SAND2021-2579J

Ionizing radiation effects in SONOS-based neuromorphic inference accelerators

T. Patrick Xiao, *Member, IEEE*, Christopher H. Bennett, *Member, IEEE*, Sapan Agarwal, *Member, IEEE,* David R. Hughart, *Member, IEEE*, Hugh J. Barnaby, *Fellow, IEEE*, Helmut Puchner, Venkatraman Prabhakar, A. Alec Talin and Matthew J. Marinella, *Senior Member, IEEE*

Abstract— **We evaluate the sensitivity of neuromorphic inference accelerators based on Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) charge trap memory arrays to total ionizing dose (TID) effects. Data retention statistics were collected for 16 Mbit of 40 nm SONOS digital memory exposed to ionizing radiation from a Co-60 source, showing good retention of the bits up to the maximum dose of 500 krad(Si). Using this data, we formulate a rate-equation-based model for the TID response of trapped charge carriers in the ONO stack, and predict the effect of TID on intermediate device states between 'program' and 'erase'. This model is then used to simulate arrays of low-power, analog SONOS devices that store 8-bit neural network weights and support *in situ* matrix-vector multiplication. We evaluate the accuracy of the irradiated SONOS-based inference accelerator on two image recognition tasks – CIFAR-10 and the challenging ImageNet dataset – using state-of-the-art convolutional neural networks, such as ResNet-50. We find that across the datasets and neural networks evaluated, the accelerator tolerates a maximum TID between 10 krad(Si) and 100 krad(Si), with deeper networks being more susceptible to accuracy losses due to TID.**

Index Terms— **Charge trap memory, SONOS, total ionizing dose, ionizing radiation, neuromorphic computing, neural networks, inference accelerators.**

I. Introduction

NON-VOLATILE memory arrays that compute large matrix operations in the analog domain have emerged as strong candidates for high-throughput and energy-efficient hardware accelerators of deep neural networks. Synaptic weights within a neural network layer can be stored in the conductance states of the memory elements, as shown in Fig. 1(a). By driving the rows with a vector of input activations, the full matrix-vector multiplication (MVM) can be performed locally within the array – eliminating the data transfer bottleneck – with all constituent multiply-accumulate operations (MACs) executed in parallel [1, 2].

Charge trap memory is an attractive option for synaptic weights owing to its CMOS compatibility and multi-level storage capability. In SONOS (silicon-oxide-nitride-oxide-silicon) memory, which is in commercial production for data storage, a non-volatile state is represented by the quantity of electrons residing in a charge-trapping nitride layer between the channel and gate of a MOSFET structure, as shown in Fig. 1(b). SONOS devices have recently been engineered for inference applications, with low read noise and programmability to many current levels (up to 8 bits) [3]. The accuracy of SONOS-based

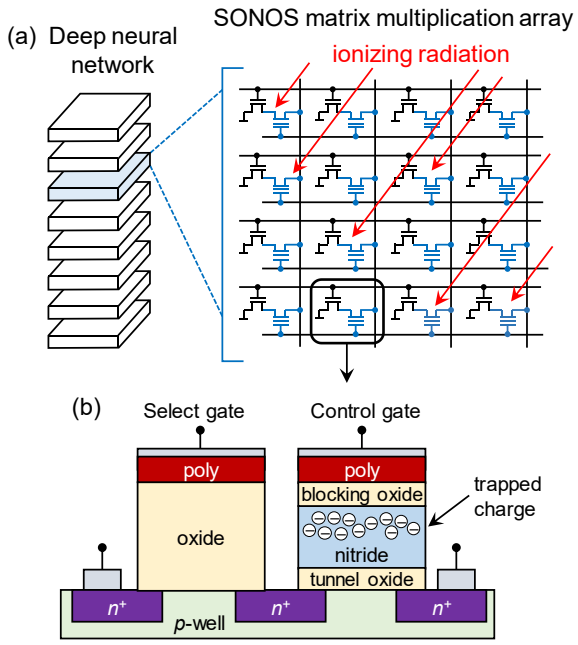

Fig. 1. (a) A matrix of weights belonging to one layer of a deep neural network can be stored in a SONOS memory array. *In situ* matrix-vector multiplication can be performed within this array by driving the rows with an input vector and summing the output currents on the columns. (b) Memory cell consisting of a SONOS transistor, which stores information in the quantity of trapped charge, and an access transistor. A signed weight can be implemented using the difference in current between two SONOS cells. When exposed to ionizing radiation, the state of the SONOS cells may be perturbed.

The work at Sandia National Laboratories was supported by the Laboratory-Directed Research and Development (LDRD) Programs. This work was funded in part by DTRA under grant no. HDTRA1-17-1-0038. The authors would like to thank Jacob Calkins of DTRA for his support of this work. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

T. P. Xiao, C. H. Bennett, S. Agarwal, D. Hughart, A. A. Talin and M. Marinella are with Sandia National Laboratories, Albuquerque, NM 87185 USA (e-mail: txiao@sandia.gov, mmarine@sandia.gov).

H. Barnaby is with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85287 USA.

H. Puchner and V. Prabhakar are with Infineon Technologies, San Jose, CA 95037 USA.

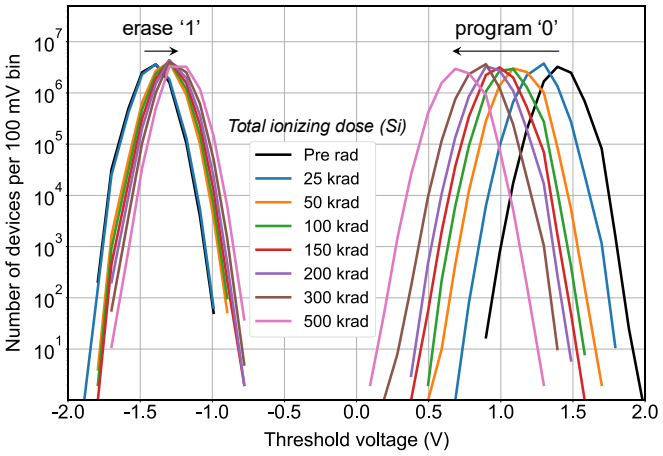

Fig. 2. Measured V_T distribution of the SONOS test chips at varying levels of TID. The V_T levels are grouped into 100 mV bins; the number of devices in each bin is shown.

neuromorphic accelerators, accounting for device properties, has been explored by previous work [4, 5]. Several accelerator architectures have been also demonstrated using charge trap memory and the closely related floating-gate memory [6-8].

