This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2020.3002539, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

Robust Deep Reservoir Computing through Reliable Memristor with Improved Heat Dissipation Capability

Hongyu An, *Student Member, IEEE*, Mohammad Shah Al-Mamun, *Student Member, IEEE*, Marius K. Orlowski, *Fellow, IEEE,* Lingjia Liu, *Senior Member, IEEE,* Yang Yi, *Senior Member, IEEE*

***Abstract*— Deep Neural Networks (DNNs), a brain-inspired learning methodology, requires tremendous data for training before performing inference tasks. The recent studies demonstrate a strong positive correlation between the inference accuracy and the size of the DNNs and datasets, which leads to an inevitable demand for large DNNs. However, conventional memory techniques are not adequate to deal with the drastic growth of dataset and neural network size. Recently, a resistive memristor has been widely considered as the next generation memory device owing to its high density and low power consumption. Nevertheless, its high switching resistance variations (cycle-to-cycle) restrict its feasibility in deep learning. In this work, a novel memristor configuration with the enhanced heat dissipation feature is fabricated and evaluated to address this challenge. Our experimental results demonstrate our memristor reduces the resistance variation by ~ 30% and the inference accuracy increases correspondingly in a similar range. The accuracy increment is evaluated by our Deep Delay-feed-back (Deep-DFR) reservoir computing model. The design area, power consumption, and latency are reduced by ~48%, ~42%, and ~67%, respectively, compared to the conventional SRAM memory technique (6T). The performance of our memristor is improved at various degrees (~13%-73%) compared to the state-of-the-art memristors.**

***Index Terms*—Memristor, Reservoir Computing, Artificial Neural Networks, Deep delay-feed-back Reservoir Computing**

I. Introduction

Deep Neural Networks (DNNs) inspired by the high-degree structure of neural networks in mammalian brains have accomplished remarkable success in many applications, such as image recognition , natural language processing , machine neural translation [1], etc. A pristine DNN with random synaptic weights has no remarkable capability until its weights are trained by tremendous data. The larger sizes of the datasets and the neural networks lead to a higher inference accuracy [2, 3]. Thereby, the demand for excessively large datasets and neural networks is becoming inevitable. As illustrated in Figure 1, the size of datasets is almost linearly increasing over the years, while the neural networks double their size roughly every two years [3, 4]. Accompanying the growth of the scale of hypermeters, the capacity of the GPU memory has only increased by a factor of three [2]. Hence, there is an urgent need for novel and reliable devices with higher capacity and lower power consumption, fulfilling the tremendous data storage demand for deep learning.

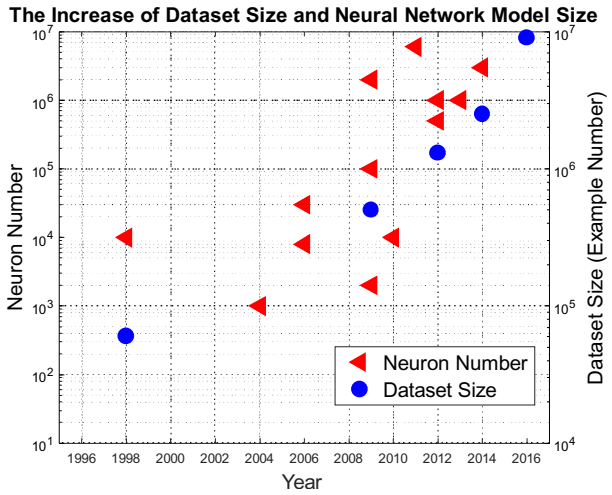


Figure 1: Increase trend of datasets and DNNs sizes [3]

Nowadays, memristors are widely considered as one of the most promising candidates for next-generation memory because of its high density and low power consumption [6]. However, its wide distribution of resistance variation restricts its feasibility in deep learning as weight storing devices [5, 6], since the weight variation significantly reduces the inference accuracy [6-11]. Several methods involving circuit and algorithm optimizations have been proposed to mitigate this shortcoming. However, these methods entail inevitable drawbacks, like the large latency and circuit design overhead [12-14].

In this work, we study the switching mechanism of the memristor and reveal the heat accumulated in the cell during the switching leads to a substantial metal atom diffusion effect. The metallic atoms diffusion at the tip ends of the conductive filaments (CFs) influences the gap size among of the filament in the off-regime when the filaments are ruptured [10]. As a result, the resistance variation increases significantly when heat is accumulated interiorly [11, 15, 16]. In order to mitigate the resistance variation, we designed and fabricated a novel configuration of a memristor with an additional heat dissipation layer integrated into the cell's electrodes, which alleviates the heat-related switching variation by more than 30% (Table I). Unlike using low thermal conductivity material for subduing heat transfer between layers [17], our approach dissipates the accumulated heat both on the metal and insulator layers. The candidates of the heat dissipation layer need to satisfy several

This work was supported by the National Science Foundation under Grant NSF 1750450 and NSF 1731928.

Hongyu An, Mohammad Shah Al-Mamun, Marius Orlowski, Lingjia Liu, and Yang Yi are with the Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: hongyu51@vt.edu, samamun@vt.edu, marius@vt.edu, ljliu@vt.edu, cindy_yangyi@vt.edu)

0278-0070 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on January 04,2021 at 19:52:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2020.3002539, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

requirements, such as high thermal conductivity, low cost, fabrication compatibility, electrochemistry stability at high temperature, etc. Several materials (Rh, Cr, Pt, Ti, Cu) have been tested for heat dissipation efficiency. It turned out that the Ti glue layer used for the adhesion of the inert electrode had to be supplanted by Cr with the most thermal conductivity to render the Joules heating effects less severe.

Furthermore, an experimentally verified memristor model capturing the electrical characteristics has been built. This memristor model is incorporated in our deep delay-feed-back reservoir computing (Deep-DFR) model for evaluation. The Deep-DFR is established by the system-level simulation platforms comprising PyTorch and *NeuroSIM* [9]. The parameters of our memristors in NeuroSIM are extracted from the measurement data (Figure 4). Through our Deep-DFR model, the impact of reducing the switching variations of the memristor on a deep learning system is analyzed. The simulation results demonstrate that the accuracy has been increased by ~30% accompanying the reduction of the resistance variation of the memristor (TABLE I). In order to eliminate the interference from other parameters of memristors and reveal the cause-and-effect relationship between resistance variation (cycle-to-cycle) and inference accuracy, we keep other nonideal parameters of memristors constant in this work.

The accuracy improvement, power consumption, design area, and latency reduction are evaluated with CIFAR-10 and CIFAR-100 datasets (Figure 9).

