This is the author-created version of the following work:


Access to this file is available from:

https://researchonline.jcu.edu.au/68688/

© 2021 IOP Publishing Ltd. Accepted Author version: CC BY-NC-ND 4.0

Please refer to the original source for the final version of this work:

https://doi.org/10.1088/1361-2125/abf29d
Empirical Metal-Oxide RRAM Device Endurance and Retention Model for Deep Learning Simulations

Corey Lammie¹, Mostafa Rahimi Azghadi¹, and Daniele Ielmini²

¹College of Science and Engineering, James Cook University, Townsville, QLD 4814, Australia
²Politecnico 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.

This is the Accepted Manuscript version of an article accepted for publication in Semiconductor Science and Technology. IOP Publishing Ltd is not responsible for any errors or omissions in this version of the manuscript or any version derived from it. The Version of Record is available online at https://doi.org/10.1088/1361-6641/abf29d.
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 [8-13, 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 principles, 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, $V_{\text{STOP}}$; and

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 $\ddagger$, it is highly integrable $\ddagger$, 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.

![Figure 1](image-url)  
**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.

| Table 1: Comparison of RRAM endurance and retention models. $^*$Models are defined independently. |
|---|---|---|
| **Model** | **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 | ✓ |

1. A conductive filament within metal-oxide RRAM devices results in low resistive states, whereas its partial destruction increases the resistivity to high resistive states.

2. Defects are gradually created within the conductive filament, which cause endurance losses.

3. Oxygen ions return to the previous thermal equilibrium state during the baking process, which causes retention losses.

4. The proposed model accurately models device endurance and retention behavior, both before and after the conductance window begins to collapse.
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

\[
R(x, s, T) = \begin{cases} 
R_0 + 10^{p_1 s + p_2 T_c} \log(x) + \log(R_0) & x \leq e_{th} \\
\text{otherwise},
\end{cases}
\]

and the sudden convergence of states is modeled as

\[
R(x, s, T) = \begin{cases} 
R_0 & x \leq e_{th} \\
R_e & \text{otherwise},
\end{cases}
\]
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

\[
e_{th} = p_0 e^{p_1 s + p_2 T_c}.
\]

The temperature constant, \( T_c \), is expressed as

\[
T_c = \min\left(\frac{T_{th}}{T}, 1\right),
\]

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, \( p \). Here, \( R_{x} \), the collapsed resistive state to which \( R_{ON} \) and \( R_{OFF} \) converge, is bounded to the range \([R_{ON}, R_{OFF}]\) [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, \( p_1 \) is used to model 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, \( \Delta H \), 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 \( V_{\text{stop}} \), the most negative voltage in the negative voltage sweep during the SET-RESET cycle [23].

\[
p_0 = \frac{10^{K(1-(2V_{\text{stop}}-1)^2)}}{e^{p_1 s + p_2 T_c}}
\]

\( K \) is used to modulate the amplitude, and \( V_{\text{stop}} \) is mapped to \( V_{\text{stop}} \in [0, 1] \). Fig. 4 demonstrates the inclusion of \( V_{\text{stop}} \) 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 \( V_{\text{stop}} \) value, i.e., the \( V_{\text{stop}} \) value for a given device that maximizes \( e_{th} \). Given sufficient experimental data observing the relationship between \( V_{\text{stop}} \) and \( e_{th} \), the mapping bounds of \( V_{\text{stop}} \) 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 \( V_{\text{stop}} \) is mutable; and
2. The impacts of the compliance current, \( I_c \), and the maximum set voltage magnitude are considered negligible [22]; and
3. Resistance states converge to \( R_{x} \) 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 is 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, \( \tau_{R} \), and the conductive filament diameter, \( \phi \), 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] \( V_{\text{stop}} \) dependence is validated in TiN/HfO\textsubscript{x}/TiN devices [23].

![Figure 4: The window function used to model \( V_{\text{stop}} \) dependence. Experimental data is extracted from TiN/HfO\textsubscript{x}/TiN devices [23].](image)
Figure 5: An overview of the mapping process of Linear (dense) and Conv2d (convolutional) layers onto a $3 \times 2$ 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.

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/HfO$_x$/TiN, TiN/ETML/HfO$_x$/TiN, and TiN/Hf$_x$/TiN RRAM devices using the VTEAM model 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

Algorithm 1 Adopted simulation methodology.

1. Map Network Parameters
   for each convolutional and linear layer do
     $W_{\text{max}} = \text{descending}_{\text{order}}(\text{abs}(W)\left[\text{size}(W)\right])$
     $W_{\text{min}} = W_{\text{max}} / (R_{\text{OFF}} / R_{\text{ON}})$
     $W_{\text{pos}} = W_{\text{pos}} \left[ W_{\text{pos}} \geq 0 \right]$, \( W_{\text{neg}} = W_{\text{neg}} \left[ W_{\text{neg}} < 0 \right] \)
   for each device, \( R_{\text{pos}}[i,j], R_{\text{neg}}[i,j] \) in \( W_{\text{pos}} \)
     \( R_{\text{pos}}[i,j] = \frac{\left( R_{\text{ON}} - R_{\text{OFF}} \right) \left| W_{\text{pos}}[i,j] - W_{\text{min}} \right|}{W_{\text{max}} - W_{\text{min}}} + R_{\text{OFF}} \)
     \( R_{\text{neg}}[i,j] = \frac{\left( R_{\text{ON}} - R_{\text{OFF}} \right) \left| W_{\text{neg}}[i,j] - W_{\text{max}} \right|}{W_{\text{max}} - W_{\text{min}}} + R_{\text{OFF}} \)
   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
     \( \beta_0 \) 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, \( R_{\text{pos}}[i,j], R_{\text{neg}}[i,j] \) in \( W_{\text{pos}}, W_{\text{neg}} \)
         \( R_{\text{pos}}[i,j], R_{\text{neg}}[i,j] \) in \( W_{\text{pos}}, W_{\text{neg}} \)
       end for
       determine the test set accuracy for the given \( x \) value
     end for

† https://github.com/coreylammie/SST-Reproducibility
circuity 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 x 128 devices, representing positive and negative parameters, respectively. Unfolded inputs are scaled and encoded using voltage signals between $\pm 0.3 V$ [35].

After network parameters are mapped, to tune each memristive layer, random inputs of variable size that are sampled from uniform distributions between $\pm 1.0$ 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 exemplary 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 $\max(V_{stop})$ 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/HfO$_x$/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/No/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 DNNs, 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.

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
[18] Chen P and Yu S 2018 Reliability Perspective of Resistive Synaptic Devices on the Neuronomorphic System Performance IEEE International Reliability Physics Symposium
[22] Wei Z et al 2012 Retention Model for High-Density ReRAM IEEE International Memory Workshop
[26] Hangxing L et al 2015 Scientific Reports 5 7764 ISSN 2045-2322
[31] Cabot T et al 2013 Temperature Impact (up to 200 °C) on Performance and Reliability of HfO2-based RRAMs IEEE International Memory Workshop