A Case for Lifetime Reliability-Aware Neuromorphic Computing

Shihao Song and Anup Das
Electrical and Computer Engineering Drexel University, Philadelphia, PA, USA
Email: {shihao.song,anup.das}@drexel.edu

***Abstract*—Neuromorphic computing with non-volatile memory (NVM) can significantly improve performance and lower energy consumption of machine learning tasks implemented using spike-based computations and bio-inspired learning algorithms. High voltages required to operate certain NVMs such as phase-change memory (PCM) can accelerate aging in a neuron's CMOS circuit, thereby reducing the lifetime of neuromorphic hardware. In this work, we evaluate the long-term, i.e., lifetime reliability impact of executing state-of-the-art machine learning tasks on a neuromorphic hardware, considering failure models such as negative bias temperature instability (NBTI) and time-dependent dielectric breakdown (TDDB). Based on such formulation, we show the reliability-performance trade-off obtained due to periodic relaxation of neuromorphic circuits, i.e., a stop-and-go style of neuromorphic computing.**

***Index Terms*—Neuromorphic Computing, Non-Volatile Memory (NVM), Phase-Change Memory (PCM), NBTI, TDDB**

I. Introduction

Spiking neural network (SNN) [1] is a machine learning technique designed using spike-based computation and bio-inspired learning algorithms [2]. Neuromorphic hardware such as DYNAP-SE [3], TrueNorth [4], and Loihi [5] can execute SNN-based machine learning tasks in an energy-efficient manner, thanks to low-power neuron circuits [6], distributed implementation of computing and storage as crossbars [7], and the integration of non-volatile memory (NVM) for synaptic storage [8], [9]. Several techniques are recently proposed to map and execute SNNs on to neuromorphic hardware [10]–[15]. These techniques mostly target performance (e.g., accuracy) and energy of neuromorphic computing. Unfortunately, neuromorphic hardware are prone to reliability issues such as limited programming endurance, read disturbance of NVM cells, and aging of CMOS-based neuron circuits [16]–[18]. In this work, we focus on the circuit aging due to negative bias temperature instability (NBTI) and time-dependent dielectric breakdown (TDDB) failure mechanisms [19]–[21].

Due to the high voltage operating requirement of NVM, CMOS devices in a neuron circuit are exposed to high-voltage induced stress when propagating excitation (i.e., current) through an NVM synapse. This impacts the long-term, i.e., lifetime reliability of neuromorphic hardware. As memory process technology scales down to smaller dimensions, reliability issues are expected to exacerbate due to the following three reasons. First, the electric field and power density increase in scaled nodes, exceeding their corresponding maximum value for reliable operation. Second, increasing power density also leads to higher chip temperatures and consequently, an even faster acceleration of the degradation mechanisms. Third, new materials like high-k dielectrics and novel devices such as multi-gate field-effect transistor (FET) that are commonly used for the neuron circuit in neuromorphic hardware have unknown reliability behavior and they introduce new failure mechanisms at scaled nodes. In our recent work [22], we have analyzed NBTI failure in neuromorphic computing. This work extends our earlier work in the following three directions. First, we consider other failure mechanisms such as TDDB and show the impact of system-level design decisions on the circuit aging in neuromorphic hardware. Second, we consider aging in a neuron circuit, which drives current into a crossbar to read synaptic weights stored in its NVM cells. Third, we show the performance-reliability trade-off in periodic relaxation of neuron excitations in neuromorphic hardware using state-of-the-art machine learning applications.

II. Modeling Reliability of Crossbars

A. NBTI Issues in Neuromorphic Computing

This is a failure mechanism of CMOS devices inside a neuron, when positive charges are trapped at the oxide-semiconductor boundary underneath the gate of a CMOS [23]. NBTI manifests as 1) decrease in drain current and transconductance, and 2) increase in off current and threshold voltage. The lifetime of a CMOS device is measured in terms of its *mean time to failure* (**MTTF**) as $\text{MTTF}_{\text{NBTI}} = \frac{A}{V^{\gamma}} e^{\frac{E_a}{KT}}$, where A and γ are material-related constants, E_a is the activation energy, K is the Boltzmann constant, T is the temperature, and V is the overdrive gate voltage of the CMOS device.