Our contributions can be summarized as follows:

- A novel memristive device configuration with higher immunity to degradation induced by thermal effects has been fabricated and evaluated. The experiment results demonstrate a ~30% reduction in switching variation (TABLE I);
- The competent material for heat dissipation layer of our new memristor configuration is determined (TABLE I);
- The accuracy improvement (~30%) on classification tasks is demonstrated through our Deep-DFR model, which deploys our memristor model;
- The hardware performance improvement, e.g., power efficiency and design area reduction, is evaluated and analyzed through a co-simulation paradigm with PyTorch and the macro-circuit simulator *NeuroSIM* [9].

This paper is organized as follows, Section II introduces our memristor fabrication and modeling methodology, Section III presents the hardware performance evaluation method using our memristor model, Section IV summarizes the conclusions.

II. RELIABLE MEMRISTOR DESIGN AND MODELING

As one of the most promising candidates of next-generation memory, memristive devices suffer a critical issue of low reliability, which diminishes its practicability for massive deployment [5, 6]. The low reliability of a memristor stems from the high variation on its on-state resistance (R_{on}) value [11]. Through the comprehensive study of the switching mechanism of a memristor [18, 19], we have discovered that the heat-related metal atom diffusion of conductive filaments (CFs) increases the resistive switching variation [20]. In order to address this issue, we designed and fabricated a novel configuration of a memristor, which can effectively mitigate the heat-related resistive switching variation.

A memristor is typically fabricated using a metallic oxide layer as a solid electrolyte sandwiched between an oxidizable active anode electrode and an inert cathode electrode. As illustrated in Figure 2, there are four resistively switching phases of a memristor. Initially, the atomic structure of the metallic oxide layer is intact. At this stage, the bonding between oxygen ions and metal atoms of the metallic oxide is strong. However, under the high electric field established by the applied voltage to the cell's electrodes, the oxygen ions in the metallic oxide could be dislodged from the constraint of the bonding force and migrate to one of the terminals of the memristor. Consequently, the removal of oxygen ions leaves the oxygen vacancies behind leading to a build-up of conductive filament connecting the two electrodes. In another mode, the atoms of the active electrode are ionized and under the applied electric field migrate to the inert electrode where they are stopped and electrically reduced. Over time the active electrode metal atoms pile up on each other leading to a formation of metallic filament connecting the two electrodes. When this happens, the cell is in an on-state characterized by an on-resistance R_{on}. Otherwise, the cell is in the off-state characterized by the off-resistance R_{off}. The ratio between R_{off} and R_{on} is large and exceeds in many cases 10^3. The switching process of the resistance from R_{off} to R_{on} is referred to as a set process. In contrast, the transition from R_{on} to R_{off} is called the reset process.

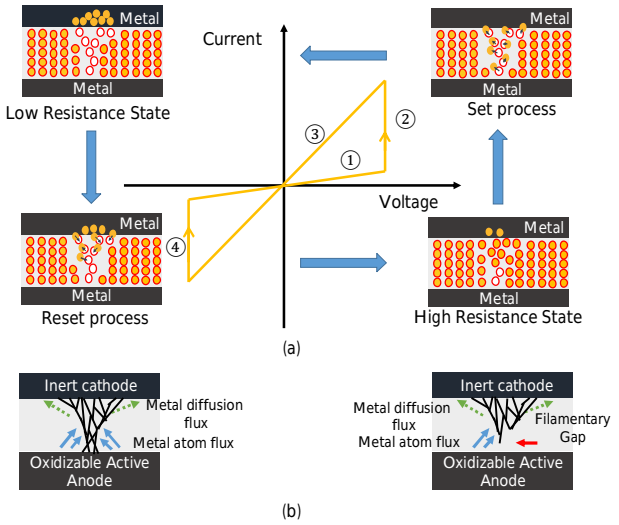


Figure 2: (a) Four typical switching phases of a memristor; (b) Formation mechanism of conductive filaments. The variation of the on-state resistance of a memristor results from a competition between the constructive metal atom flux and destructive metal atoms diffusion flux [11, 18].

As illustrated in Figure 2 (a), the switching capability of memristors attributes to the construction and rupture of the conductive filaments. The shape and the size of the filaments could significantly influence the switching characteristic of a memristor.

0278-0070 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on January 04,2021 at 19:52:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2020.3002539, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

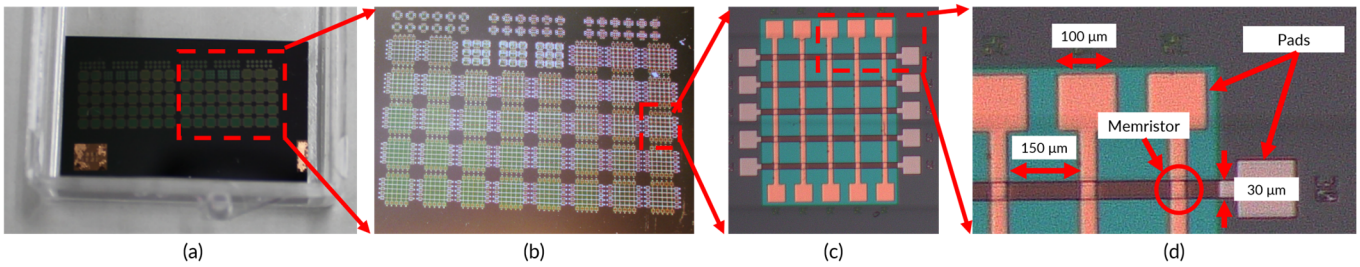


Figure 3: Our fabricated memristor die: (a) The overview of our memristor die; (b) The zoom-in view of our memristor; (c) The five by five crossbar structure of our memristor; (d) The memristor located at the cross-point of the crossbar.

TABLE I: COMPARISON OF THE MEMRISTOR RESISTANCE SWITCHING VARIATION

Device	Size (nm)	Thermal Conductivity	R_{on} (I_{cc} = 5 uA)	Target Value[1]	Variation[2]	R_{on} (I_{cc} = 50 uA)	Target Value	Variation
Cu/TaOx/Rh/Cr	150/25/50/20	Rh:150 Cr: 94	2.5 ± 0.1 KΩ	2.4 KΩ	~4 %	500 ± 5 Ω	500 Ω	~1 %
Cu/TaOx/Rh/Ti	150/25/50/20	Rh:150 Ti: 20	2.3 ± 0.12 KΩ	2.4 KΩ	~5 %	225 – 750 Ω	500 Ω	~35 %
Cu/TaOx/Pt/Cr	150/25/50/20	Pt: 72 Cr: 94	2.1 ± 0.1 KΩ	2.1 KΩ	~4.7 %	331 – 1000 Ω	400 Ω	~33.4 %
Cu/TaOx/Pt/Ti	150/25/50/20	Pt: 72 Cr: 20	2.1 ± 0.9 KΩ	2.1 KΩ	~42.8 %	230 – 1000 Ω	400 Ω	~61.5 %

[1] The target value of the resistance is estimated by the Eq. (6)
[2] The variation is cycle-to-cycle variation that is measured by percent deviation.

