Empirical Metal-Oxide RRAM Device Endurance and Retention Model for Deep Learning Simulations

Corey Lammie[1], **Mostafa Rahimi Azghadi**[1], **and Daniele Ielmini**[2]

[1]College of Science and Engineering, James Cook University, Townsville, QLD 4814, Australia
[2]Politecnico di Milano, 20133 Milano, Italy

E-mail:
corey.lammie,mostafa.rahimiazghadi@jcu.edu.au;daniele.ielmini@polimi.it

Abstract. Memristive devices including Resistive Random Access Memory (RRAM) cells are promising nanoscale low-power components projected to facilitate significant improvement in power and speed of Deep Learning (DL) accelerators, if structured in crossbar architectures. However, these devices possess non-ideal endurance and retention properties, which should be modeled efficiently. In this paper, we propose a novel generalized empirical Metal-Oxide RRAM endurance and retention model for use in large-scale DL simulations. To the best of our knowledge, the proposed model is the first to unify retention-endurance modeling while taking into account time, energy, SET-RESET cycles, device size, and temperature. We compare the model to state-of-the-art and demonstrate its versatility by applying it to experimental data from fabricated devices. Furthermore, we use the model for CIFAR-10 dataset classification using a large-scale Deep Memristive Neural Network (DMNN) implementing the MobileNetV2 architecture. Our results show that, even when ignoring other device non-idealities, retention and endurance losses significantly affect the performance of DL networks. Our proposed model and its DL simulations are made publicly available.

Keywords: Metal-Oxide RRAM, Deep Learning, Endurance, Retention, Simulation

Accepted by Semiconductor Science and Technology.

1. Introduction

RRAM devices have attracted significant attention for use in next generation DL and neuromorphic architectures to perform in-memory computing operations, which can reduce power usage and time complexity, massively augmenting performance [1–4]. However, RRAM is an emerging technology with a number of limitations including endurance and retention losses, as depicted in Fig. 1. Consequently, significant research efforts are being made to efficiently and accurately model device limitations to improve the reliability and robustness of RRAM-based DL architectures [5–7].

In this paper, we propose a generalized empirical Metal-Oxide RRAM device endurance and retention model. We compare our model to related works and demonstrate its versatility by using it to fit experimental data from several devices. We then deploy the model within large-scale DL simulations to implement the MobileNetV2 Convolutional Neural Network (CNN) architecture to investigate how device endurance and retention losses affect inference performance using the CIFAR-10 dataset.

2. Related Work

Previous works have investigated Metal-Oxide RRAM endurance and retention losses experimentally [8–15], numerically [16], and analytically [17–23]. Table 1 compares the proposed model with previous numerical and analytical RRAM device-level endurance and retention models. Given the increasing popularity of RRAM-based DMNNs, a number of works [8, 11–13, 17–20] specifically consider endurance and retention loss effects on DMNNs performance. While most models [17, 21–23] are inherently physics-based and model various phenomena and internal device mechanics using fundamental physics principals, others [16, 18–20] adopt a generalized high-level approach, and model device behavior empirically. Our model fits into the latter group, and is the first to:

1. Accurately model device endurance and retention behavior, before and after the conductance window begins to collapse; and
2. Model both gradual and sudden window collapse; and
3. Model temperature, cell size, and when modeling endurance, the voltage dependence, V_{stop}; and

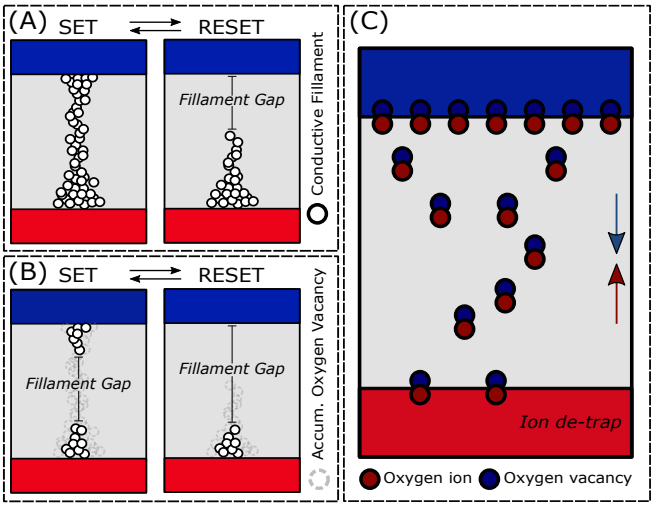

Figure 1: (A) The formation of a conductive filament within metal-oxide RRAM devices results in low resistive states, whereas its partial destruction increases the resistivity to high resistive states. (B) When a voltage is applied, defects are gradually created within the conductive filament [24], which cause endurance losses. (C) Oxygen ions return to the previous thermal equilibrium state during the baking process, which causes retention losses.