In high-radiation environments, such as space missions, non-volatile memories are subject to cumulative damage induced by ionizing radiation. Total ionizing dose (TID) effects on neural network accuracy has been previously studied in neuromorphic accelerators based on resistive random-access memory (ReRAM), both for inference [9] and for training [10]. In flash and charge trap memories, TID can lead to the addition or removal of stored charge, as described theoretically [11, 12] and characterized experimentally [13-16]. We measure the state distribution across 16 Mbit of 40nm SONOS digital memory devices when exposed to varying levels of TID and combine this with a device physics model to derive the irradiated response of analog SONOS weights. We use this model to project the resilience of SONOS-based inference accelerators, running state-of-the-art convolutional neural networks (CNNs) pre-trained on image recognition tasks, when operated under high levels of ionizing radiation.

II. TID Experiments on SONOS Digital Memory

In a SONOS memory device, stored electrons within the nitride charge trap layer modify the threshold voltage (V_T) needed to turn on the channel. When operated as digital memory, electrons can be injected into the nitride layer to set the device to the 'program' state with high V_T (low current), or the nitride can be depleted of electrons to reach the 'erase' state with low V_T (high current). To function reliably as digital memory, the V_T distributions of the 'program' and 'erase' states across a large array of SONOS devices must have a sufficiently large separation to minimize bit errors.

TID experiments were conducted on memory arrays fabricated by Infineon Technologies in a 40nm SONOS technology. Test chips containing 8 Mbit of un-cycled SONOS memory were used for these tests. Of these, 4 Mbit was used to measure the effect of the radiation on the device V_T states; 2 Mbit was set to the 'program' state and 2 Mbit was set to the 'erase' state immediately prior to irradiation. The remaining 4 Mbit of the chip was used to verify the ability to reprogram the cells after irradiation. The chips were mounted on a test board and exposed to ionizing radiation from a Co-60 source in a sequence of dose steps at room temperature. While exposed, power was supplied to the test chip, including the peripheral circuits, and all memory cells were left in standby mode; the cells were not read, programmed, or erased during the tests and therefore not under any applied bias. The V_T distribution across all 8 Mbit of memory per chip was profiled after each step, at TID levels of 25k, 50k, 100k, 150k, 200k, 300k, and 500k rad(Si). A constant dose rate of 50 rad(Si)/s was used during exposure. The full duration of the test, including V_T characterization after each radiation step, was approximately three hours.

Fig. 2 shows the aggregated V_T distribution of the irradiated devices across four test chips, containing a total of 16 Mbits, at varying TID levels. Only the cells that were not reprogrammed are shown. Devices in the 'program' state show a decrease in V_T with TID exposure on average, while the 'erase' state V_T increases. The 'program' V_T shows a more pronounced response to radiation. These trends are consistent with previous reports in the literature for SONOS and flash memory [13, 14, 17]. Up to the maximum TID of 500 krad(Si), the test chips retain a large enough separation in V_T between the 'program' and 'erase' distributions to reliably support the full program/erase window of the digital memory.

The remaining 4 Mbit of SONOS cells on each exposed test chip, separate from the cells characterized in Fig. 2, were erased and programmed after each radiation step. The V_T distribution of these reprogrammed cells, which continued to support the full program/erase window without any bit errors, showed minimal damage to the peripheral circuits up to 500 krad(Si).

III. Modeling Radiation-Induced Decay in SONOS Synapses

While the tested SONOS devices retain their digital state under a total dose of at least 500 krad(Si), TID will affect the devices more strongly when they are operated as multi-level synaptic elements for analog *in situ* MVM operations. We consider an inference accelerator where neural network weights are stored at 8-bit precision, which is a commonly used resolution for neural network inference [18, 19], using the difference in current of two SONOS devices to represent positive and negative weights. For this use case, each SONOS device needs to be programmable to 128 distinct current levels, and each level must have good retention to maintain sufficient data precision in the neural network. The accuracy of the analog accelerator is thus more sensitive to TID-induced V_T shifts than a binary digital memory array.

To predict the TID response of intermediate states of the SONOS device between 'program' and 'erase', we use a rate-equation model for the quantity of trapped electrons and holes within the ONO (oxide-nitride-oxide) gate stack. We fit this model to the measured TID response of the 'program' and 'erase' states, which represent the two extremes of a continuum of analog device states. Since we measured the V_T distribution

at several intermediate radiation levels below 500 krad(Si) (see Fig. 2), the experimental data also includes the response of several intermediate V_T states to further TID exposure.

In a SONOS gate stack, charge is stored in traps within the bulk of the nitride layer, as well as in traps at the interface of the nitride and the oxide. We assume that the threshold voltage is modified by the trapped charge as:

$$V_T = V_{T,neu} + q(n_T - p_T)\left(\frac{d_{BO}}{\epsilon_{ox}} + \frac{d_n}{2\epsilon_n}\right) + qn_{TO}\left(\frac{d_{BO}}{\epsilon_{ox}} + \frac{d_n}{\epsilon_n}\right) + qn_{BO}\frac{d_{BO}}{\epsilon_{ox}}, \quad (1)$$

where $V_{T,neu}$ is the value of the threshold voltage when the ONO stack is neutral, i.e. contains net zero trapped charge. n_T and p_T are the densities (per area) of electrons and holes stored in bulk nitride traps, and n_{TO} and n_{BO} are densities of electrons trapped at the silicon nitride interface with the tunneling oxide (TO) and blocking oxide (BO), respectively. We do not include interface hole traps in the present model as these were not needed to describe the available measured data. d_n and d_{BO} are the thicknesses of the nitride and BO, respectively, ε_n and ε_{ox} are the dielectric constants of Si_3N_4 and SiO_2, and q is the electron charge. Equation (1) assumes that the bulk nitride traps are distributed uniformly through the thickness of the nitride storage layer [20].