During the set and reset switching processes, the considerable current flows through the CFs generally leads to a significant Joules heat dissipation. The temperature of the memristor cell is governed by the Joules heating and the rate of heat removal, which is determined by the thermal conductivity of the surrounding metallic oxide and the thermal conductivities of the electrodes. If the surrounding metallic oxide or the two electrodes cannot dissipate the heat fast enough, the temperature of the filament is bound to increase. Eventually, the high temperature of the CFs triggers a substantial metal diffusion. The metallic atoms of the filament, particularly at the tip of the cone-shaped CFs, diffuse out of the CFs consequently determining the size in the filament [10]. Macroscopically, the on-state resistance variation increases significantly [11, 15, 16, 21]. This phenomenon is even more severe during the rupturing process as the reset is dominated by a thermal dissolution effect [20]. When current flows through the memristive cell, Joule heat is deposited in the conductive filament. As a result, the temperature in the narrowest part (highest resistance) of the filament can reach 1000 ℃ [22, 23]. Such a high temperature triggers Cu atom diffusion from the constriction of filaments.

In order to address this issue, we proposed and investigated a solution of adding an extra metallic layer for facilitating heat dissipation. The copper (Cu) is selected as an oxidizable active anode due to its medium activation energy-yielding ions readily [24] Cu ←→ Cu++e-. The rhodium (Rh) is used for inert cathode since it is compatible with the back-end-of-line (BEOL) integration technique and potentially can be integrated on the top of the metal-oxide-semiconductor field-effect transistors (MOSFETs) for a three-dimensional structure [25]. Furthermore, the Rh-Cu material configuration demonstrates a negligible solid solubility between two elements, rendering Rh an ideal inert electrode for Cu ions (Cu+). In addition, the Rh is 45 times less expensive than Pt with similar characteristics [20].

The oxygen-deficient tantalum oxide (TaOx) is used as the metallic oxide. In this work, the memristor Cu/TaOx/Pt is used as a benchmark device. Our memristive devices have been fabricated in a crossbar configuration on a thermally oxidized silicon wafer. The metal electrodes and solid electrolytes are deposited through e-beam evaporation. The TaOx layer was deposited by evaporating the Ta_2O_5 pellets with no oxygen injection at the evaporation chamber. A thin Ti layer was added between Pt and SiO_2 to improve the adhesion of Pt to the substrate. All the layers (Cu, TaO_x, Pt) are deposited by e-beam PVD in a Kurt Lesker PVD-250 chamber. The fabricated memristor die and the detailed geometry are illustrated in Figure 3. The range of the high resistance state (HRS) is ~1-900 MΩ, yielding a ratio of $R_{off}/R_{on} \approx 10^3 - 10^7$, which effectively avoids the negative effect caused by the sneak path.

The reliability of the memristive devices with different inert cathodes is evaluated by the variation of their on-state resistance. The testing results are summarized in TABLE I. In TABLE I, the cycle-to-cycle variation is measured by percent deviation. The precise temperature control is not practical in real measurement setups. Thus, we distinguish different temperatures (high and low) by applying different compliance currents during the set operation; they are I_{cc} = 5 uA and 50 uA respectively. The heat generated by the different currents, assuming constant current in the time interval t, is governed by:

$$w = I^2 R_{on} t \quad (1)$$

TABLE I demonstrates that the memristive device exhibit a higher spread of on-state resistance (R_{on}) values with higher temperatures (larger compliance current). For example, the on-state resistances of the Rh/Ti configuration are at the range of 225 Ω to 750 Ω for I_{cc} = 50 uA. This instability phenomenon comes from the competition between the constructive Cu^+ electro-migration flux and the destructive Cu diffusion flux, illustrated in Figure 2(b). Our measurements demonstrate an

0278-0070 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on January 04,2021 at 19:52:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2020.3002539, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

effective metal dissipation layer (Cr) could effectively suppress the heat-related metal atom diffusion phenomenon, resulting in a significant reduction of switching variation (by ~30%).

The measurement is performed by applying a positive voltage to the electrode of the device and the voltage is swept at a constant voltage ramp rate (0.2V/s). Initially the value of current remains small until the set voltage of the memristor is reached. The current switching is caused by the conductive filaments (CFs) formation when the applied voltage exceeds the set voltage of the memristor. The measurement usually performed more than 100 times. The variation is measured by the percent deviation from average, which shows the average percentage that a data point differs from the mean value.

The endurance of the devices depends on the compliance current (I_{cc}). For the I_{cc} is at the range of 10 μA and 5mA, the device can be switched more than 150 times. For smaller compliance current, like 1 μA, the endurance of our memristor device can be more than 1000 times switching. During the measurement, no incorrect switching of the unselect and adjacent memristor cells was detected which potentially caused by the sneak path issue. The high ratio of off-state and on-state resistances of our memristor device (more than 10^3) effectively avoids the negative impact of sneak path issue.

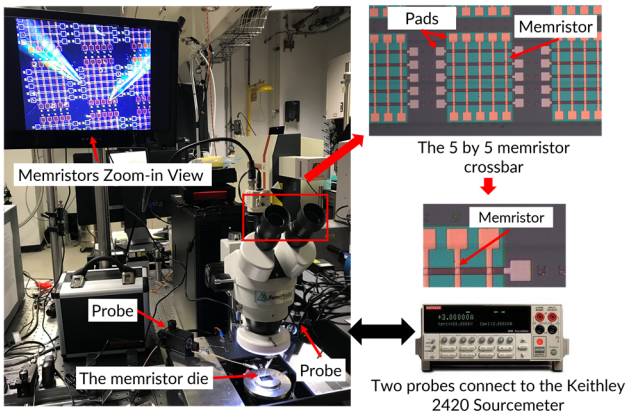


Figure 4: The testing setup of our memristor

Furthermore, to analyze the effect of resistance variation reduction of our memristor on deep learning at a system level, a corresponding Verilog-A memristor model is built upon the filament growing method [26]. In the set process, the w, and x are growing under the stimulus voltage by the following equations:

$$I_{hop} = I_0 (\pi w^2/4) exp(-x/x_T) sinh(V_{gap}/V_T), \quad (2)$$

$$I_{CF} = \frac{\pi w^2 V_{CF}}{4\rho(x_0 - x)}, \quad (3)$$

where x_0 is the initial value of gap distance, x_T and V_T are the characteristic length and voltage in hopping. V_{gap} and V_{CF} are the voltage over the gap region and conductive filament region, respectively. W denotes the Joules heat dissipated in the filament. In the reset process, the w, and x are growing under the stimulus voltage by the following equations:

$$dx/dt = af\ exp(-(E_a - \alpha_a ZeE)/k_B T), \quad (4)$$

$$dw/dt = \left(\Delta w + \frac{\Delta w^2}{2w}\right) f exp\left(-\frac{E_a - \alpha_a ZeE}{k_B T}\right). \quad (5)$$