4. Model endurance and retention interchangeably using a unified methodology.

The proposed model is well suited toward DL modeling using memristors, as the behavior of new devices can easily be modeled using tools provided in our supplementary materials ‡, it is highly integrable ‡, and it is able to capture a large range of Metal-Oxide RRAM device behavior, as depicted in Fig. 2, Fig. 3, and Fig. 4.

Table 1: Comparison of RRAM endurance and retention models. †Models are defined independently.

Model	Models	
	Endurance	Retention
Endurance Statistical [16]	✓	
Statistical State Instability and Retention [17]		✓
Reliability Perspective [18–20]	✓†	✓†
Endurance, Retention and Window Margin [21]	✓	
Retention Model for High-Density RRAM [22]		✓
Voltage-Controlled Cycling Endurance [23]	✓	
Proposed	✓	✓

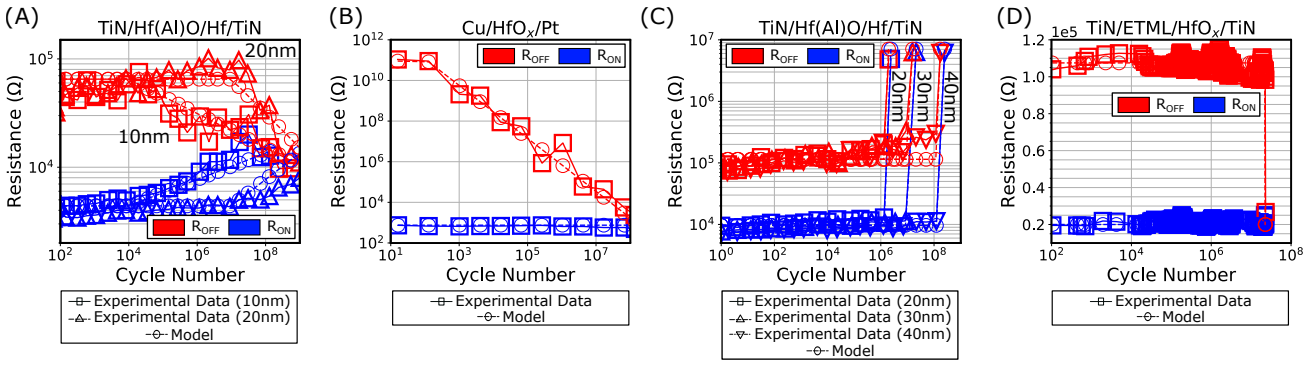

Figure 2: Experimental endurance data from various Metal-Oxide RRAM device types, and the behavior of our proposed model. (A) TiN/Hf(Al)O/Hf/TiN [25] devices with different cell sizes, (B) Cu/HfO$_x$/Pt [26] devices, and results from the proposed model in gradual resistance convergence operation mode; (C) TiN/Hf(Al)O/Hf/TiN [25] devices with different cell sizes, (D) TiN/Electro-thermal Modulation Layer (ETML)/HfO$_x$/TiN [8] devices, and results from the proposed model in sudden resistance convergence operation mode.