The three types of traps considered (bulk electron traps, bulk hole traps, and interface electron traps) each respond differently to ionizing radiation owing to their position, energy distribution, and total density of traps. Since the density of traps is not directly known for our devices, we treat these as fitting parameters of the model. We do not model the energy distributions of the traps directly, since these are also not known *a priori*, and instead capture their effect using a different emission coefficient e_c for each group of traps in Equation (1). Shallow traps close to the conduction or valence bands are more easily ionized by radiation than traps that lie deep inside the bandgap, and traps at the interface can more readily be evicted from the gate stack than traps in the interior of the nitride.

We use a separate rate equation to describe how each of the four trapped charge densities (n_T, p_T, n_{TO}, and n_{BO}) changes with an increment of ionizing dose γ (rad). The processes that add or remove charge from these traps are depicted in Fig. 3(a) and (b) for the program and erase state, respectively. We describe these mechanisms below.

A. Oxide charge generation and injection into nitride

Ionizing radiation generates electrons and holes in the oxide layers. After generation, some of these electron-hole pairs very quickly recombine before they drift in opposite directions under the influence of an electric field. The fraction of carriers that remain after recombination is the charge yield. The depedence of the charge yield on the electric field and the energy and type of incident radiation has been extensively studied for SiO_2 [21]. We use the germinate recombination model [22], which accurately describes the carrier dynamics under exposure Co-60 gamma rays, to compute the charge yield in the TO and BO. Since the TO tends to be significantly thinner, it has a lower carrier generation rate per area but a higher charge yield due to the larger field.

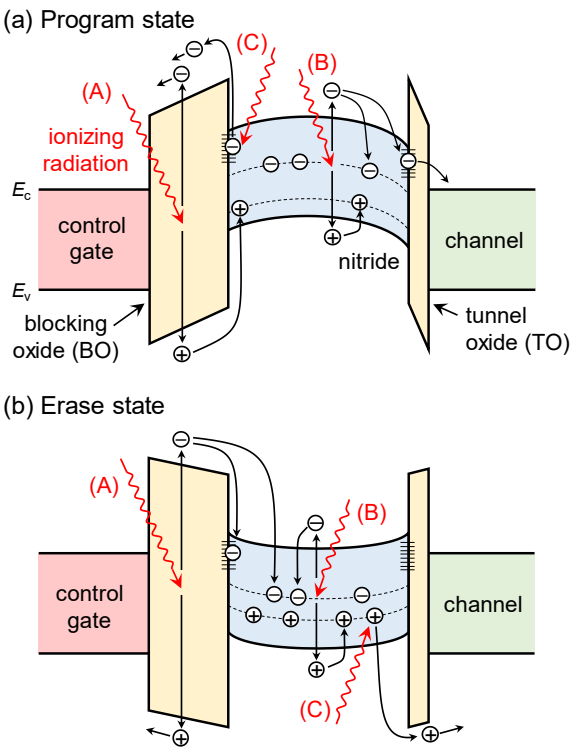

Fig. 3. Modeled processes of trapped charge injection and removal when the SONOS device is exposed to ionizing radiation in (a) the program state and (b) the erase state. The effects of ionizing radiation labeled (A)–(C) are described under the corresponding subsection headings in Section III.

When there is net negative charge in the ONO stack, as in Fig. 3(a), the oxide fields drive the holes toward the nitride while electrons are swept out to the electrodes. These holes can be captured by unfilled traps in the nitride, decreasing V_T so that the device approaches the neutral state $V_{T,neu}$. Holes that do not find an available trap recombine after some time in the nitride layer and have no effect on V_T. Trap saturation is modeled by a trapping probability $1 - p_T/P_T$, where P_T is the density of available hole traps. We do not include the trapping of carriers within the oxide, away from the interface, as this effect tends to be minor due to the small oxide thickness in scaled SONOS technology [13].

When there is net positive charge in the ONO stack, as in Fig 3(b), the generated electrons migrate into the nitride layer. These electrons can populate either the traps at the nearest interface (TO or BO) or the traps in the bulk nitride, increasing V_T and again driving the device toward the neutral state $V_{T,neu}$. Injected electrons that are not trapped eventually recombine and have no effect on V_T.

B. Charge generation in the nitride

Ionizing radiation can also generate electrons and holes in the nitride. Unlike SiO_2, the process of electron-hole generation and subsequent recombination resulting from Co-60 irradiation has not been well studied in silicon nitride. To model it, we first scale the number of electron-hole pairs generated per dose in SiO_2 by the density and bandgap of Si_3N_4, and use the same

charge yield model in Si_3N_4 as in SiO_2. The charge yield is calculated as a function of depth in the Si_3N_4 layer, where the electric field varies linearly following the assumption of uniform charge distribution in the nitride layer. We then scale this effective charge generation rate at each position by a fixed value α, which is left as a fitting parameter to the measured data in Fig. 2. This parameter reflects our uncertainty in the charge generation and separation processes as a result of gamma ray irradiation in the nitride layer compared to SiO_2 stemming, among other factors, from different carrier mobilities and charge screening properties.

