Statistical temperature coefficient distribution in analog RRAM array: impact on neuromorphic system and mitigation method

Heng Xu¹, Yue Sun², Yangyang Zhu², Xiaohu Wang²,* and Guoxuan Qin¹,*

¹ School of Microelectronics, Tianjin University, Tianjin 300072, People’s Republic of China
² School of Mechanical Engineering, Dalian University of Technology, Dalian, Liaoning 116024, People’s Republic of China

E-mail: wangxiaohu@dlut.edu.cn and gqin@tju.edu.cn

Received 10 June 2021, revised 18 September 2021
Accepted for publication 23 September 2021
Published 8 October 2021

Abstract
Emerging analog resistive random access memory (RRAM) based on HfOₓ is an attractive device for non-von Neumann neuromorphic computing systems. The differences in temperature dependent conductance drift among cells hamper computing accuracy, characterized by the statistical distribution of temperature coefficient ($T_{\alpha}$). A compact model was presented in order to investigate the statistical distribution of $T_{\alpha}$ under different resistance states. Based on this model, the physical mechanism of thermal instability of cells with a positive $T_{\alpha}$ was elucidated. Furthermore, this model can also effectively evaluate the impact of conductance distribution of different levels under various temperatures in artificial neural networks. A current compensation scheme and hybrid optimization method were proposed to reduce the impact of the distribution of $T_{\alpha}$. The simulation results showed that recognition accuracy was improved from 79.8% to 91.3% for the application of Modified National Institute of Standards and Technology handwriting digits classification with a two-layer perceptron at 400 K after adopting the proposed optimization method.

Keywords: RRAM, temperature coefficient, neuromorphic computing, array

(Some figures may appear in color only in the online journal)

1. Introduction

In recent years, brain inspired neuromorphic computing has demonstrated promising characteristics in terms of computing efficiency and energy consumption compared with conventional von-Neumann architecture [1]. Non-volatile memory represented by resistive random access memory (RRAM) has been extensively studied as synaptic elements in brain-inspired computing, which is mainly used to build a high-speed and low-power neuromorphic computing system [2–5].

HfOₓ-based RRAM has recently been widely utilized in neuromorphic hardware systems due to its fast switching speed, low power consumption, high reliability, excellent analog switching properties and great compatibility with the mainstream CMOS fabrication process [6, 7].

However, the calculation accuracy of the neuromorphic computing system based on the memristor is restricted by the non-ideal effect of the memristor, such as endurance and retention degradation, read/write noises, the intrinsic non-linearity of conductance update [8]. Therefore, the impact of device-level and array-level non-ideal effects on the accuracy of neuromorphic computing systems has been widely studied [9–12]. Meanwhile, the thermal stability of analog...
or multi-stage RRAM is crucial for neuromorphic terminal devices that work in a wide temperature range. Different from the well-studied binary RRAM used for storage, the resistance of the analog memristor acting as synaptic weights in neuromorphic computing system will overlap significantly when temperature changes and thus which will cause a decrease in inference accuracy [13].

Low-temperature characteristics and the impact of temperature on the reset operation of HfO₂-based RRAM have been studied [14, 15]. Meanwhile, the impact of operating temperature on the read/write reliability has been examined [16]. The temperature dependent statistical model has been proposed to predict the transport dependence in the temperature range below 300 K [17]. The temperature coefficient (\(T_\alpha\)) of resistance is one of the significant indicators for evaluating thermal stability. In contrast to the high resistance state (HRS), the conduction mechanism of HfO₂-based RRAM in the low resistance state (LRS) is considered to be metallic with a positive \(T_\alpha\). The study found that different cells with the same resistance range in the RRAM array will show either positive or negative \(T_\alpha\) [18]. Therefore, the synaptic weight in the neuromorphic system will further deviate from the initial value with the temperature changes.

In this work, a compact model is proposed to predict the statistical conductance evolution with temperature changes, which can explain the physical origin of the unstable properties of cells with positive \(T_\alpha\). The compact model demonstrates the excellent consistency between experimental data and simulations. Meanwhile, the impact of the statistical distribution of the \(T_\alpha\) at the array level on the accuracy of neuromorphic computing systems is effectively evaluated. By selecting the conductance mapping range and the current compensation scheme, the calculation accuracy of the system can be effectively improved from 79.8% to 89.6% at 400 K. The effect of this method is not obvious when the temperature is lower than 350 K. Therefore, maintaining excellent heat dissipation should be given priority to improve the calculation accuracy for neuromorphic systems with a temperature below 350 K.