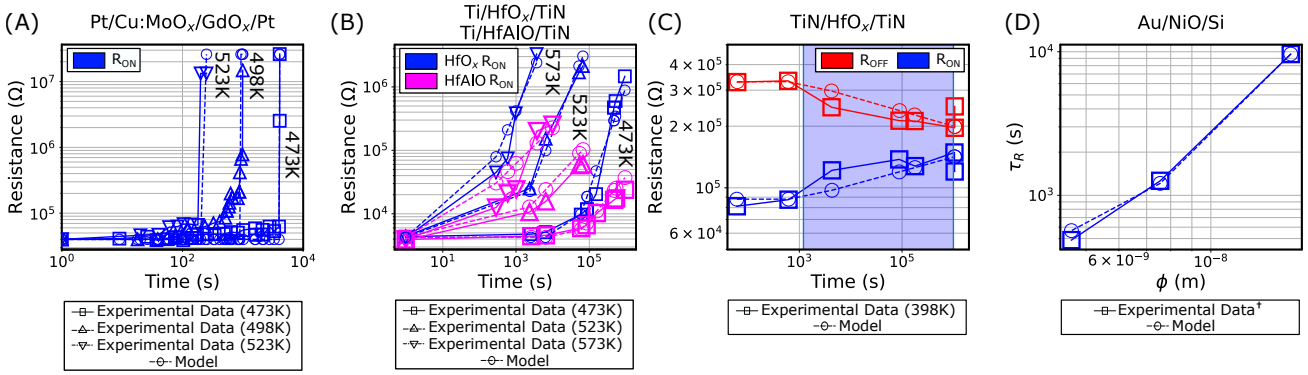

Figure 3: Experimental retention data from various Metal-Oxide RRAM device types, and the behavior of our proposed model. (A) Pt/Cu:MoO$_x$/GdO$_x$/Pt [27] devices operating at different temperature points, and results from the proposed model in sudden resistance convergence operation mode; (B) TiN/HfO$_x$/TiN and Ti/HfAlO/TiN devices [28] operating at different temperature points, and results from the proposed model in sudden resistance convergence operation mode; (C) TiN/HfO$_x$/TiN [29] devices and results from the proposed model in gradual operation mode, where the temperature was elevated from room temperature (25°C) to 125°C depicted using a blue background segment between 1200s and 10^6s; (D) Relationship between the retention time to failure, τ_R, and conductive filament diameter, ϕ, of Au/NiO/Si [30] devices, and results from the proposed model, where ϕ is substituted for the cell size, and $\tau_R \propto e_{th}$, i.e., $\tau_R = p_0 e^{p_1\phi + p_2 T_c}$.† The conductive filament size, which can be representative of device dimension, was obtained using a piecewise linear fit of the mean activation energy, E_{AC}, as done in [30].

3. Proposed Model

The proposed model has two modes of operation. The first mode assumes that resistance states gradually converge after a device-specific threshold energy level is exceeded, and can be used to model device endurance and retention, as depicted in Fig. 2 (A,B) and Fig. 3 (B,C). The second mode, on the other hand, assumes sudden failure, and can be used to model device endurance, as depicted in Fig. 2 (C,D) and Fig. 3 (A). The gradual convergence of resistance states is modeled as

$$R(x,s,T) = \begin{cases} R_0 & x \leqslant e_{th} \\ 10^{p_3(p_1 s + p_2 T_c)\log(x) + \log(R_0) - p_3(p_1 s + p_2 T_c)\log(e_{th})} & \text{otherwise,} \end{cases} \quad (1)$$

and the sudden convergence of states is modeled as

$$R(x,s,T) = \begin{cases} R_0 & x \leqslant e_{th} \\ R_\infty & \text{otherwise,} \end{cases} \quad (2)$$

where the device-specific threshold energy level, e_{th}, if exceeded, causes the resistance window of a device to collapse either gradually or suddenly, is modeled using

$$e_{th} = p_0 e^{p_1 s + p_2 T_c}. \quad (3)$$

The temperature constant, T_c, is expressed as

$$T_c = \min(\frac{T_{th}}{T}, 1), \quad (4)$$

which is used to introduce temperature dependence to the model. Using (1)–(4), the resistance state of a device, R, is determined using four parameters, x, s, T, and T_{th}, and various fitting parameters, $\mathbf{p}$. Here, R_∞, the collapsed resistive state to which R_{ON} and R_{OFF} converge, is bounded to the range $[R_{\text{ON}}, R_{\text{OFF}}]$ [8, 15, 23, 25, 29, 31]. x denotes either the time (s), the energy (J), or the number of SET-RESET cycles, s denotes the device cell size (nm), when the depth and width are fixed, or the filament volume (nm^3) when they are not, T denotes the operating temperature (K), and T_{th} denotes the temperature threshold, that if exceeded, accelerates device failure. For both modes of operation, p_0 modulates the magnitude of e_{th}, and p_1 and p_2 modulate the strength of the dependence on s and T, respectively. For instances where s is fixed, $p_1 = 0$, and for instances where T is fixed, $p_2 = 0$. When modeling the gradual convergence of resistance states, p_3 is used to modulate the rate of failure once e_{th} is exceeded. We believe that, given sufficient data, all fitting parameters could be related to physical device parameters, such as those determined using ab initio calculations in [21], including formation enthalpy energies, ΔH, migration barriers, E_d, and hopping distances between sites during ion migration, d_h.