Recent studies suggest that a portion of the threshold voltage can be recovered by annealing at high temperatures if the NBTI stress voltage is removed. Figure 1 illustrates the stress and recovery of threshold voltage of a CMOS device due to NBTI failure mechanism on application of a high (V_{read} = 1.8V) and a low voltage (V_{idle} = 1.2V) to a CMOS device in a neuron circuit. We observe that both stress and recovery depends on the time of exposure to the corresponding voltage [24].

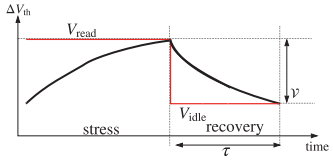


Fig. 1. Demonstration of degradation due to NBTI.

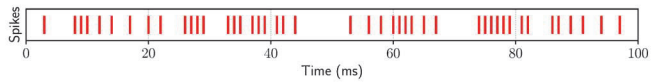
B. TDDB Issues in Neuromorphic Computing

This is a failure mechanism in a CMOS device, when the gate oxide breaks down as a result of long-time application of relatively low electric field (as opposed to immediate breakdown, which is caused by strong electric field) [25]. The TDDB lifetime of a CMOS device is $\text{MTTF}_{\text{TDDB}} = A.e^{-\gamma\sqrt{V}}$, where A and γ are material-related constants, and V is the overdrive gate voltage of the CMOS device [26].

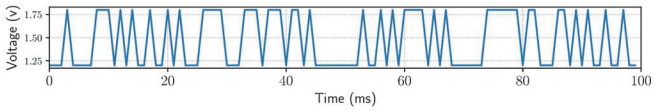
C. Circuit Aging in Neuromorphic Computing

To illustrate the degradation caused by these failure mechanisms, we take the example of a single neuron of the LeNet convolutional neural network (CNN) [27] used for handwritten digit recognition and illustrate its spike times within the first

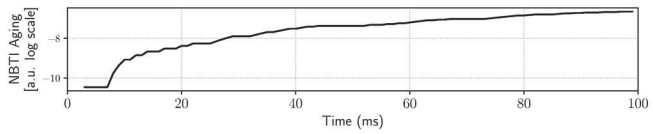
100ms in Figure 2a. The voltage required to propagate these spikes through the neuron's fanout synapses are shown in Figure 2b. Figures 2c and 2d show the NBTI and TDDB aging of a CMOS device inside the neuron's circuit, respectively. As can be clearly seen, both aging increases with time as more spikes are generated by the neuron. If CMOS devices in the neuron circuit are not de-stressed regularly, the aging (both NBTI and TDDB) in a neuron continues to increase, eventually leading to transient, intermittent, or permanent faults in the neuromorphic hardware.



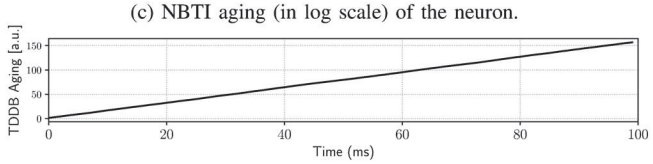
(a) Spike times of a neuron in LeNet CNN.



(b) Voltage of the neuron to process the spike train.



(c) NBTI aging (in log scale) of the neuron.



(d) TDDB aging (in log scale) of the neuron.

Fig. 2. (a) Spike times of a neuron in LeNet, (b) voltages needed to propagate these spikes through its fanout synapses, (c) NBTI degradation (in arbitrary units), and (d) TDDB degradation (in arbitrary units) of CMOS devices.

To de-stress a neuron, all CMOS devices in the neuron must be programmed with a voltage lower than the threshold voltage V_{th}, which forces them to operate in the sub-threshold region, relieving their stress. Once discharged, a neuron requires several clock cycles to boost its voltage back to the required voltage level, before it can safely be used to generate spikes again. This introduces performance overhead.