2. Experiments

The 1 kbit HfO₂-based RRAM 1T1R array consists of 1024 cells with 128 rows and eight columns and the major fabrication processes of the array are as follows [19, 20]. The RRAM devices are formed on the drain of the transistors by using the following processes. The 8 nm HfO₂ was deposited with the atomic layer deposition method on the TiN bottom electrode. Then, a 60 nm TaO₂ layer was deposited by the physical vapor deposition method. The top electrode is TiN/AI, which was deposited by sputtering and electron beam evaporation, respectively. The array was placed on Cascade Summit 11 000 probe station and connected to probe station with a probe card. Electrical tests and temperature dependent tests were performed with a Nextest array testing system and an ERS SP72 temperature controller, respectively. For device modeling, an atomistic simulation method was used to establish a 2D resistance network in MATLAB to simulate the concentration and distribution of oxygen vacancy (\(V_O\)) in the filament region. The filament region was equivalent to a stochastic distribution of different concentrations of \(V_O\) in a 40 × 32 matrix, which was proportional to the physical dimensions of an actual device. Kirchhoff’s law could solve the electric potential and current distribution in the filament area, and the overall resistance of the network can be further calculated. Similar to Conductive Bridge RAM, filamentary RRAM exhibits negative and positive \(T_\alpha\) corresponding to conduction behavior of semiconductors and metals in high and LRSs, respectively [21]. Therefore, different adjacent atom connection types in the 2D resistance network where \(V_O–V_O, V_O–O^{2-}\) and \(O^2–O^2–\) were regarded as metallic, semiconductor and insulator resistance, respectively. Meanwhile, distinct adjacent atom connection types have corresponding temperature coefficients according to their different conductivity types. \(T_\alpha\) was calculated by fitting the overall resistance of the 2D network at different temperatures. Since the distribution of \(T_\alpha\) should be correlated with statistical results, the \(V_O\) were randomly distributed at a fixed \(V_O\) concentration. The above experimental process was repeated 300 times to simulate the differences of cells in the array.

3. Results and discussion

Figure 1(b) shows the temperature dependent measurement results of two different resistance value patterns of cells in the 1 kbit HfO₂-based RRAM array [18]. The initial resistances of \(R_1\) and \(R_2\) were 28.5 kΩ and 40 kΩ, respectively. The resistance of \(R_{H1}\) and \(R_{H2}\) decreases with increasing temperature exhibiting an explicitly doped semiconducting behavior. However, \(R_{L1}\) and \(R_{L2}\) indicated metallic characteristics or semiconducting behaviors. The relationship between resistance and temperature can be described by the approximation below [22],

\[
R(T) = R_0 \cdot [1 + T_\alpha \cdot (T - T_0)].
\]

Here, \(R_0\) is the resistance at a reference temperature \(T_0\) and \(T_\alpha\) is the temperature coefficient. The statistical distribution results of \(T_\alpha\) versus resistance for HfO₂-based 1 kbit RRAM array were presented in figure 1(b). The conduction mechanism of a minority of low-resistance cells transformed to be metallic. The conductance in RRAM array is simply proportional to the represented weight in neural networks [23]. Conductance fluctuations can reduce the recognition accuracy of the neuromorphic computing system.

Figure 2(a) shows the simulation flow. Different from the conventional strong-filament based RRAM, in multiple-weak-filaments based RRAM, the oxygen vacancies were distributed at a nanoscale region in multiple-weak-filaments based RRAM [24]. The performance of HfO₂-based RRAM was related to the concentration and random distribution of \(V_O\). Simulation results for the statistical distribution of the \(T_\alpha\) is presented in figure 2(b). As the oxygen vacancy concentration increases from 50% to 58%, the \(T_\alpha\) of all cells moved to a positive value, and more low-resistance cells obtained positive \(T_\alpha\). A cell with a larger resistance at the same concentration...
corresponds to a smaller $T_\alpha$. The simulation result consistent trend with the experimental data. For the same resistance state cells, the lower $V_O$ concentration value in the filament region corresponds to a higher $T_\alpha$.