The parameter p_0 in (3) can be modulated using (5) when modeling endurance to introduce dependence to V_{stop}, the most negative voltage in the negative voltage sweep during the RESET cycle [23].

$$p_0 = \frac{10^{K(1-(2\bar{V}_{\text{stop}}-1)^2)}}{e^{p_1 s + p_2 T_c}} \quad (5)$$

K is used to modulate the amplitude, and V_{stop} is mapped to $\bar{V}_{\text{stop}} \in [0, 1]$. Fig. 4 demonstrates the inclusion of V_{stop} dependence in the proposed model, where x is assumed to denote the number of SET-RESET cycles, and the model to be used will operate in sudden resistance convergence mode. In this figure, the optimal point corresponds to the optimal V_{stop} value, i.e., the V_{stop} value for a given device that maximizes e_{th}. Given sufficient experimental data observing the relationship between V_{stop} and e_{th}, the mapping bounds of V_{stop} can be determined, and the K parameter can be determined using Nonlinear Least Squares Regression (NLSR).

The following assumptions are made in our modeling:

1. The waveform used to program each device is constrained, and only V_{stop} is mutable; and
2. The impacts of the compliance current, I_c, and the maximum set voltage magnitude are considered negligible [23]; and
3. Resistance states converge to R_∞ when device failure occurs; and
4. Resistance states are stable until a device-specific threshold energy level, e_{th}, is exceeded; and
5. (5) is constrained to be symmetrical around the optimal point.

4. Model Validation

To validate the proposed model, we fit it to experimental data from various fabricated devices, indicative of a variety of use cases, as shown in Fig. 2, Fig. 3, and Fig. 4. NLSR is used to fit the model empirically to each device type. In Fig. 2 and Fig. 3 (C), two sets of parameters are used to model R_{OFF} and R_{ON}, respectively, for each simulated device. In Fig. 3 (A,B), one set of parameters are used to model R_{ON}. To the best of our knowledge experimental data for R_{OFF} is currently not available in literature.

In Fig. 3 (D), we model the relationship between the retention time to failure, τ_R, and the conductive filament diameter, ϕ, of Au/NiO/Si [30] devices. The conductive filament size, which can be representative of device dimension, was obtained using a piecewise linear fit of the mean activation energy, E_{AC}, which accounts for metallic and semiconductor-like behavior, as done in [30]. $\bar{V}_{\text{stop}}$ dependence is validated in

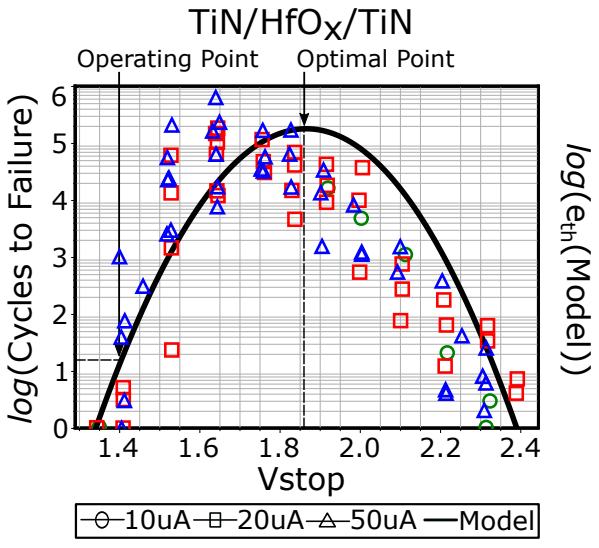

Figure 4: The window function used to model v_{stop} dependence. Experimental data is extracted from $\text{TiN/HfO}_\text{x}\text{/TiN}$ devices [23].