III. Periodic Relaxation of Neuromorphic Circuits

To improve the long-term, i.e., the lifetime reliability of neuromorphic computing, we propose periodic relaxation of a neuromorphic architecture, where we de-stress all neurons in the hardware at fixed intervals. To compute the overhead due to such de-stress operations, we assume that the controller issues a de-stress command to a crossbar once every $tDSI$, which is known as the *de-stress interval*. Each de-stress operation completes within a time interval $tDSC$, known as the *de-stress cycle time*. Hence, the performance overhead (i.e., spike throughput loss) due to periodic de-stress is

$$\text{de-stress overhead} = tDSC/tDSI. \tag{1}$$

Figure 3 shows an example where four spikes (S1, S2, S3, & S4) generated by a neuron. These spikes have some idle time between them. The neuron circuits are de-stressed after every $tDSI$, such that the aging due to NBTI and TDDB (indicated by $\mathcal{A}_{\text{TDDB}}$ and $\mathcal{A}_{\text{NBTI}}$, respectively) are lower than 1000 units. Using this approach, the de-stress operation is initiated upon generating S3, which increases the latency of S4 due to the non-zero latency of the de-stress operation (indicated by $tDSC$). Increase in spike latency can lead to information loss in SNNs and degrade the quality of response.

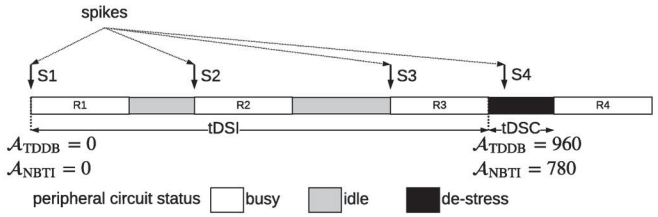


Fig. 3. Performance impact due to periodic relaxation.

We introduce two key performance metrics in SNNs that are affected due to periodic de-stressing of neuromorphic architectures – inter-spike interval (ISI) and disorder spike count. These are defined as follows.

- **Inter-spike interval distortion:** Performance of supervised machine learning is measured in terms of *accuracy*, which can be assessed from inter-spike intervals (ISIs) [28]. To define ISI, we let $\{t_1, t_2, \cdots, t_K\}$ be a neuron's firing times in the time interval $[0, T]$. The average ISI of this spike train is given by [28]:

$$\mathcal{I} = \sum_{i=2}^{K} (t_i - t_{i-1})/(K-1). \tag{2}$$

- **Disorder spike count:** This is defined for SNNs where information is encoded in terms of spike rate. We formulate spike disorder as follows. Let $F^i = \{F^i_1, \cdots, F^i_{n_i}\}$ be the expected spike arrival rate at neuron i and $\hat{F}^i = \{\hat{F}^i_1, \cdots, \hat{F}^i_{n_i}\}$ be the actual spike rate considering de-stress latencies. The spike disorder is computed as

$$\text{spike disorder} = \sum_{j=1}^{n_i} [(F^i_j - \hat{F}^i_j)^2]/n_i \tag{3}$$

IV. Evaluation

We evaluate 10 standard machine learning applications, which are listed in Table I.

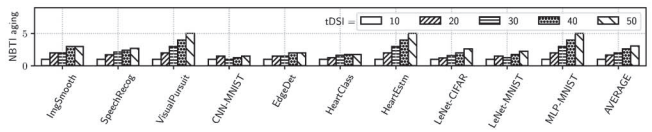
TABLE I
Applications used to evaluate our approach [10].

Class	Applications	Synapses	Neurons	Topology	Accuracy
MLP	EdgeDet	272,628	1,372	FeedForward (4096, 1024, 1024, 1024)	100%
	ImgSmooth	136,314	980	FeedForward (4096, 1024)	100%
	MLP-MNIST	79,400	984	FeedForward (784, 100, 10)	95.5%
CNN	CNN-MNIST	159,553	5,576	CNN	96.7%
	LeNet-MNIST	1,029,286	4,634	CNN	99.1%
	LeNet-CIFAR	2,136,560	18,472	CNN	84.0%
	HeartClass [29], [30]	2,396,521	24,732	CNN	85.12%
RNN	HeartEstm [31]	636,578	6,952	Recurrent Reservoir	99.2%
	SpeechRecog	636,578	6,952	Recurrent Reservoir	96.8%
	VisualPursuit	636,578	6,952	Recurrent Reservoir	89.0%