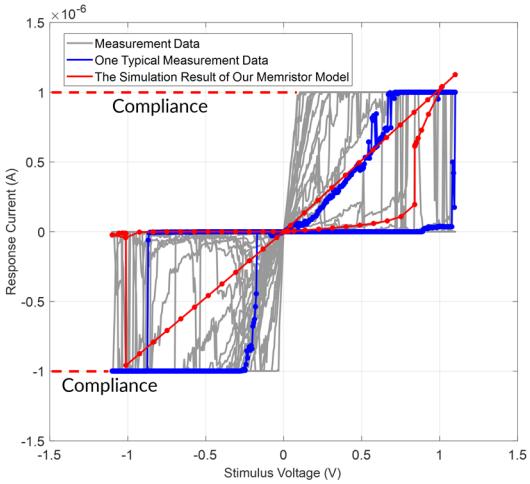


Figure 5: V-I switching characteristics of our memristor (Cu/TaOx/Rh/Cr): The gray lines represent the measurement data, the blue line shows one typical measurement data, and the red line depicts our memristor. Note: the compliance current is 1uA in this case.

Figure 5 illustrates the V-I characteristic curve comparison of our memristor model and the measurement data of our memristors. As depicted in Figure 5, the resistance of the memristor model switches from ~1 MΩ to ~940 MΩ at V_{set} ~0.8V, which matches the measurement data. The sudden current cut-off at ± 1 μA in Figure 5 comes from the compliance current setting. The inconsistency of on-state resistance in Figure 5 and Table 1 comes from the different compliance current [6]. The relationship between R_{on} (low resistance state) and compliment current can be estimated by the equation:

$$R_{on} = \frac{K}{I_{cc}^n}, \quad (6)$$

where n and K are fitting parameters and I_{cc} is the compliance current [20]. Equation (6) indicates the negative correlation between the compliance current and R_{on}.

III. Performance Evaluation of the Memristor on Deep Delay Feedback Reservoir Computing

The emerging Deep-DFR demonstrates a strong capability of processing spatiotemporal data due to its recurrent loop and multiple layer structure [27, 28]. This specific structure allows the system to have more remarkable performance compared to other conventional reservoir computing system. Deep-DFR models demonstrate more than 50% better performance than the typical leaky echo state network (ESN) model [29]. Furthermore, the Delay Feedback Reservoir (DFR) has a simplified structure, which merely consists of one nonlinear neuron in the reservoir [30, 31]. On the contrary, the traditional reservoir system requires numerous nonlinear neurons that demand more hardware resources increasing the hardware design challenge.

0278-0070 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on January 04,2021 at 19:52:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2020.3002539, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

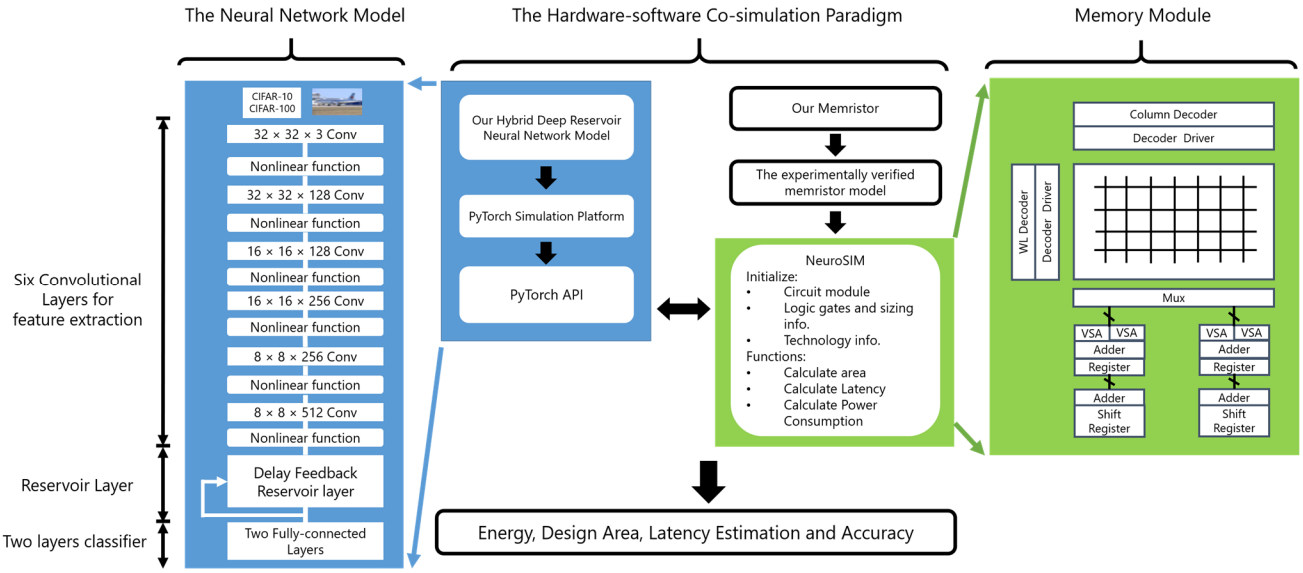


Figure 6: The diagram of our hardware-software co-simulation paradigm with NeuroSIM and PyTorch.

In this work, our Deep-DFR model (Figure 6) is used for evaluating the impact of resistance variation reduction (cycle-to-cycle) of our memristor on inference accuracy. In order to focus on studying the cause-and-effect between the resistance variation and the inference accuracy, other nonideal parameters of memristors that may influence the inference accuracy, e.g. device-to-device variation, are excluded (keeping constant) in this work. At last, The hardware performance improvement, e.g., power efficiency, latency, and design area, is evaluated through a co-simulation paradigm with PyTorch and *NeuroSIM* [9].