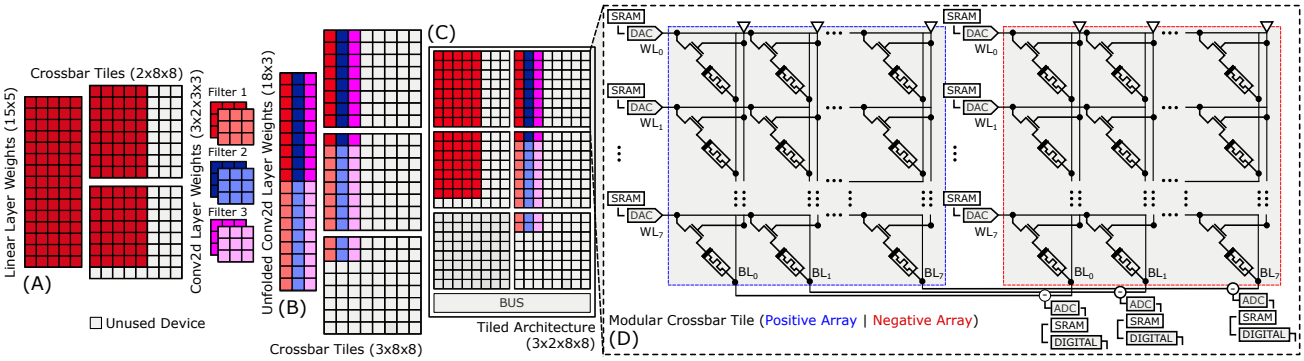

Figure 5: An overview of the mapping process of Linear (dense) and Conv2d (convolutional) layers onto a 3×2 tiled architecture with tiles constructed using 128 ($2 \times 8 \times 8$) devices. (A) Linear layers are mapped directly onto crossbar tiles. (B) Convolutional layers are unfolded before being mapped onto crossbar tiles. (C) Tiled architectures contain several modular crossbar tiles connected using a shared bus. (D) Modular crossbar tiles consist of crossbar arrays with supporting peripheral circuitry, and can represent weights using a dual-array scheme (as depicted), a dual row scheme, where double the number of rows are required, or a current-mirror scheme, that is capable of operation using a singular device to represent each weight [32].

Fig. 4 using TiN/HfO_x/TiN devices [23]. We note that variations between experimental data and the behavior of the proposed model could be further reduced by simulating other device non-idealities, such as conductance drift, evident in Fig. 2 (C) and Fig. 3 (A), however, this is beyond the scope of this paper. In favour of reproducible research, our model, its fitting parameters, and all of the information required to reproduce the reported results are made publicly available ‡.

5. Large-scale Deep Learning Simulations

Exemplar large scale DL simulations were performed that modeled the gradual and sudden resistance state convergence on account of endurance and retention losses of TiN/Hf(Al)O/Hf/TiN, TiN/ETML/HfO_x/TiN, and TiN/Hf_x/TiN RRAM devices using the VTEAM model [33] within layers of a DMNN employing 1-Transistor 1-Resistor crossbars. These crossbars were constructed by converting linear and unfolded convolutional layers from a pre-trained MobileNetV2 CNN that achieved 91.93% accuracy on the CIFAR-10 test set. In Fig. 5, we overview the mapping process of linear and convolutional layers onto a modular tiled architecture. Batch-normalization, pooling, and activation functions, which are simulated in our experiments, should be implemented using additional circuitry to realize the other computations required for a DL task. Inputs are unfolded and scaled, prior to being presented to the network. By generalizing this approach, modular crossbar tiles and digital

‡ https://github.com/coreylammie/SST-Reproducibility

Algorithm 1 Adopted simulation methodology.

1. Map Network Parameters
  for each convolutional and linear layer **do**
    $\boldsymbol{W}_{\max} = \text{descending\_order}(\text{abs}(\boldsymbol{W})[\text{size}(\boldsymbol{W})])$
    $\boldsymbol{W}_{\min} = \boldsymbol{W}_{\max}/(R_{\text{OFF}}/R_{\text{ON}})$
    $\boldsymbol{W}_{\text{pos}} = \boldsymbol{W}[\boldsymbol{W} \geqslant 0]$, $\boldsymbol{W}_{\text{neg}} = \boldsymbol{W}[\boldsymbol{W} < 0]$
    for each device, $\boldsymbol{R}_{\text{pos}}[i,j]$, $\boldsymbol{R}_{\text{neg}}[i,j]$ in $\boldsymbol{W}_{\text{pos}}$, $\boldsymbol{W}_{\text{neg}}$ **do**
      $\boldsymbol{R}_{\text{pos}}[i,j] = \frac{(R_{\text{ON}}-R_{\text{OFF}})(\boldsymbol{W}_{\text{pos}}[i,j]-\mathbf{w}_{\min})}{|\boldsymbol{w}|_{\max}-\mathbf{w}_{\min}} + R_{\text{OFF}}$
      $\boldsymbol{R}_{\text{neg}}[i,j] = \frac{(R_{\text{ON}}-R_{\text{OFF}})(\boldsymbol{W}_{\text{neg}}[i,j]-\mathbf{w}_{\min})}{|\boldsymbol{w}|_{\max}-\mathbf{w}_{\min}} + R_{\text{OFF}}$
    end for
  end for
2. Tune Memristive Layers
  for each converted memristive layer **do**
    if the layer is convolutional **then**
      $P = (8 \times \text{in\_channels} \times 32 \times 32)$
    else if the layer is linear **then**
      $P = (8 \times \text{in\_features})$
    end if
    determine β_0 for $\tilde{Y} = \beta_0 \tilde{X}$, where $\tilde{Y}$ denotes the legacy layer's output and $\tilde{X}$ denotes the converted layer's output when a randomly generated tensor of size P is propagated.
  end for
3. Model Device Endurance and Retention
  for each value of x to simulate **do**
    for each converted memristive layer **do**
      for each device, $\boldsymbol{R}_{\text{pos}}[i,j]$, $\boldsymbol{R}_{\text{neg}}[i,j]$ in $\boldsymbol{W}_{\text{pos}}$, $\boldsymbol{W}_{\text{neg}}$ **do**
        $\boldsymbol{R}_{\text{pos}}[i,j], \boldsymbol{R}_{\text{neg}}[i,j] = R(x, s, T, \tilde{V}_{\text{stop}})$
      end for
    end for
    determine the test set accuracy for the given x value
  end for