For the situation in Fig. 3(a), electrons that survive the initial recombination drift toward either the TO or BO, depending on their initial position, and will either fill these interface traps, the bulk electron traps, or recombine. Trapped electrons at the interface can escape out of the ONO stack via tunneling, whose probability is calculated using a modified Fowler-Nordheim expression [23]. The tunneling probability is much higher through the thinner TO than the BO, though in general electron tunneling is a small effect due to the relatively small oxide field (compared to that seen during the program/erase operation). If the net charge in the ONO stack is positive, as in Fig. 3(b), the electrons must be trapped or recombine in the interior of the nitride. Holes that are generated in the nitride are likewise trapped or recombine, though there is less dependence on the polarity of the field in this case since hole interface traps are not included, and hole tunneling is even smaller than that of electrons due to the considerably larger offset in the valence band.

C. Radiation-assisted trap emission

Trapped charge in the nitride layer and at the interfaces can be emitted out of the ONO stack upon receiving energy from ionizing radiation. Following [11], we model this emission rate to be proportional to the trapped charge density with an emission coefficient e_c. As described previously, different emission rates are chosen for the bulk nitride electron traps, bulk hole traps, and interface electron traps which have different energy and spatial distributions. We expect the interface traps to be shallower on average than the bulk traps, which can lie in a defect band that is deep inside the bandgap, particularly for electrons [24].

D. Fit to measured data

We choose the trap densities and emission coefficients for each group of traps (bulk electron traps, bulk hole traps, and interface electron traps), the neutral-point voltage $V_{T,neu}$, and the scale factor α for the nitride charge generation to fit the measured data in Fig. 2. These parameters are listed in Table I. We also use additional parameters to convert a given initial threshold voltage to an initial distribution of charge among the different traps in Equation (1). The resulting fit to the data is shown in Fig. 4(a), where the blue and red points denote the mean values of the measured V_T distribution for the program and erase states, respectively, at each TID level in Fig. 2. The solid curves are the V_T evolutions with TID predicted for several intermediate values of initial V_T. The model gives an accurate fit to the measured data using a neutral-point voltage of $V_{T,neu} = -0.907$V, which is closer to the erase state than to the program state.

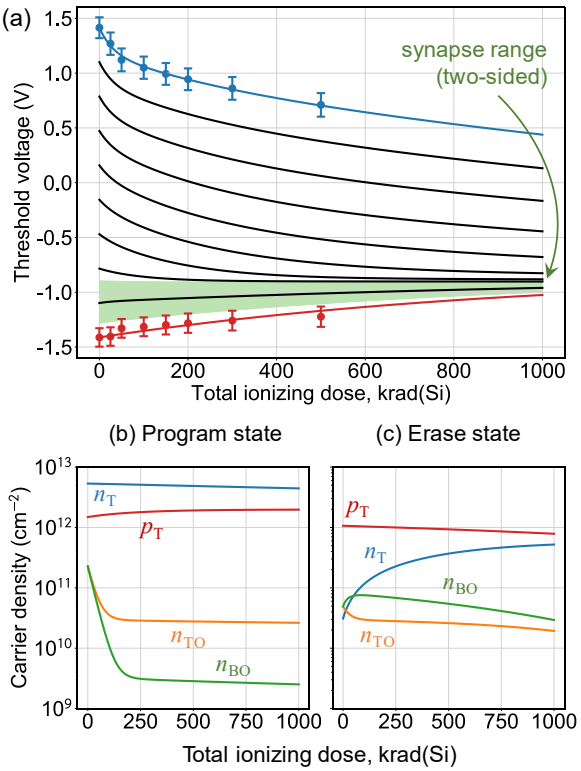

Fig. 4. (a) The rate equation based model of V_T as a function of TID is plotted alongside the means of the V_T distributions in Fig. 2 for the 'program' and 'erase' states. The data points and error bars denote the mean and standard deviations of the V_T distributions in Fig. 2 at different TID levels. The green region denotes the range used to represent synaptic weights with a two-sided mapping scheme as in Fig. 5(a). (b) The evolution of the trapped carrier densities with increasing TID for the program state. (c) The evolution of the trapped carrier densities with increasing TID for the erase state.

TABLE I
TID MODEL FITTING PARAMETERS

Parameter	Value
Neutral point threshold voltage $V_{T,neu}$	–0.907V
Nitride charge generation scaling factor, α	0.150
Trap densities:	
Bulk electron traps	1.0×10^{13} cm^{-2}
Bulk hole traps	2.4×10^{12} cm^{-2}
Interface electron traps	9.1×10^{11} cm^{-2}
Emission coefficients:	
Bulk electron traps	1.9×10^{-7} rad^{-1}
Bulk hole traps	5.8×10^{-7} rad^{-1}
Interface electron traps	3.3×10^{-5} rad^{-1}

Fig. 4(b) and (c) show the evolution of trapped carrier densities corresponding to the predicted V_T evolution of the program and erase states, respectively, in Fig. 4(a). In the program state, Fig. 4(b), the sharp decrease in V_T that is observed at low TID is explained by a rapid de-filling of the interface traps, which are shallow in energy and readily ionized

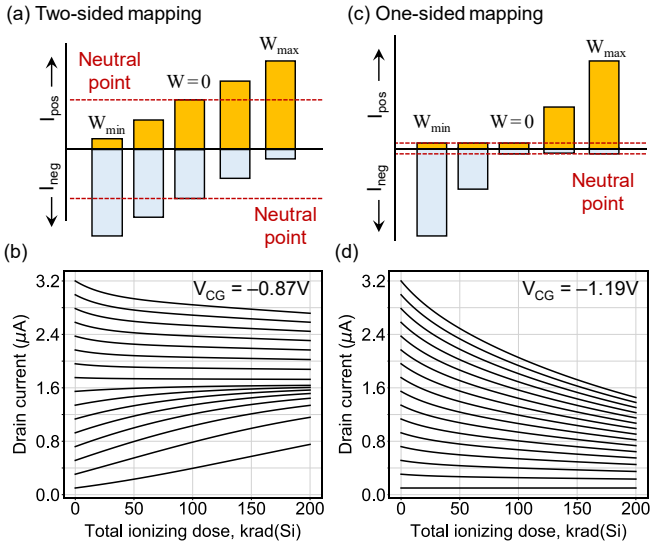

Fig. 5. Two schemes of mapping V_T values to currents and neural network weights. The weight represented by a pair of SONOS devices is proportional to $I_{pos} - I_{neg}$. (a) In the two-sided mapping scheme, a zero weight is mapped to two devices with currents set to the midpoint of the range from 0.01 μA to 3.2 μA. (b) Decay in drain current with TID when the neutral point is set to the midpoint current. (c) In the one-sided mapping scheme, a zero weight is mapped to two devices programmed to the lowest current level. (d) Decay in drain current with TID when the neutral point is set to the lowest current.