In this section, the crossbar configuration of the memristor as a memory array is introduced. Next, our Deep-DFR model is introduced in detail. At last, the hybrid simulation paradigm is presented, combining our experimentally verified memristor model and the Python-based Deep-DFR model.

A. Weight Storage in Memristor Crossbar

Memristors typically are fabricated in a crossbar structure massively. As illustrated in Figure 7, the nanowires built with the inert cathodes and oxidizable active anodes are placed at the top and bottom of the crossbar, respectively. The metallic oxide layer is located at the cross points of the top and bottom nanowires. This crossbar structure is similar to the conventional memory array. As illustrated in Figure 7(b), each memory cell of the memory array connects to a *wordline* and a *bitline.*

For example, the DRAM (Dynamic Random-access Memory) uses a capacitor for each memory cell, and the SRAM (Static Random-access Memory) generally has six transistors as one memory cell. The stored information is represented by the voltage states at the terminals of capacitor or transistor. For memristor, the values are encoded in the resistance of a memristor and the nanowires serve as the bitline and wordline for accessing the memristive memory cells. Figure 7 depicts the writing and reading phases of a memristive memory cell. In the writing phase, a voltage pulse, larger than set voltage, is applied to the nanowire of the crossbar structure and modifies the resistance value of the memristor. In the reading stage, the applied voltage is much smaller than the set voltage in order to preserve the resistance of the cell unaltered. The resistance value of the selected memristor equals the applied voltage divided by the measured current at the end of the nanowire. The weight matrices are mapped to the passive memristor crossbar with the memory cell selection devices, such as transistor or diode. The decoder of the system uses the *wordline* and *bitline* to access to every single memory cell. As illustrated in Figure 7 (a), the operations of weight sum and update in NeuroSIM are row-by-row-based write and reading [9]. The row selection is activated through the WL decoder. Then the BLs are precharged to each cell access. The memory data are captured by the sense amplifier (S/A). After that, the adder and register are used to sum the weight values in a row-by-row style. By replacing the SRAM core memory with the memristors, the architecture is not significantly modified (Figure 7). But the size of the memory cell reduces due to the intrinsic nanoscale of memristors. The weighted sum operation in the memristor-based synaptic core is also a row-by-row style expect the use of multiplexers (Mux) [9].

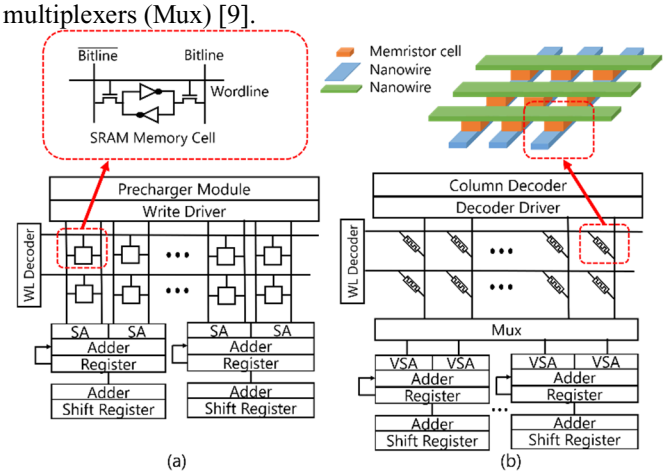


Figure 7: Configuration comparison between the memristive crossbar and the memory array with SRAM memory cells in NeuroSIM [9]: (a) the traditional memory array with SRAM (6T); (b) the structure of the memristive crossbar.

0278-0070 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on January 04,2021 at 19:52:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2020.3002539, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

B. Deep Reservoir Neural Network

Nowadays, hardware-friendly DFR demonstrates an impressive capability of processing temporal information [27, 28]. In this work, several convolutional layers are added for constructing a deep DFR structure. Figure 6 illustrates the details of our Deep-DFR structure. The six convolutional layers serve as feature extractor, which is followed with a delay-feed-back layer extracts the one-dimensional time series characteristics. Two fully connected layers are used for reducing the output dimensional serving as a classifier. The number of time delay reservoir layers matches the output of the convolutional layer. Initially, the weights in the reservoir ($\boldsymbol{W}^{res}$) layer is assigned as zeros. During the training process, the updating equation of the reservoir state is expressed as:

$$Res(t) = \alpha \times Res(t-1) + f_{nonlinear}(\boldsymbol{H}^{in}(t)), \quad (7)$$

where t is the time step, Res(t) is the reservoir state, α is the decay factor, $f_{nonlinear}$ is the nonlinear activation function, and $\boldsymbol{H}^{in}$ is the hidden layer. This equation reveals that the current state of the reservoir is not only determined by the current input but also highly related to the last time step.

Algorithm 1: Performance Estimation

Initialize: The configuration of the Deep-DFR and the corresponding weights $\boldsymbol{W}_{i,j}$ with small random numbers
Initialize: $\mathbf{W}^{\mathrm{res}}$ of the reservoir as all zeros
Initialize: Memory cell configuration
Initialize: Peripheral circuits configuration
1 For epoch = 1, M do
2 While batch in dataset do
4 $\boldsymbol{y}_{conv}^{out} \leftarrow$ six convolutional layers to batch (input)
3 $\boldsymbol{h}_{res\_1} = \boldsymbol{W}_{res}^{in} \times \boldsymbol{y}_{conv}^{out} + bias$
4 $\boldsymbol{W}^{res} = \alpha \times \boldsymbol{W}^{res} + nonlinear(\boldsymbol{h}_{res\_1})$
5 $\boldsymbol{h}_{res\_2} = \boldsymbol{W}_{res}^{out} \times \boldsymbol{W}^{res} + bias$
6 $\boldsymbol{y}_{res}^{out} = f_{nonlinear}(\boldsymbol{h}_{res\_2})$
7 $\boldsymbol{y}_{classifier}^{out} \leftarrow$ full-connected layer as classifier to $\boldsymbol{y}_{res}$
8 $\hat{\boldsymbol{y}} = softmax\ (\boldsymbol{y}_{classifier}^{out})$
7 $loss = cross\_entropy\ (\hat{\boldsymbol{y}}, \boldsymbol{y})$
8 *Minimize(loss)*
9 End While
10 End For
11 Store weights and neural network configuration
12 Calculate Area of Peripheral circuits based on their configuration
13 Calculate total area = memristor memory array area + Σ area of the peripheral circuits
14 Recall Stored weights
15 For number of the weight index = 1, N do
16 Calculate latency of Peripheral circuits with RC as load parameters
17 Total latency = Σ (latency) of peripheral circuits in each operation
18 Total energy = array dynamic/static energy + Σ (dynamic energy) of peripheral circuits in each operation
19 End For

To evaluate our memristor performance, e.g., design area, accuracy, power consumption, a hardware-software co-simulation is established with *PyTorch* and *NeuroSIM* [9], as illustrated in Figure 6. The model is built as follows steps:

Firstly, our Deep-DFR model is built of six convolutional layers for extracting features, followed by a Delay Feedback Reservoir Layer, and two full-connected layers. There are no weights within the delay feedback loop [30]. The Deep-DFR model is trained on the PyTorch platform with CIFAR-10 and CIFAR-100 datasets. During the training progress, the weights and neural network configuration are monitored and stored.