The analog switch capability of RRAM is essential for realizing high-density weight storage of neuromorphic computing. To further study the effect of temperature coefficient on the analog switch capability of RRAM, we simulated the SET process of the RRAM cells using incremental step pulse program scheme. The cell conductance gradually increased with the application of the SET pulse, which stopped after the device conductance exceeded the target value of 60 $\mu$S. Each SET voltage pulse would increase the concentration of $V_O$ in the filament area and redistributes the $V_O$.

As shown in figures 3(a) and (b), for the target conductance value, the SET pulse number of negative $T_\alpha$ cell ($T_\alpha = -0.0014$) is 163, which is significantly larger than that (149) for the positive $t$-cell ($T_\alpha = 0.0012$). The illustration shows the statistical distribution of the number of pulses required for 100 $T_\alpha$ and $100 - T_\alpha$ cells to be set to the target resistance state. The average number of pulses (150) of all $+ T_\alpha$ units is less than the average number of pulses (161) of $- T_\alpha$ units. Meanwhile, compared with a negative $T_\alpha$ cell, the conductance of a positive $T_\alpha$ cell generally had more fluctuations after 120 pulses. Therefore, the analog switching performance of the negative $T_\alpha$ cell is preferable to that of the positive $T_\alpha$ cell. Figures 3(c) and (d) show the simulated current density distribution of the negative $T_\alpha$ cell and the positive $T_\alpha$ cell. It can be seen that multiple weak CFs are formed due to the percolation effect in the negative $T_\alpha$ cell. In contrast, the apparent conductive path is formed in the positive $T_\alpha$ cell, similar to strong-filament based RRAM. An order parameter of $V_O$ can
be used to evaluate the disorder effect of $V_O$ distribution [23], which can be described as:

$$O_V = \frac{2N_{V-V}}{zC_V N}$$

where $N_{V-V}$ is the number of $V_O-V_O$ bond, $C_V$ is $V_O$ concentration, $N$ is the total number of oxygen sites in the filament region, and $z$ is coordinate number of lattice. The disparity of $V_O$ in the filament region increases as $O_V$ decreases. $O_V$ is lower for the cell with negative $T_o$ than the cell with positive $T_o$.

Previous research has shown that the retention property of cells with a positive $T_o$ was much worse than that of cells with a negative value [18]. In order to further reveal the correlation between $T_o$ and retention characteristics, the perturbation process of the RRAM cells was simulated. As shown in figure 4(a), after the cells were disturbed, the $V_O$ randomly hops at adjacent lattice sites, or $V_O$ can hop many times when this process is in analogy to Brownian motion. The final distribution of $V_O$ was more dispersed. The simulation results of the redistribution of $T_o$ are presented in figure 4(b). The $T_o$ of cells at LRS was effectively limited to values below zero, which meant that the number of cells with poor retention properties was reduced. The underlying mechanism of the correlation between $T_o$ and retention properties can be explained as follows. The strong-like filament formed in the filament region of HfO$_2$-based RRAM with TEL is unstable. The $V_O$ distribution in the filamentous region tended to be disordered and scattered, which cause $T_o$ to tend to be negative. Therefore, repeated write-verify and heating after the programming can redistribute oxygen vacancies, which reduces the probability of cells with positive $T_o$.

In order to display the potential of the compact model to evaluate and optimize neural networks, a standard multi-layer perceptron was used as an example to illustrate the influence of the distribution of $T_o$ in the array on neuromorphic computing systems. This $784 \times 100 \times 10$ fully-connected neural network was used to recognize images from Modified National Institute of Standards and Technology (MNIST) database on handwritten digits, as shown in figure 5(a). The activation functions of the hidden layer and the output layer were rectified linear unit (ReLU) [26] and softmax function [4], respectively. The real-valued weights were linearly mapped to the conductance difference of two RRAM cells, which was composed of the corresponding positive and negative weight rows

$$I_j = \sum_{i=1}^{n} G_{ij} \cdot V_i = \sum_{i=1}^{n} \left( G_{ij}^+ - G_{ij}^- \right) \cdot V_i.$$  (3)

where $G_{ij}$ is the conductance of the memory element at array position $(i, j)$. The initial accuracy achieved using software was 97.36% (training with 32-bit single-precision floating-point weights), which was degraded to 94.48% after the quantization using eight-level weights. In order to exclusively investigate $T_o$ impact on neural networks, non-ideal factors, such as quantization, circuit parasitic, and retention degradation caused by temperature changes are ignored. The eight-level weight was mapped in the maximum conductance range (12.5 $\mu$S–100 $\mu$S). Generally, the highest operating temperature for mobile device chips is 345 K, while the highest temperature for computer chips can reach 400 K. When only considering the statistical distribution of $T_o$, the calculation accuracy with $T_o$ decreased from 94.48% to 79.8% for the mean when the temperature was changed from 300 K to 400 K.