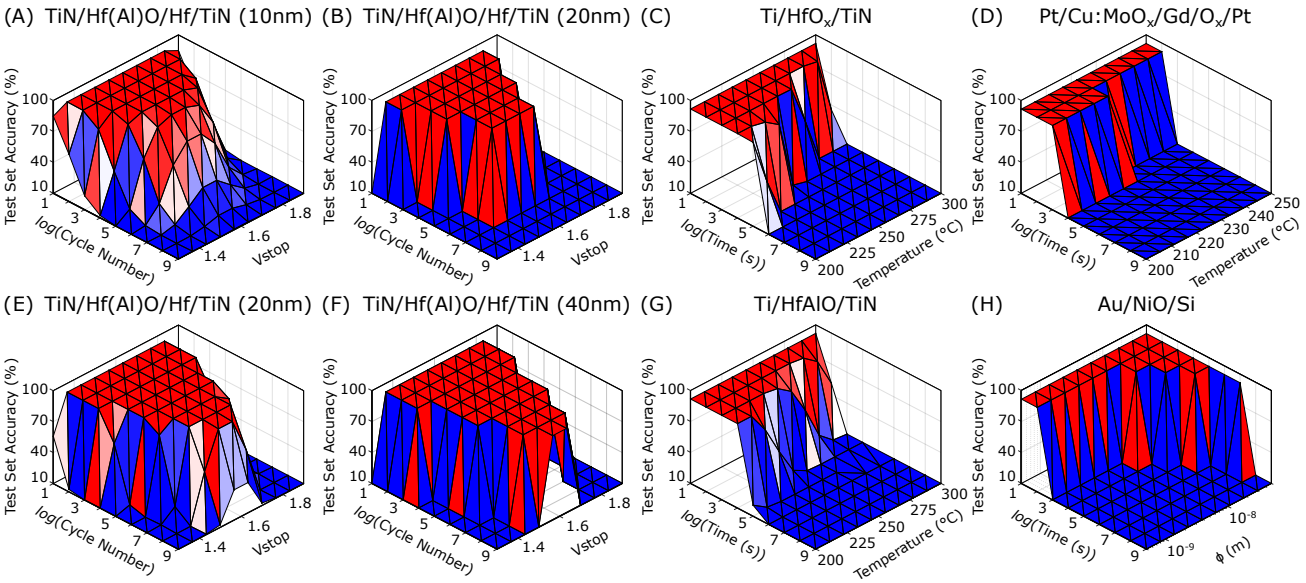

Figure 6: Large-scale DL simulations of TiN/Hf(Al)O/Hf/TiN, TiN/HfO$_x$/TiN, Pt/Cu:MoO$_x$/GdO$_x$/Pt, TiN/HfAlO/TiN, and Au/NiO/Si devices. (A,E) gradual endurance failure; (B,F) sudden endurance failure; (C,G) gradual retention failure; (D,H) sudden retention failure.

circuitry can be used to perform inference of any arbitrary Deep Neural Network (DNN) [34].

Algorithm 1 details our simulation methodology, in which a double-column scheme is used to represent network weights within memristive crossbars, i.e., a dual-array scheme is adopted. All RRAM devices are assumed to operate as fully analog devices, and other device non-idealities are ignored. Analog to Digital Converters (ADCs) are assumed to have a bit-length of 8, and modular crossbar tiles are constructed using two arrays of 128×128 devices, representing positive and negative parameters, respectively. Unfolded inputs are scaled and encoded using voltage signals between $\pm 0.3V$ [35].

After network parameters are mapped, to tune each memristive layer, random inputs of variable size that are sampled from uniform distributions between ± 1.0 are presented to each layer. The readout currents of each column associated with each layer are linearly related to each layer's desired output. Prior to endurance and retention losses, our RRAM-based networks achieved 91.69% accuracy on the CIFAR-10 test set. We attribute the small performance degradation to quantization noise introduced from ADCs and the non-ideal mapping and tuning methodologies employed. The results from six exemplar large-scale DL simulations are presented in Fig. 6. Each surface plot is constructed from the results of 100 individual simulations (one per point).

