec. 1986.
[11] E. S. Snyder, P. J. McWhorter, T. A. Dellin, and J. Sweetman, "Radiation response of floating gate EEPROM memory cells," IEEE Trans. Nucl. Sci., vol. 36, no. 6, pp. 2131–2139, Dec. 1989.
[12] S. Gerardin, M. Bagatin, A. Paccagnella, K. Grürmann, F. Gliem, T. R. Oldham, F. Irom, and D. N. Nguyen, "Radiation effects in Flash memories," IEEE Trans. Nucl. Sci., vol. 60, no. 3, pp. 1953-1969, Jun. 2013.
[13] D. M. Fleetwood, "Radiation-induced charge neutralization and interface-trap buildup in metal-oxide-semiconductor devices," J. Appl. Phys., vol. 67, no. 1, pp. 580-583, Jan. 1990.
[14] I. S. Esqueda et al. "Modeling the radiation response of fully-depleted SOI n-Channel MOSFETs," IEEE Trans. Nucl. Sci., vol. 56, no. 4, pp. 2247-2250, Aug. 2009.
[15] N. N. Mahatme et al., "Impact of back-gate bias and device geometry on the total ionizing dose response of 1-transistor floating body RAMs," IEEE Trans. Nucl. Sci., vol. 59, no. 6, pp. 2966-2973, Dec. 2012.
[16] E. Simoen et al., "Radiation effects in advanced multiple gate and silicon-on-insulator transistors," IEEE Trans. Nucl. Sci., vol. 60, no. 3, pp. 1970-1991, Dec. 2013.
[17] R. Carteret al., "22nm FDSOI technology for emerging mobile, Internet-of-Things, and RF applications," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 2.2.1-2.2.4.
[18] N. Rezzak et al., "Total-ionizing-dose radiation response of 32 nm partially and 45 nm fully-depleted SOI devices," 2012 IEEE International SOI Conference (SOI), NAPA, CA, 2012.
[19] Kothandaraman, *et al.*, "Oxygen vacancy traps in Hi-K/Metal gate technologies and their potential for embedded memory applications," 2015 IEEE International Reliability Physics Symposium, April 2015.
[20] Faraz Khan, E. Cartier, C. Kothandaraman, J. C. Scott, J. C. S. Woo, and S. S. Iyer, "The Impact of Self-Heating on Charge Trapping in High- k -Metal-Gate nFETs," Electron Device Letters, IEEE, vol. 37, no. 1, pp. 88-91, Jan. 2016.

What do you plan to do during the next reporting period to accomplish the goals?

If we were to be granted the 4th year extension, we plan to make our CIRCUS model more robust with extensive hardware validation with our third-generation chip. We will realize inference engines to address real life problems and validate our models with experimental results after radiation exposure. This would be the first radiation studies on analog neuromorphic engines to the best of our knowledge and these results would determine the limits on the use of analog in-memory computing in radiation rich environments.