The existing array programming methods, such as write-verify and periodic-refreshing are ineffective in eliminating the impacts of $T_o$ distribution on neuromorphic computing systems because the $T_o$, as an intrinsic property of a material, is closely related to the device microstructure. Hence, a simple current compensation method is proposed to offset temperature-induced conductance change and recover the network performance. Firstly, a reasonable conductance mapping interval was selected according to the distribution of $T_o$. The coefficient of variation ($c_v$) is used to evaluate the dispersion of the $T_o$ distribution in three conductance ranges (high: 50 $\mu$S–100 $\mu$S; middle: 25 $\mu$S–50 $\mu$S; and low: 12.5 $\mu$S–25 $\mu$S),

$$c_v = \frac{\sigma}{\mu} \times 100\%$$  (4)

where $\sigma$ is standard deviation, and $\mu$ is mean of $T_o$. The $C_V$ of $T_o$ in the low conductance ranges (5.48%) was less than that in the middle (16.3%) and the high conductance ranges (32.62%). $T_o$ in low conductance ranges were similar, so the cells programmed in such a range can be treated as having the same $T_o$ ($-0.004$ K$^{-1}$). The output current ($I_{L,0}$) of each column at room temperature was stored in an integrated non-volatile register. The compensation current value can be calculated by measuring the operating temperature.
The sensitivity of the ANN neural network structure to the hardware resources than the current compensation scheme and modules and data transmission modules, it will save more hardware resources than the current compensation scheme. Therefore, it is crucial for the neuromorphic computing system to have temperature control. Meanwhile, for multi-layer neural networks implemented by analog interfaces, the errors in the front layer accumulate to the next layer [27]. Thus, it is crucial for the neuromorphic computing system to have temperature control of the front layer or the optimization of the $T_\alpha$ distribution in the front layer to mitigate accuracy loss.

The sensitivity of the ANN neural network structure to the weight drift caused by the $T_\alpha$ was then discussed. The effect of $T_\alpha$ was re-examined by setting the first layer and the second layer to the ideal weight value (the weight does not change with temperature). Figure 6(a) shows the influence of the $T_\alpha$ distribution in different layers on the inference accuracy with a changing temperature. The $T_\alpha$ distribution of the first layer led to a rapid decline in the overall inference accuracy of the neural network with the temperature increasing. The average inference accuracy of 100 experiments at 400 K was 83.7%, which was only an increase of 3.9% compared with the case where both layers were affected by the temperature coefficient (79.8%). When the weight of the first layer was ideal and the $T_\alpha$ distribution was present in the second layer, the accuracy was almost unchanged as the temperature increased and dropped by 1.68 at 400 K. The $T_\alpha$-induced inference accuracy error in the first layer was more sensitive to weight drift than that in the second layer. The computing errors caused by $T_\alpha$ in the first layer will also be further amplified by the nonlinear activation function and incorporated to the latter layer. Meanwhile, for multi-layer neural networks implemented by analog interfaces, the errors in the front layer accumulate to the next layer [27]. Thus, it is crucial for the neuromorphic computing system to have temperature control of the front layer or the optimization of the $T_\alpha$ distribution in the front layer to mitigate accuracy loss.