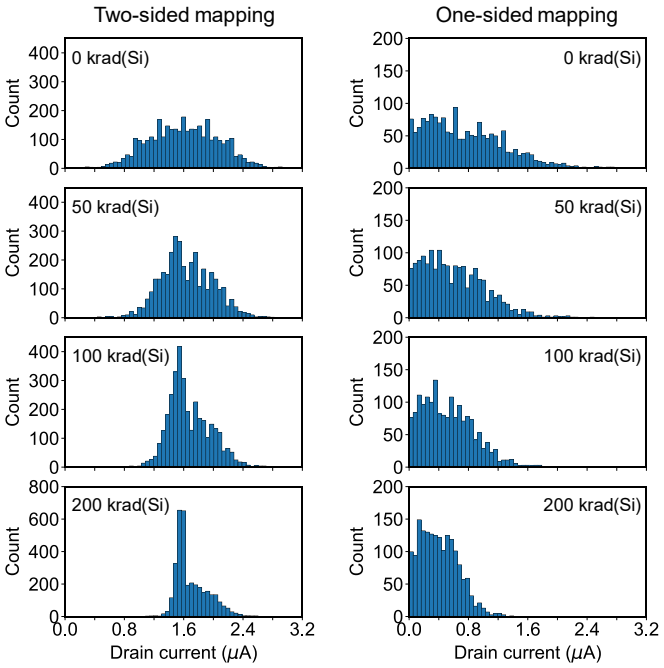

Fig. 6. Distribution of device currents in the first layer of the CIFAR-10 CNN in Fig. 7 at several levels of TID. For the two-sided mapping scheme, zero weights are represented using the midpoint current (1.6 μA) while for the one-sided mapping scheme zero weights correspond to zero current. For the one-sided case, the current of the "unused" device in the pair, which is always set to the minimum value, is not included in the distribution. With increasing TID, the weights are compressed towards zero in both cases.

by radiation. This de-filling is eventually balanced by electrons that are generated in the nitride and trapped at the interfaces, reaching a steady state. Afterwards, V_T declines more gradually toward the neutral point by the slow de-filling of the deeper electron traps in the bulk. The trapped hole population remains relatively stable due to saturation of the available traps.

In the erase state, Fig. 4(c), the net charge in the ONO stack is initially positive due to the trapped holes and low population of trapped electrons. Upon exposure to ionization radiation, the trapped electron and hole populations both increase due to carrier generation in the nitride and electron injection from the oxides. This increases V_T, but as the device approaches the neutral point where the bands are flat, these processes are slowed or reversed as a result of the decreasing charge yield and the removal of trapped charge by radiation. The smaller response in the erase state V_T is attributed to its relative proximity to the neutral point $V_{T,neu}$ of the device.

We note that while the set of fitting parameters in Table I yields a strong fit to the data, it is not guaranteed to be a unique solution. The values of these parameters, and the model itself, can be clarified with additional irradiation data for intermediate device states, as well as by independent experiments on unexposed devices to ascertain the values of $V_{T,neu}$ and the trap densities.

E. Mapping of device state to neural network weights

Since the matrix-vector product is given by a sum of currents along a column (or bit line) of the array, the values of the neural network weights are proportional to the currents passed by the memory cells. To operate the cells as energy-efficient multi-level synapses, we use the SONOS transistors in the subthreshold regime. We allocate 128 current levels in the range from 0.01 μA to 3.2 μA to represent the weights, which is similar to the range previously used in Ref. [3] for low-current multi-bit SONOS devices optimized for neural inference. We use measured I_D-V_{CG} data on devices programmed to different threshold voltages as the basis of the conversion from V_T to I_D. The drain-source voltage is fixed at 0.1V.

The control gate voltage V_{CG} can be used as a degree of freedom to set the range of V_T values that will be mapped to a desired range of drain currents I_D. This bias point should be chosen in order to minimize the disturbance to I_D, and therefore the neural network weights, caused by ionization radiation. The response to TID is weakest near the neutral point of the device, where the charge yield is near zero (due to the absence of an electric field) and the trapped carrier densities are also near zero. Thus, V_T values near $V_{T,neu}$ should be mapped to the weights; in particular, since synaptic weights tend to be distributed around zero, the zero weight should be mapped exactly to the neutral point.

We investigate two possible schemes of mapping V_T values that satisfy the above condition, shown in Fig. 5. In the "two-sided" scheme in Fig. 5(a), the neutral point $V_{T,neu}$ is set to correspond to the midpoint current of 1.6 μA. When using the difference in current of two memory cells to represent a signed weight, writing both devices to the midpoint current sets the weight to zero. Writing one device to a higher (lower) current than the midpoint and the other device to a lower (higher) current yields a progressively more positive (negative) weight value. This scheme has the advantage that neither the largest nor the smallest current used will decay very strongly with TID, as shown in Fig. 5(b), because they remain somewhat close to the

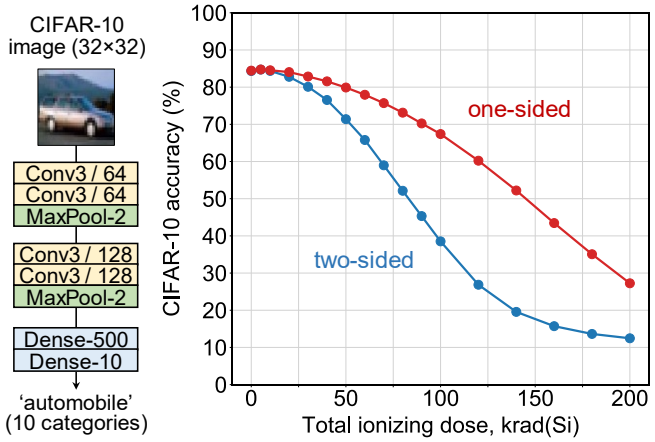

Fig. 7. TID-induced accuracy degradation of a plain CNN trained on the CIFAR-10 task. The network has 4.36M total weights and 100.4M multiply-accumulate operations (MACs). A ReLU activation with an upper bound of 1 is used between layers [4].