Secondly, our experimentally verified memristor model is incorporated into the micro-architecture simulator *NeuroSIM* [9] including the set voltage, on-state resistance, off-state resistance. The resistance variation with different levels (TABLE I) is incorporated in the memristor model in *NeuroSIM*. To reveal intently the cause-and-effect relationship between resistance variation (cycle-to-cycle) and inference accuracy, other nonideal parameters of memristors are not included for eliminating the interference from them.

Thirdly, the Python API deploys the saved weights and configurations of the Deep-DFR to the *NeuroSIM* for hardware performance inference. The deployment method evaluates the performance of the neural network system on an offline training environment which demands a local computation. Compared to Online Learning, Offline Learning training keeps the trained neural network at the client-side and perform all prediction computation locally [32], due to the limited energy and space budget at the client-side.

At last, the performance improvements of our memristor on energy, design area, execution latency, and accuracy are estimated through the co-simulation paradigm. The pseudocode of our hardware-software co-simulation paradigm is introduced in Algorithm 1.

C. Performance Evaluation

Using our co-simulation paradigm introduced in the previous subsection, the performance improvement of our memristor on deep learning at the system level is evaluated and estimated. The inference accuracy degrades significantly while the resistance variation of the memristor increase [6, 10, 11]. Figure 8 presents a correlation analysis between the variation of the weights and the inference accuracy of our Deep-DFR model. The Deep-DFR models are trained with the CIFAR-10 and CIFAR-100 datasets in 150 epochs. The model structure details are depicted in Figure 6. The simulation results demonstrate a strong negative correlation between the testing accuracy and the variation of the weights. For example, in Figure 8 (a), the testing accuracy significantly reduces while the variation of the weight increases, specifically in the range from 0.2 to 0.6. After the weight variations reach the range larger than 0.6, the testing accuracies tend to be stable and are at low levels (lower than 13%). The testing accuracies with different memristive devices, associating with their variations, are marked in the testing accuracy curve. Our memristive device (Cu/TaOx/Rh/Cr) reaches the highest testing accuracy (~90%) due to its lower variation compared to other devices. The simulation results using the CIFAR-100 dataset (Figure 8 (b)) illustrates a similar degradation trend of the testing accuracy. The difference is the testing accuracy on CIFAR-100 reduces faster than CIFAR-10 and reaches its stable range on 0.4 weight variation.

The simulation results with CIFAR-10 and CIFAR-100 both demonstrate the accuracies of the Deep-DFR models constituted of our memristor (1% variation) outperform the

0278-0070 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on January 04,2021 at 19:52:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2020.3002539, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

other state-of-the-art memristors, and other material configurations we explored (listed in TABLE I).

The main advantage of storing weights of the neural networks in memristors is to enhance hardware performance. In this work, we compared our memristor with SRAM and other state-of-the-art memristor reported, which are implemented with other materials, such as Ag:SiGe [33] and AlOx/HfO_2 [34].

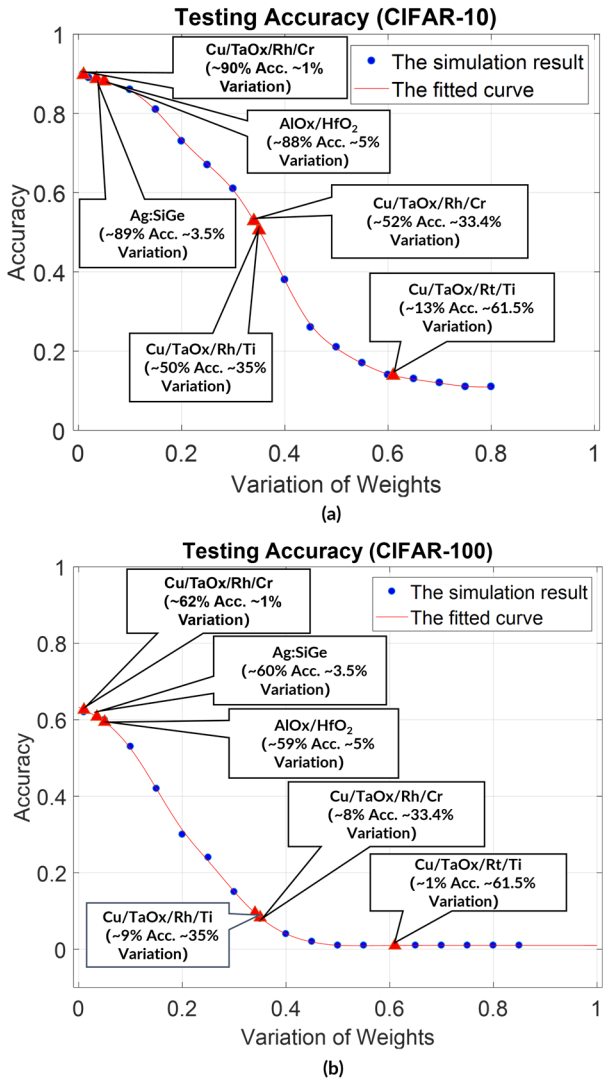


Figure 8: The reduction in the accuracy accompanying the increase of the weight variation: (a) CIFAR-10 and (b) CIFAR-100. The neural network model is our Deep-DFR. The blue cycles indicate the simulation results and the red line represents the fitted curve. The memristive device of Ag:SiGe and AlOx/HfO_2 are from [33] and [34] respectively.

The hardware performance enhancement with different memory techniques in the design area, power consumption, and computing latency are inferred and compared using *NeuroSIM* [9]. The settings of the model are summarized in Table II. The SRAM is implemented in the typical six-transistor cell (6T) with 32 nm technology. The weights are stored in memristors in digital format since the analog memristive synapse degrades the learning accuracy [9]. The weights are stored in 4-bit precision. The feature size of the memristor is assigned at 40 nm because the current industry technology of integrating memristors and the transistors is at the range of 40 nm to 28 nm [9]. The configuration detail of *NeuroSIM* is illustrated in Figure 7, which includes the essential modules for estimating the writing/reading performance parameters of accessing the memory array, such as decoder, encoder, adder, register, and so on. The simulation calculates all the latency, design area, and power consumption from different function modules, including the main memory module (SRAM and memristors) and the periphery circuits. The breakdown results of each module are listed in TABLE III, which uses CIFAR-10 dataset.