The error caused by the $T_\alpha$ of the front layer cannot be completely eliminated and the error will propagate down and magnify step-by-step in a multi-layer or more complex neural network. Re-adjusting the weight of the last layer can compensate for the error propagated by the front layer [6]. Hence, a hybrid optimization method that combines conductivity mapping optimization and in situ training was adopted to alleviate the accuracy loss caused by the $T_\alpha$. The weight of the front layer is trained in the software, and the low resistance range (50 $\mu$S–100 $\mu$S) is selected when the weight is mapped to the conductance, so that the temperature coefficient distribution is as dense as possible and tends to zero. As shown in figure 6(b), only optimizing the mapping range of the first layer of the memristor can increase the average inference accuracy from 79.8% to 85.6% at 400 K. Meanwhile, when the accuracy loss was greater than the set threshold with the temperature increasing, only the RRAM in the second layer were trained in situ, i.e. the conductance was updated on-chip. The results show that the average inference accuracy can be restored to 91.3% at 400 K after using the hybrid optimization scheme, which is better than the current compensation scheme. The hybrid optimization provides better universality for different neural network structures, such as multi-layer neural networks or CNN. Although the physical implementation of the hybrid optimization method may also require corresponding memory modules and data transmission modules, it will save more hardware resources than the current compensation scheme and

$$\Delta I_{j,T} = \sum_{i=1}^{n} G_i \cdot V_j - \sum_{i=1}^{n} \frac{G_i}{1 + T_\alpha \cdot \Delta T} \cdot V_j$$

$$\Delta I_{j,T} = \frac{T_\alpha \cdot \Delta T}{1 + T_\alpha \cdot \Delta T} \cdot I_{j,T_0}.$$ (5)

The recovered sum-of-product value ($I_{out,T}$) can be derived by summing the actual output current ($I_{j,T}$) and compensation current ($\Delta I_{j,T}$),

$$I_{out,T} = I_{j,T} + \Delta I_{j,T}.$$ (6)
Figure 6. (a) The statistical distribution of inference accuracy with temperature increasing in two-layer (artificial neural network) ANN with the first layer ideal weight + the second layer weight affected by $T_\alpha$, and the first layer weight affected by $T_\alpha$ + the second layer ideal weight affected by $T_\alpha$. (b) The statistical distribution of inference accuracy with temperature increasing in two-layer ANN with the conductance mapping range optimization and hybrid optimization method.

in situ updates of all conductance at different temperatures. Thus, excellent heat dissipation capacity is essential for future applications of neuromorphic computing.

4. Conclusion

In conclusion, thermal instability based on temperature coefficient ($T_\alpha$) was comprehensively investigated in the HfO$_x$-based RRAM array. A compact model was proposed by a 2D atomistic simulation to study the statistical distribution of $T_\alpha$ on the array level. Based on the simulation of the SET process and perturbation process of the RRAM cells in the array, the physical mechanism of instability of cells with positive $T_\alpha$ was elucidated. A compensation scheme and a hybrid optimization method were proposed in this paper by selecting the appropriate conductance range for the weight-conductance mapping and adding compensation current, which can effectively recover the inference accuracy. With this scheme, the mean MNIST inference accuracy of a multi-layer neural network can be improved remarkably. Our results are crucial for the evaluation and optimization of RRAM-based neuromorphic computing systems.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

We thank Laboratory of Emerging MemOry and Novel computing and School of Integrated Circuits, Tsinghua University for providing HfO$_x$-based 1 kbit RRAM array to be test.

ORCID iDs

Xiaohu Wang  https://orcid.org/0000-0001-6154-2020
Guoxuan Qin  https://orcid.org/0000-0002-6259-8517

References

[1] Sebastian A, Le Gallo M, Khaddam-Aljameh R and Eleftheriou E 2020 Memory devices and applications for in-memory computing Nat. Nanotechnol. 15 529–44
[2] Zhang W, Gao B, Tang J, Yao P, Yu S, Chang M-F, Yoo H-J, Qian H and Wu H 2020 Neuro-inspired computing chips Nat. Electron. 3 371–82
[3] Zidan M A, Strachan J P and Lu W D 2018 The future of electronics based on memristive systems Nat. Electron. 1 22–29
[4] Cai F, Correll J M, Lee S H, Lim Y, Bothra V, Zhang Z, Flynn M P and Lu W D 2019 A fully integrated reprogrammable memristor—CMOS system for efficient multiply—accumulate operations Nat. Electron. 2 290–9
[5] Zhang Y, Huang P, Gao B, Kang J and Wu H 2021 Oxide-based filamentary RRAM for deep learning J. Appl. Phys. 54 083002
[6] Yao P, Wu H, Gao B, Tang J, Zhang Q, Zhang W, Yang J J and Qian H 2020 Fully hardware-implemented memristor convolutional neural network Nature 577 641–6
[7] Liu Q et al 2020 33.2 A fully integrated analog RRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing 2020 IEEE Int. Solid- State Circuits Conf.—(ISSCC) (San Francisco, CA, USA: IEEE) pp 500–2
[8] Zhao M, Gao B, Tang J, Qian H and Wu H 2020 Reliability of analog resistive switching memory for neuromorphic computing Appl. Phys. Rev. 7 011301
[9] Zhao M et al 2017 Investigation of statistical retention of filamentary analog RRAM for neuromorphic computing 2017 IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA, USA: IEEE) pp 39.4.1–4