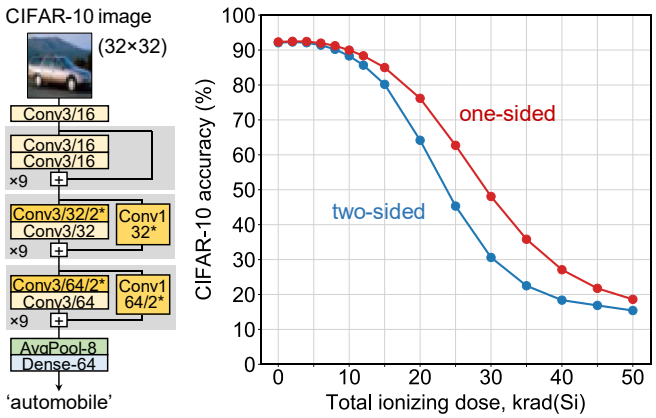

Fig. 8. TID-induced accuracy degradation of a deep residual network (ResNet-56v1 [26]) trained on the CIFAR-10 task. The network has 0.86M total weights and 126.3M MACs. On the left, asterisk refers to a shortcut layer or a stride of 2 that is present only in the first residual block of the group.

neutral point. Since the currents are subtracted, the TID-induced decay in the two devices reinforce each other. We note that the degradation in current is not symmetric about the neutral point; the states with $V_T > V_{T,neu}$ (I_D below the mid-point) are more sensitive to TID.

In the "one-sided" scheme in Fig. 5(c), the neutral point $V_{T,neu}$ is set to the lowest current level in the desired range. When mapping to weights, one of the two devices in the pair will always be set to the lowest current, which is at the neutral point and is therefore relatively immune to TID. However, the other device has a larger sensitivity to TID if the programmed current is high, as shown in Fig. 5(d). The one-sided scheme has the advantage of significantly lower energy consumption, since the majority of devices in the system will be set to the lowest current level. It may also be more robust to programming errors, since it has been found that SONOS devices with lower target currents can be programmed more precisely [3].

In both mapping schemes in Fig. 5, zero weights (which appear most frequently in a neural network) are unperturbed by TID while the weights at the extremes of the distribution will be compressed toward zero, leading to a loss of inference accuracy. This is shown in Fig. 6 for the first layer of the CNN to be shown in Fig. 7. As TID increases, the device currents decay toward the neutral point and the weight values are compressed toward zero.

IV. Neural Inference Accuracy Loss Under Radiation Exposure

In this section, we evaluate the image recognition accuracy of a SONOS-based neural inference engine operating in a high-radiation environment using CrossSim, a simulation tool for computational memory arrays [25]. Sensitivity to TID is evaluated on CNNs trained on two image classification tasks: CIFAR-10, which has 10 categories of objects, and the much more difficult ImageNet Large-Scale Visual Recognition Challenge dataset (ILSVRC-2012, or ImageNet for short), which has 1000 categories. For CIFAR-10, we evaluate two networks: a plain, strictly sequential CNN (Fig. 7) and a deep residual network, ResNet-56v1 (Fig. 8), that is trained using the procedure detailed in [26]. For ImageNet, we evaluate the pre-trained ResNet-50 network (Fig. 9) that is available in the Keras machine learning library [27]. Since ResNet-56v1 and ResNet-50 share similar topologies, we expect that our evaluations across these three networks will reveal the role of both task complexity and network topology on the sensitivity of the inference accelerator to TID. For CIFAR-10, we evaluate inference on the full dataset (10,000 images) while for ImageNet we use the same subset of 5,000 images (10% of validation set) in each inference simulation.

We follow the scheme proposed in Ref. [28] for mapping the weights of a convolutional layer onto a memory array. Batch normalization parameters, if present, are folded into the weights of the convolutional layers [29]. For each layer, the full range of weight values is mapped to the current range from 0.01 μA to 3.2 μA using one of the two schemes in Fig. 5. To isolate the effect of TID on the SONOS device synapses, our simulations do not include the effect of random programming errors or read noise, array parasitic resistances, the resolution of the analog-to-digital converter (ADC), or any radiation-induced damage to the peripheral circuits or access devices. We do, however, quantize the floating-point weights from the neural network model to 8-bit precision so that they are compatible with the programmable resolution of the SONOS devices. The effect of TID, based on the model in Section III, is then applied onto the programmed current values in each device. Bias weights are not included in the array and are instead stored digitally at higher precision to maintain high accuracy [29]. We assume that these digitally stored weights are unaffected by TID.