Table II: SIMULATION SETTING OF NEUROSIM MODEL

Device	SRAM	Memristors
Frequency	1 GHz	1 GHz
Temperature	301 K	301 K
Subarray size	64 × 64	64 × 64
Read Voltage	1.1 V	0.5 V
Read Pulse Width	N/A	10 ns
Structure	6T	1R
Technology	32 nm	40 nm

TABLE III: SIMULATION RESULT BREAKDOWN OF CHIP PERFORMANCE

Device	[34]	[35]	SRAM	VT Memristor
Chip Area (mm^2)	98.05	138.83	166.17	85.97
IC Area on chip (mm^2)	2.90	3.50	4.24	2.70
ADC Area on chip (mm^2)	14.03	14.03	42.68	14.03
Periphery circuits (mm^2)	47.50	84.66	52.89	39.05
Chip total Read Latency (us)	423.34	1082.33	803.38	264.97
Chip total Read Dynamic Energy (uJ)	44.1533	55.48	70.88	41.51
Chip total Leakage Energy (nJ)	223.37	699.91	966.03	108.90
Chip total Leakage Power (uW)	791.09	791.09	3074.87	791.09
Chip buffer Read Latency (us)	12.36	12.36	12.36	12.36
Chip buffer read Dynamic Energy (uJ)	4.16	4.16	5.87	4.16
Chip IC Read Latency (us)	36.22	49.40	28.40	32.77
Chip IC Read Dynamic Energy (uJ)	24.39	34.38	25.94	21.92
ADC Read Latency (us)	39.93	39.34	81.66	42.25
Periphery circuits read Latency (us)	214.10	873.62	93.78	53.48
ADC Read Dynamic Energy (uJ)	3.77	3.39	13.12	4.01
Periphery circuits read Dynamic Energy (uJ)	30.60	42.30	35.01	27.71

0278-0070 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on January 04,2021 at 19:52:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2020.3002539, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

Figure 9 demonstrates that our memristor reduces chip area, power consumption and latency reduce by ~48%, ~42%, and ~67% with respect to SRAM, respectively. Furthermore, the performance is improved at various degrees compared to other state-of-the-art memristors [33, 34]. The improvements show similar levels with the datasets of CIFAR-10 and CIFAR-100 in Figure 9 (a) and Figure 9 (b). This phenomenon probably stems from the same neural network model (Deep-DFR) and a similar value range of data (CIFAR-10 and CIFAR-100), which leads to a similar number and values of the weights.

The area difference of memristors in Figure 9 mainly comes from the periphery circuits. The larger area of periphery circuits of memristors of Ag:SiGe and AlOx/HfO_2 [33, 34] stem from their small on-state resistance [33-35]. The small on-state resistance requires the larger size (W/L) of transistors in peripheral circuits, e.g., Mux or switch matrixes, to avoid the significant current drop and impedance mismatch [35]. Accordingly, the latency of periphery circuits also increases due to the large size of the transistors, which needs a longer time for charging and discharging.

As a non-volatile device, the memristors store the data in their resistances. Unlike SRAM, the non-volatile memory cores do not need a power supply to retain the data in memory cells thus their leakage power is much smaller than a typical SRAM. The energy reduction of other state-of-the-art memristors (Ag:SiGe and AlOx/HfO_2 [33, 34]) is much less than our memristors because of their smaller on-state resistance (R_{on}). The small on-state resistance leads the array static energy (consumed by cells) dominates rather than the dynamic energy in the system. The static energy consumes more energy in the system, which leads our memristor is much energy efficient compared to Ag:SiGe and AlOx/HfO_2 [33, 34].

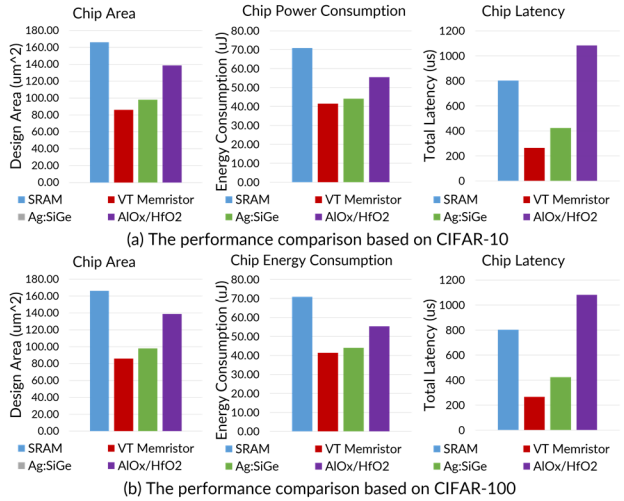


Figure 9: Performance evaluation on the different memory techniques: (a) CIFAR-10 and (b) CIFAR-100. The memristive device of Ag:SiGe and AlOx/HfO_2 are from [33] and [34] respectively

IV. Conclusion

In this work, a novel memristor configuration with the enhanced heat dissipation feature is designed and fabricated. The measurement data demonstrate our memristor has higher immunity to degradation induced by the thermal effect. The on and off resistance variations are reduced correspondingly, leading to an increase of the testing accuracy within the same range. Our Deep-DFR model is used for evaluating our memristor as the weight storing devices. The datasets CIFAR-10 and CIFAR-100 are used for training the Deep-DFR model. The design area, power consumption, and latency of the system using our memristor are reduced by ~48%, ~42%, and ~67% compared to conventional SRAM memory technique. At last, these hardware parameters are also improved at various degrees (~13%-73%) compared to other state-of-the-art memristors [33, 34].

Acknowledgment

The measurement of our memristive devices was conducted at the Center for Nanophase Materials Sciences (CNMS), which is a Department of Energy Office of Science User Facility. We deeply appreciate their valuable assistance on this project.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[3] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016.

[4] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," *Proceedings of the IEEE International Conference on Computer Vision,* vol. 2017-Octob, pp. 843-852, 2017.

[5] B. Govoreanu *et al.*, "10× 10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, 2011: IEEE, pp. 31.6. 1-31.6. 4.

[6] H. S. P. Wong *et al.*, "Metal-oxide RRAM," *Proceedings of the IEEE,* vol. 100, pp. 1951-1970, 2012.