[10] 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 2018 IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA: IEEE) pp 40.4.1–4

[11] Chen P-Y, Peng X and Yu S 2017 NeuroSim+: an integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures 2017 IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA, USA: IEEE) pp 6.1.1–4

[12] Calixto M, Maldonado D, Miranda E and Roldán J B 2020 Modeling of the temperature effects in filamentary-type resistive switching memories using quantum point-contact theory J. Phys. D: Appl. Phys. 53 295106

[13] Lin Y, Wang C, Lee M, Lee D, Lin Y, Lee F, Lung H, Wang K, Tseng T and Lu C 2019 Performance impacts of analog RRAM non-ideality on neuromorphic computing IEEE Trans. Electron Devices 66 1289–95

[14] Fang R, Chen W, Gao L, Yu W and Yu S 2015 Low-temperature characteristics of HfO$_x$-based resistive random access memory IEEE Electron Device Lett. 36 567–9

[15] Puglisi F M, Qafa A and Pavanello P 2015 Temperature impact on the reset operation in HfO$_x$ RRAM IEEE Electron Device Lett. 36 244–6

[16] Schultz T and Jha R 2019 Operating temperature based vulnerabilities in RRAM 2019 IEEE 62nd Int. Midwest Symp. on Circuits and Systems (MWSCAS) (Dallas, TX, USA: IEEE) pp 464–7

[17] Wang L, Thean A V Y and Liang G 2018 Percolation theory based statistical resistance model for resistive random access memory Appl. Phys. Lett. 112 253505

[18] Wang X H, Wu H, Gao B, Li X, Deng N and Qian H 2018 Thermal stability of HfO$_x$-based resistive memory array: a temperature coefficient study IEEE Electron Device Lett. 39 192–5

[19] Yao P et al 2017 Face classification using electronic synapses Nat. Commun. 8 15199

[20] Huang X et al 2016 HfO$_x$/Al$_2$O$_3$ multilayer for RRAM arrays: a technique to improve tail-bit retention Nanotechnology 27 395201

[21] Guan W, Liu M, Long S, Liu Q and Wang W 2008 On the resistive switching mechanisms of Cu/ZrO$_2$-Cu/Pt Appl. Phys. Lett. 93 223506

[22] Wang Y et al 2010 Investigation of resistive switching in Cu-doped HfO$_x$ thin film for multilevel non-volatile memory applications Nanotechnology 21 045202

[23] Xiao T P, Bennett C H, Feinberg B, Agarwal S and Marinella M J 2020 Analog architectures for neural network acceleration based on non-volatile memory Appl. Phys. Rev. 7 031301

[24] Gao B et al 2017 Modeling disorder effect of the oxygen vacancy distribution in filamentary analog RRAM for neuromorphic computing 2017 IEEE Int. Electron Devices Meeting (IEDM) (San Francisco, CA, USA: IEEE) pp 4.4.1–4

[25] Wu W, Wu H, Gao B, Deng N, Yu S and Qian H 2017 Improving analog switching in HfO$_x$-based resistive memory with a thermal enhanced layer IEEE Electron Device Lett. 38 1019–22

[26] Chang C- C et al 2017 Challenges and opportunities toward online training acceleration using RRAM-based hardware neural network 2017 IEEE Int. Electron Devices Meeting (IEDM) (pp 11.6.1–4

[27] Hu M, Strachan J P, Li Z, Grafals E M, Davila N, Graves C, Lam S, Ge N, Yang J J and Williams R S 2016 Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication Proc. of the 53rd Annual Design Automation Conf. DAC ’16: The 53rd Annual Design Automation Conf. 2016 (Austin Texas: ACM) pp 1–6