Fig. 7 shows the TID-induced loss in classification accuracy of a plain CNN trained on the CIFAR-10 task, which has a floating-point baseline accuracy of 84.5%. The accuracy falls off gradually with TID due to the induced decay in the weight values. For the one-sided mapping scheme, the accuracy falls by 10% relative to the floating-point baseline after a TID of 70 krad(Si), while for the two-sided scheme the same point is reached at 40 krad(Si). The inferior performance of the two-

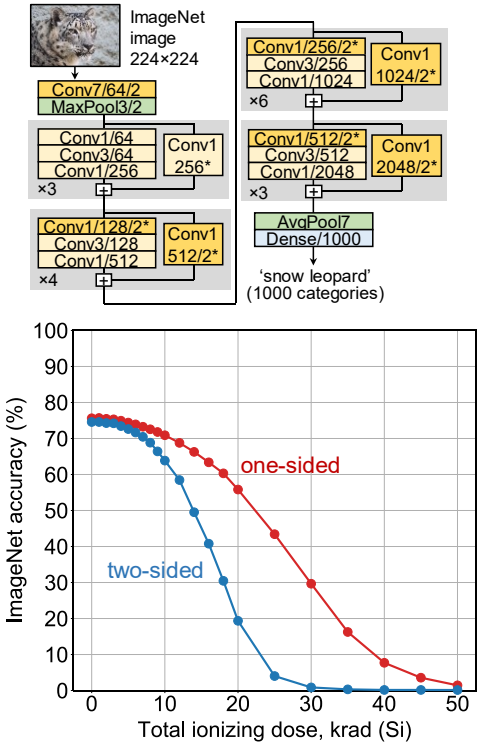

Fig. 9. TID-induced degradation in the top-1 accuracy on ImageNet using ResNet-50 [26]. The network has 25.6M total weights and 4.1B MACs.

sided scheme results from the fact that devices with $V_T > V_{T,neu}$, which correspond to the lower half of Fig. 5(b), are more sensitive to TID in our model. This can also be observed in Fig. 6 (left) where the smaller currents collapse more swiftly than the larger currents. The one-sided scheme performs better since it uses only devices with $V_T < V_{T,neu}$, which are more stable upon exposure to radiation, and only one of two devices in each memory cell pair is degraded by radiation. Only the large outlier weights, which map to states far from the neutral point, are prone to degrade more rapidly in the one-sided scheme than the two-sided scheme.

Fig. 8 shows the TID response of a CNN that is trained on the same task but has a more complex structure. The network, which implements the ResNet-56v1 topology in Ref. [26], is 56 layers deep when counting the convolutional and dense layers along the main path, but also employs frequent skip connections. The floating-point accuracy of this network is 92.3%, which is close to the state-of-the-art for CIFAR-10. The accuracy falls by 10% relative to floating-point at ~16 krad(Si) for the one-sided mapping scheme and ~13 krad(Si) for the two-sided scheme. Despite being evaluated on the same task, the residual network is significantly more sensitive to TID than the plain network. Since the weight matrices (i.e. filter sizes) are smaller in ResNet-56v1 than the plain CNN in Fig. 7, we believe that the added sensitivity can be largely attributed to the depth of the residual network, which is greater by about a factor of 9. Due to the many sequential matrix-vector multiplication operations encountered in a deep network, the effects of TID accumulate from layer to layer. Unlike the effects of random noise or programming errors, these TID-induced decays do not experience any cancellation during propagation. In fact, the decayed weights in one layer are propagated to the next layer as decayed activations, and the effects of TID may be multiplicative from one layer to the next. We note it is also possible that since the network in Fig. 7 has about 5× more weights, it gains some resilience simply from having more redundant fitting parameters.

Fig. 9 shows the TID sensitivity of ResNet-50, a state-of-the-art CNN for the ImageNet dataset. This network attains 75.6% floating-point top-1 accuracy (92.3% top-5 accuracy) on the 5,000 validation images used for our evaluation. The accuracy falls by 10% relative to this value at 12 krad(Si) for the one-sided mapping scheme and 9 krad(Si) for the two-sided mapping scheme. Remarkably, despite the much greater difficulty of the ImageNet task compared to CIFAR-10, the ResNet-50 network is only slightly more sensitive to TID compared to ResNet-56v1. This is in stark contrast to the large difference in noise sensitivity seen between CIFAR-10 and ImageNet when running the two tasks using similar ResNet topologies [30]. This suggests that the TID sensitivity seen in Fig. 8 and Fig. 9 originate primarily from the topology of the network rather than the complexity of the task. ResNet-50 and ResNet-56v1 are similar in the total depth of the network and therefore the number of cascaded matrix-vector multiplications, even though they differ dramatically in the network size as measured by the total number of weights or arithmetic operations. Therefore, if sufficiently deep, the network's depth appears to be the dominant factor that determines its TID sensitivity. Further evaluation of networks with different depths will be needed to fully validate this finding.

In addition to TID, a SONOS-based inference accelerator may be affected by single event effects, such as upsets in random individual synaptic cells due to heavy ion irradiation. Such upsets would affect the inference accuracy in a different way from TID, which gradually degrades the information stored in all of the cells. Investigation of single event effects, which are also relevant for geosynchronous and low-earth orbit applications, is an important subject of future work.

V. Conclusion

We have examined the resilience of neural network inference accelerators implemented with scaled SONOS memory when exposed to ionizing radiation. Our evaluation is based on a device-level model of TID response, which is fit to the measured characteristics of irradiated SONOS digital memory. The sensitivity of inference accuracy to TID depends upon the proximity of the selected synaptic weight levels to the neutral point of the device, which is immune to radiation. Considering the effect of radiation on the SONOS devices alone, the magnitude of the inference accuracy degradation at a given TID depends on the neural network topology, particularly the depth of the network, and varies from 10 krad(Si) to 100 krad(Si) for networks that we have evaluated on CIFAR-10 and ImageNet. This level of radiation hardness may be suitable for deployment at geosynchronous orbit. In addition, the weights in the SONOS devices can be periodically refreshed to further extend the life of the accelerator for deep space missions.

ACKNOWLEDGMENT

We would like to acknowledge Ben Feinberg for useful comments on the manuscript.

REFERENCES

[1] G. W. Burr *et al.*, "Neuromorphic computing using non-volatile memory," *Adv. Phys. X,* vol. 2, no. 1, pp. 89-124, 2017/01/02 2017.

[2] M. J. Marinella *et al.*, "Multiscale Co-Design Analysis of Energy, Latency, Area, and Accuracy of a ReRAM Analog Neural Training Accelerator," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.,* vol. 8, no. 1, pp. 86-101, 2018.