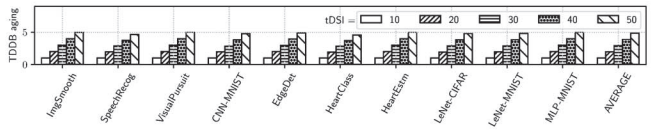
A. Reliability

Figures 4a and 4b plot respectively, the NBTI and TDDB aging of the 10 machine learning applications when increasing the tDSI from 10ms to 50ms. We make the following three key observations. First, both NBTI and TDDB aging increases with increase in tDSI. This is because, a neuron accrues higher aging when its CMOS devices are kept active for longer duration (i.e., for higher tDSI). Second, the increase in aging is application-dependent. For CNN-MNIST, increasing tDSI from 10ms to 50ms leads to 50% increase in NBTI aging, compared to VisualPursuit, where the NBTI aging increase by 5x. This is because, the number of spikes generated in CNN-MNIST is far fewer than in VisualPursuit, which leads to lower aging in neuron circuits. Therefore, the impact of increasing tDSI for CNN-MNIST is less significant compared

to VisualPursuit. Third, compared to NBTI, the increase of TDDB aging is consistent across different applications for the same range of tDSI. This is due to the difference in the two mechanisms. NBTI-induced stress (e.g., V_{th} shift) recovers partially when the neuron is idle. On the other hand, a CMOS devices encounters low-voltage TDDB stress even when idle.



(a) NBTI aging for 10 machine learning applications.

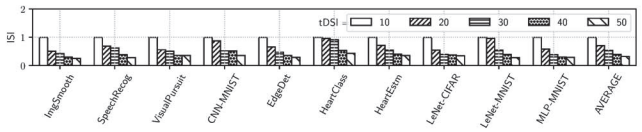


(b) TDDB aging for 10 machine learning applications.

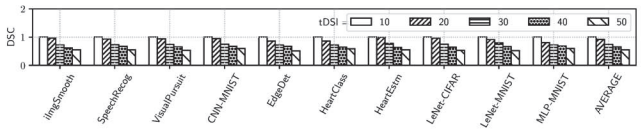
Fig. 4. (a) Normalized NBTI aging, and (b) Normalized TDDB aging for tDSI of 10ms, 20ms, 30ms, 40ms, and 50ms.

B. Performance

Figures 5a and 5b plot respectively, the ISI distortion and disorder spike count (DSC) of the 10 machine learning applications when increasing the tDSI from 10ms to 50ms. We observe that both ISI and DSC reduce with increase in tDSI. This reduction is due to the reduction of the de-stress overhead (Equation 1) with an increase in tDSI.



(a) ISI distortion for 10 machine learning applications.



(b) Disorder spike count for 10 machine learning applications.

Fig. 5. ISI, disorder for tDSI of 10ms, 20ms, 30ms, 40ms, and 50ms.

C. Thermal Impact

The results of Sections IV-A are obtained at nominal temperature of 300K. Prior works such as [32] show the impact of temperature on reliability of conventional multiprocessor system. Figure 6 shows the increase of aging with temperature. Average circuit aging at 325K and 350K is higher than that at 300K by an average of 7% and 26%, respectively.

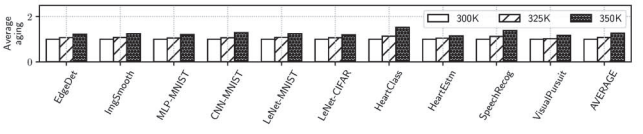


Fig. 6. Average circuit aging at 325K and 350K normalized to aging at 300K.

V. Conclusion

We evaluate circuit aging in the neurons of neuromorphic architectures considering NBTI and TDDB failure mechanisms. We then propose a simple approach to improve reliability by periodically de-stressing its neurons. This introduces latency, which degrades key performance metrics such as inter-spike interval and disorder spike count, which correlates directly to the performance of machine learning models. We evaluate reliability-performance trade-offs for 10 state-of-the-art machine learning applications. We **conclude** that the proposed work will enable intelligent reliability optimization strategies in neuromorphic computing.

Acknowledgment

This work is supported by the National Science Foundation Faculty Early Career Development Award CCF-1942697 (CAREER: Facilitating Dependable Neuromorphic Computing: Vision, Architecture, and Impact on Programmability).