In Fig. 6 (A,B,E,F), the CIFAR-10 test set accuracy is reported after each SET-RESET cycle to investigate the performance degradation on account of endurance losses, i.e., we assume massive reprogramming in the DNN accelerator is performed. v_{stop} was extrapolated using (5), where $\max(v_{\text{stop}})$ was arbitrarily chosen to be 1.6, due to the unavailability of experimental data on v_{stop}. K and the mapping bounds of v_{stop} were determined using operational points from each device. In Fig. 6 (C,D,G,H), the CIFAR-10 test set accuracy is reported at each time-step to investigate the performance degradation on account of retention losses. TiN/Hf(Al)O/Hf/TiN devices from Fig. 2 (A) are modeled to achieve the results in Fig. 6 (A) and Fig. 6 (E), for devices with cell sizes of 10nm and 20nm, respectively; TiN/Hf(Al)O/Hf/TiN devices from Fig. 2 (C) are modeled to achieve the results in Fig. 6 (B) and Fig. 6 (D), for devices with cell sizes of 20nm and 40nm, respectively; Ti/HfO$_x$/TiN devices from Fig. 3 (B) are modeled to achieve the results in Fig. 6 (C); Pt/Cu:MoO$_x$/GdO$_x$/Pt devices from Fig. 3 (A) are modeled to achieve the results in Fig. 6 (D); Ti/HfAlO/TiN devices from Fig. 3 (B) are modeled to achieve the results in Fig. 6 (G), and Au/NiO/Si devices from Fig. 3 (D) are modeled to achieve the results in Fig. 6 (H). From Fig. 6, it can be observed that the proposed model is capable of robustly modeling endurance and retention losses of Metal-Oxide RRAM devices within large-scale DL simulations.

6. Discussion and Conclusion

We proposed a novel generalized empirical Metal-Oxide RRAM device endurance and retention model for use in large-scale simulations. We demonstrated

its versatility by fitting it to experimental data from various devices, and using it for large DL simulations. Our findings show that, even when other device non-idealities are ignored, endurance and retention losses significantly affect the reprogrammability of DMNNs, degrading their learning and inference accuracy. A limitation of the proposed model is the lack of a clear link between its parameters and physical device parameters. This is mainly due to unavailability of experimental data, which resulted in developing an empirical, rather than a physics-based model. Additionally, while this work only focuses on endurance and retention and their impact on memristive deep learning networks performance, future improvements of our model can account for modelling a finite number of conductance states and other device non-idealities [32, 36].

Acknowledgments

C. Lammie acknowledges the James Cook University (JCU) DRTPS and an IBM PhD Fellowship. M. Rahimi Azghadi acknowledges a JCU Rising Star ECR Fellowship.

References

[1] Mittal S 2018 *Machine Learning and Knowledge Extraction* **1** 75

[2] Chi P, Li S, Xu C, Zhang T, Zhao J, Liu Y, Wang Y and Xie Y 2016 PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory *ACM/IEEE International Symposium on Computer Architecture*

[3] Li B, Song L, Chen F, Qian X, Chen Y and Li H H 2018 ReRAM-based Accelerator for Deep Learning *Design, Automation Test in Europe Conference Exhibition*

[4] Azghadi M R *et al* 2020 *Advanced Intelligent Systems* **2** 1900189

[5] Mao M, Cao Y, Yu S and Chakrabarti C 2016 *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* **6** 352

[6] Valad Beigi M and Memik G 2018 THOR: THermal-aware Optimizations for extending ReRAM Lifetime *IEEE International Parallel and Distributed Processing Symposium*

[7] Mao M, Cao Y, Yu S and Chakrabarti C 2015 Programming Strategies to Improve Energy Efficiency and Reliability of ReRAM Memory Systems *IEEE Workshop on Signal Processing Systems*

[8] Zhao M *et al* 2018 Characterizing Endurance Degradation of Incremental Switching in Analog RRAM for Neuromorphic Systems *IEEE International Electron Devices Meeting*

[9] Zhao M, Gao B, Xi Y, Xu F, Wu H and Qian H 2019 *IEEE Journal of the Electron Devices Society* **7** 1239

[10] Zhao M *et al* 2019 Impact of Switching Window on Endurance Degradation in Analog RRAM *Electron Devices Technology and Manufacturing Conference*