[3] V. Agrawal *et al.*, "In-Memory Computing array using 40nm multibit SONOS achieving 100 TOPS/W energy efficiency for Deep Neural Network Edge Inference Accelerators," in *2020 IEEE Int. Memory Workshop*, pp. 1-4, May 2020.

[4] C. H. Bennett *et al.*, "Device-aware inference operations in SONOS nonvolatile memory arrays," in *2020 IEEE Int. Reliability Physics Symp.*, pp. 3C.2.1-3C.2.6, Apr.-May 2020.

[5] S. Agarwal *et al.*, "Using Floating-Gate Memory to Train Ideal Accuracy Neural Networks," *IEEE J. Explor. Solid-State Computat.,* vol. 5, no. 1, pp. 52-57, 2019.

[6] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *2017 IEEE Int. Electron Devices Mtg.*, pp. 6.5.1-6.5.4, Dec. 2017.

[7] Y. J. Park *et al.*, "3-D Stacked Synapse Array Based on Charge-Trap Flash Memory for Implementation of Deep Neural Networks," *IEEE Trans. Electron Devices,* vol. 66, no. 1, pp. 420-427, 2019.

[8] Y. Du *et al.*, "An Analog Neural Network Computing Engine Using CMOS-Compatible Charge-Trap-Transistor (CTT)," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.,* vol. 38, no. 10, pp. 1811-1819, 2019.

[9] Z. Ye, R. Liu, J. L. Taggart, H. J. Barnaby, and S. Yu, "Evaluation of Radiation Effects in RRAM-Based Neuromorphic Computing System for Inference," *IEEE Trans. Nucl. Sci,* vol. 66, no. 1, pp. 97-103, 2019.

[10] R. B. Jacobs-Gedrim *et al.*, "Training a Neural Network on Analog TaOx ReRAM Devices Irradiated With Heavy Ions: Effects on Classification Accuracy Demonstrated With CrossSim," *IEEE Trans. Nucl. Sci,* vol. 66, no. 1, pp. 54-60, 2019.

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

[12] P. J. McWhorter, S. L. Miller, and T. A. Dellin, "Radiation Response of SNOS Nonvolatile Transistors," *IEEE Trans. Nucl. Sci,* vol. 33, no. 6, pp. 1413-1419, 1986.

[13] S. Gerardin *et al.*, "Radiation Effects in Flash Memories," *IEEE Trans. Nucl. Sci,* vol. 60, no. 3, pp. 1953-1969, 2013.

[14] B. Draper, R. Dockerty, M. Shaneyfelt, S. Habermehl, and J. Murray, "Total Dose Radiation Response of NROM-Style SOI Non-Volatile Memory Elements," *IEEE Trans. Nucl. Sci,* vol. 55, no. 6, pp. 3202-3205, 2008.

[15] M. Bagatin *et al.*, "Total Ionizing Dose Effects in 3-D NAND Flash Memories," *IEEE Trans. Nucl. Sci,* vol. 66, no. 1, pp. 48-53, 2019.

[16] H. Puchner, P. Ruths, V. Prabhakar, I. Kouznetsov, and S. Geha, "Impact of Total Ionizing Dose on the Data Retention of a 65 nm SONOS-Based NOR Flash," *IEEE Trans. Nucl. Sci,* vol. 61, no. 6, pp. 3005-3009, 2014.

[17] M. Li *et al.*, "Total Ionizing Dose Effects of 55-nm Silicon-Oxide-Nitride-Oxide-Silicon Charge Trapping Memory in Pulse and DC Modes," *Chinese Phys. Lett.,* vol. 35, no. 7, p. 078502, 2018/07 2018.

[18] V. J. Reddi *et al.*, "MLPerf Inference Benchmark," in *2020 ACM/IEEE Annu. Int. Symp. Computer Architecture*, pp. 446-459, May-Jun. 2020.

[19] C. Wu *et al.*, "Machine Learning at Facebook: Understanding Inference at the Edge," in *2019 IEEE Int. Symp. High Performance Computer Architecture*, pp. 331-344, Feb. 2019.

[20] M. L. French and M. H. White, "Scaling of multidielectric nonvolatile SONOS memory structures," *Solid State Electron.,* vol. 37, no. 12, pp. 1913-1923, 1994.

[21] T. R. Oldham and F. B. McLean, "Total ionizing dose effects in MOS oxides and devices," *IEEE Trans. Nucl. Sci,* vol. 50, no. 3, pp. 483-499, 2003.

[22] G. A. Ausman Jr, "Field Dependence of Geminate Recombination in a Dielectric Medium," in "Adelphi, MD, Harry Diamond Lab. Tech. Rep.HDL-TR-2097," 1987.

[23] T. H. Kim, I. H. Park, J. D. Lee, H. C. Shin, and B.-G. Park, "Electron trap density distribution of Si-rich silicon nitride extracted using the modified negative charge decay model of silicon-oxide-nitride-oxide-silicon structure at elevated temperatures," *Appl. Phys. Lett.,* vol. 89, no. 6, p. 063508, 2006.

[24] J. Robertson and M. J. Powell, "Gap states in silicon nitride," *Appl. Phys. Lett.,* vol. 44, no. 4, pp. 415-417, 1984.

[25] S. Agarwal *et al.*, "Achieving ideal accuracies in analog neuromorphic computing using periodic carry," in *2017 Symp. VLSI Technology*, pp. T174-T175, Jun. 2017.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 770-778, Jun.-Jul. 2016.

[27] F. Chollet. (2015). *Keras*. Available: https://github.com/fchollet/keras

[28] A. Shafiee *et al.*, "ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars," in *2016 ACM/IEEE Annu. Int. Symp. Computer Architecture*, pp. 14-26, Jun. 2016.

[29] B. Jacob *et al.*, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2704-2713, Jun. 2018.

[30] V. Joshi *et al.*, "Accurate deep neural network inference using computational phase-change memory," *Nat. Commun.,* vol. 11, no. 1, p. 2473, 2020.