References

[1] W. Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural Networks*, 1997.
[2] Y. Dan *et al.*, "Spike timing-dependent plasticity of neural circuits," *Neuron*, 2004.
[3] S. Moradi *et al.*, "A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs)," *TBCAS*, 2017.
[4] M. V. DeBole *et al.*, "TrueNorth: Accelerating from zero to 64 million neurons in 10 years," *Computer*, 2019.
[5] M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, 2018.
[6] G. Indiveri, "A low-power adaptive integrate-and-fire neuron circuit," in *ISCAS*, 2003.
[7] P. Merolla *et al.*, "A digital neurosynaptic core using embedded crossbar memory with 45pJ per spike in 45nm," in *CICC*, 2011.
[8] G. W. Burr *et al.*, "Neuromorphic computing using non-volatile memory," *Advances in Physics: X*, 2017.
[9] A. Mallik *et al.*, "Design-technology co-optimization for OxRRAM-based synaptic processing unit," in *VLSIT*, 2017.
[10] S. Song *et al.*, "Compiling spiking neural networks to neuromorphic hardware," in *LCTES*, 2020.
[11] A. Balaji *et al.*, "PyCARL: A PyNN interface for hardware-software co-simulation of spiking neural network," in *IJCNN*, 2020.
[12] A. Balaji *et al.*, "Mapping spiking neural networks to neuromorphic hardware," *TVLSI*, 2019.
[13] A. Das *et al.*, "Mapping of local and global synapses on spiking neuromorphic hardware," in *DATE*, 2018.
[14] A. Das *et al.*, "Dataflow-based mapping of spiking neural networks on neuromorphic hardware," in *GLSVLSI*, 2018.
[15] A. Balaji *et al.*, "Run-time mapping of spiking neural networks to neuromorphic hardware," *JSPS*, 2020.
[16] P.-Y. Chen *et al.*, "Reliability perspective of resistive synaptic devices on the neuromorphic system performance," in *IRPS*, 2018.
[17] B. Gleixner *et al.*, "Reliability characterization of phase change memory," in *NVMTS*, 2009.
[18] A. Pirovano *et al.*, "Reliability study of phase-change nonvolatile memories," *TDMR*, 2004.
[19] C. Hu, "Future CMOS scaling and reliability," *Proc. of the IEEE*, 1993.
[20] A. Das *et al.*, "Aging-aware hardware-software task partitioning for reliable reconfigurable multiprocessor systems," in *Compilers, Architecture and Synthesis for Embedded Systems (CASES)*, 2013, p. 1.
[21] S. Song *et al.*, "Exploiting inter-and intra-memory asymmetries for data mapping in hybrid tiered-memories," in *ISMM*, 2020.
[22] A. Balaji *et al.*, "A framework to explore workload-specific performance and lifetime trade-offs in neuromorphic computing," *CAL*, 2019.
[23] R. Gao *et al.*, "NBTI-generated defects in nanoscaled devices: fast characterization methodology and modeling," *TED*, 2017.
[24] S. Song *et al.*, "Improving dependability of neuromorphic computing with non-volatile memory," in *EDCC*, 2020.
[25] P. Roussel *et al.*, "New methodology for modelling MOL TDDB coping with variability," in *IRPS*, 2018.
[26] A. Das *et al.*, "Communication and migration energy aware task mapping for reliable multiprocessor systems," *FGCS*, 2014.
[27] Y. LeCun *et al.*, "LeNet-5, convolutional neural networks," 2015.
[28] S. Grün *et al.*, *Analysis of parallel spike trains*. Springer, 2010.
[29] A. Balaji *et al.*, "Power-accuracy trade-offs for heartbeat classification on neural networks hardware," *JOLPE*, 2018.
[30] A. Das *et al.*, "Heartbeat classification in wearables using multi-layer perceptron and time-freq joint distribution of ECG," in *CHASE*, 2019.
[31] A. Das *et al.*, "Unsupervised heart-rate estimation in wearables with Liquid states and a probabilistic readout," *Neural Networks*, 2018.
[32] A. Das *et al.*, "Reliability and energy-aware mapping and scheduling of multimedia applications on multiprocessor systems," *TPDS*, 2016.