[11] Xiang Y, Huang P, Zhao Y, Zhao M, Gao B, Wu H, Qian H, Liu X and Kang J 2019 *IEEE Transactions on Electron Devices* **66** 4517

[12] Zhao M *et al* 2017 Investigation of statistical retention of filamentary analog RRAM for neuromophic computing *IEEE International Electron Devices Meeting*

[13] Grossi A *et al* 2019 *IEEE Transactions on Electron Devices* **66** 1281

[14] Sharma Y, Misra P and Katiyar R S 2014 *Journal of Applied Physics* **116** 084505

[15] Chen Y Y *et al* 2012 *IEEE Transactions on Electron Devices* **59** 3243

[16] Alfaro Robayo D, Sassine G, Rafhay Q, Ghibaudo G, Molas G and Nowak E 2019 *IEEE Transactions on Electron Devices* **66** 3318

[17] Huang P, Xiang Y C, Zhao Y D, Liu C, Gao B, Wu H Q, Qian H, Liu X Y and Kang J F 2018 Analytic Model for Statistical State Instability and Retention Behaviors of Filamentary Analog RRAM Array and Its Applications in Design of Neural Network *IEEE International Electron Devices Meeting*

[18] Chen P and Yu S 2018 Reliability Perspective of Resistive Synaptic Devices on the Neuromorphic System Performance *IEEE International Reliability Physics Symposium*

[19] Peng X, Huang S, Luo Y, Sun X and Yu S 2019 DNN+NeuroSim: An End-to-End Benchmarking Framework for Compute-in-Memory Accelerators with Versatile Device Technologies *IEEE International Electron Devices Meeting*

[20] Sun X and Yu S 2019 *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* **9** 570

[21] Nail C *et al* 2016 Understanding RRAM Endurance, Retention and Window Margin Trade-off using Experimental Results and Simulations *IEEE International Electron Devices Meeting*

[22] Wei Z *et al* 2012 Retention Model for High-Density ReRAM *IEEE International Memory Workshop*

[23] Balatti S, Ambrogio S, Wang Z, Sills S, Calderoni A, Ramaswamy N and Ielmini D 2015 *IEEE Transactions on Electron Devices* **62** 3365

[24] Ambrosi E, Bricalli A, Laudato M and Ielmini D 2019 *Faraday Discuss* **213** 87

[25] Fantini A, Goux L, Redolfi A, Degraeve R, Kar G, Chen Y Y and Jurczak M 2014 Lateral and Vertical Scaling Impact on Statistical Performances and Reliability of 10nm TiN/Hf(Al)O/Hf/TiN RRAM Devices *Symposium on VLSI Technology*

[26] Hangbing L *et al* 2015 *Scientific Reports* **5** 7764 ISSN 2045-2322

[27] Park J, Jo M, Bourim E M, Yoon J, Seong D, Lee J, Lee W and Hwang H 2010 *IEEE Electron Device Letters* **31** 485–487

[28] Traoré B, Blaise P, Vianello E, Grampeix H, Jeannot S, Perniola L, De Salvo B and Nishi Y 2015 *IEEE Transactions on Electron Devices* **62** 4029–4036

[29] Ambrogio S, Balatti S, Wang Z Q, Chen Y S, Lee H Y, Chen F T and Ielmini D 2015 Data Retention Statistics and Modelling in HfO 2 Resistive Switching Memories *IEEE International Reliability Physics Symposium*

[30] Ielmini D, Nardi F, Cagli C and Lacaita A L 2010 *IEEE Electron Device Letters* **31** 353–355

[31] Cabout T *et al* 2013 Temperature Impact (up to 200 °C) on Performance and Reliability of HfO2-based RRAMs *IEEE International Memory Workshop*

[32] Lammie C, Xiang W, Linares-Barranco B and Azghadi M R 2020 *ArXiv* **abs/2004.10971**

[33] Kvatinsky S, Ramadan M, Friedman E G and Kolodny A 2015 *IEEE Transactions on Circuits and Systems II: Express Briefs* **62** 786

[34] Wang Q, Wang X, Lee S H, Meng F and Lu W D 2019 A Deep Neural Network Accelerator Based on Tiled RRAM Architecture *IEEE International Electron Devices*

Meeting (IEDM) (San Francisco, CA.)
[35] Shim W, Luo Y, Seo J and Yu S 2020 Impact of Read Disturb on Multilevel RRAM based Inference Engine: Experiments and Model Prediction *IEEE International Reliability Physics Symposium (IRPS)* (Dallas, TX)
[36] Mehonic A, Joksas D, Ng W H, Buckwell M and Kenyon A J 2019 *Frontiers in Neuroscience* **13** 593