[7] H. An, M. S. Al-Mamun, M. K. Orlowski, and Y. Yi, "Learning Accuracy Analysis of Memristor-based Nonlinear Computing Module on Long Short-term Memory," in *Proceedings of the International Conference on Neuromorphic Systems*, 2018: ACM, p. 5.

[8] Y. Long, T. Na, and S. Mukhopadhyay, "ReRAM-based processing-in-memory architecture for recurrent neural network acceleration," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 26, no. 12, pp. 2781-2794, 2018.

[9] P.-Y. Chen, X. Peng, and S. Yu, "NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning," *IEEE Trans. Comput-Aided Des. Integr. Circuits Syst.,* vol. 37, no. 12, pp. 3067-3080, 2018.

[10] X. Guan, S. Yu, and H.-S. P. Wong, "A SPICE Compact Model of Metal Oxide Resistive Switching Memory With Variations," *IEEE Electron Device Letters,* vol. 33, pp. 1405-1407, 2012.

[11] A. Chen and M.-R. Lin, "Variability of resistive switching memories and its impact on crossbar array performance," in *2011 International Reliability Physics Symposium*, 2011: IEEE, pp. MY. 7.1-MY. 7.4.

[12] B. Liu, H. Li, Y. Chen, X. Li, Q. Wu, and T. Huang, "Vortex: variation-aware training for memristor x-bar," in *Proceedings of the 52nd Annual Design Automation Conference*, 2015: ACM, p. 15.

[13] L. Chen *et al.*, "Accelerator-friendly neural-network training: Learning variations and defects in RRAM crossbar," in *Proceedings of the Conference on Design, Automation & Test in Europe*, 2017: European Design and Automation Association, pp. 19-24.

[14] Y. Long, X. She, and S. Mukhopadhyay, "Design of reliable DNN accelerator with un-reliable ReRAM," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2019: IEEE, pp. 1769-1774.

0278-0070 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on January 04,2021 at 19:52:21 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2020.3002539, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

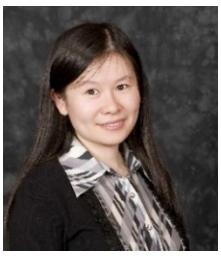
[15] D. Ielmini, "Modeling the universal set/reset characteristics of bipolar RRAM by field-and temperature-driven filament growth," *Ieee T Electron Dev,* vol. 58, no. 12, pp. 4309-4317, 2011.
[16] G. Ghosh and M. K. Orlowski, "Write and erase threshold voltage interdependence in resistive switching memory cells," *IEEE transactions on Electron Devices,* vol. 62, no. 9, pp. 2850-2856, 2015.
[17] W. Wu, H. Wu, B. Gao, N. Deng, S. Yu, and H. Qian, "Improving analog switching in HfO x-based resistive memory with a thermal enhanced layer," *IEEE Electron Device Letters,* vol. 38, no. 8, pp. 1019-1022, 2017.
[18] M. Al-Mamun, S. W. King, S. Meda, and M. K. Orlowski, "Impact of the Heat Conductivity of the Inert Electrode on ReRAM Performance and Endurance," *ECS Transactions,* vol. 85, no. 8, pp. 207-212, 2018.
[19] M. Al-Mamun, S. W. King, and M. K. Orlowski, "Impact of the Heat Conductivity of the Inert Electrode on Reram Memory Cell Performance and Endurance," in *Meeting Abstracts*, 2018, no. 24: The Electrochemical Society, pp. 1476-1476.
[20] Y. Fan, M. Al-Mamun, B. Conlon, S. W. King, and M. K. Orlowski, "Resistive Switching Comparison between Cu/TaOx/Ru and Cu/TaOx/Pt Memory Cells," *ECS Transactions,* vol. 75, no. 32, p. 13, 2017.
[21] C. Walczyk *et al.*, "Impact of Temperature on the Resistive Switching Behavior of Embedded HfO_2-Based RRAM Devices," *IEEE transactions on electron devices,* vol. 58, no. 9, pp. 3124-3131, 2011.
[22] C. D. Landon *et al.*, "Thermal transport in tantalum oxide films for memristive applications," *Applied Physics Letters,* vol. 107, no. 2, p. 023108, 2015.
[23] P. Sun *et al.*, "Physical model of dynamic Joule heating effect for reset process in conductive-bridge random access memory," *Journal of Computational Electronics,* vol. 13, no. 2, pp. 432-438, 2014.
[24] S. Kaeriyama *et al.*, "A nonvolatile programmable solid-electrolyte nanometer switch," *IEEE Journal of Solid-State Circuits,* vol. 40, no. 1, pp. 168-176, 2005.
[25] H. An, M. A. Ehsan, Z. Zhou, and Y. Yi, "Electrical modeling and analysis of 3D synaptic array using vertical RRAM structure," in *Quality Electronic Design (ISQED), 2017 18th International Symposium on*, 2017: IEEE, pp. 1-6.
[26] Z. Jiang, S. Yu, Y. Wu, J. H. Engel, X. Guan, and H. S. P. Wong, "Verilog-A Compact Model for Oxide-based Resistive Random Access Memory (RRAM)," pp. 41-44, 2014.
[27] K. Bai, Q. An, and Y. Yi, "Deep-DFR: A Memristive Deep Delayed Feedback Reservoir Computing System with Hybrid Neural Network Topology," *56th ACM/ESDA/IEEE Design Automation Conference (DAC) IEEE,* 2019.
[28] C. Gallicchio, A. Micheli, and L. Pedrelli, "Deep reservoir computing: a critical experimental analysis," *Neurocomputing,* vol. 268, pp. 87-99, 2017.
[29] J. Li, K. Bai, L. Liu, and Y. Yi. A Deep Learning Based Approach for Analog Hardware Implementation of Delayed Feedback Reservoir Computing System.
[30] L. Appeltant *et al.*, "Information processing using a single dynamical node as complex system," *Nat Commun,* vol. 2, p. 468, 2011.
[31] L. Appeltant, G. Van der Sande, J. Danckaert, and I. Fischer, "Constructing optimized binary masks for reservoir computing with delay systems," *Scientific reports,* vol. 4, p. 3629, 2014.
[32] N. D. Lane, P. Georgiev, and L. Qendro, "DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 283-294.
[33] S. Choi *et al.*, "SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations," *Nature materials,* vol. 17, no. 4, pp. 335-340, 2018.
[34] J. Woo *et al.*, "Improved synaptic behavior under identical pulses using AlOx/HfO2 bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Letters,* vol. 37, no. 8, pp. 994-997, 2016.
[35] P.-Y. Chen and S. Yu, "Technological Benchmark of Analog Synaptic Devices for Neuroinspired Architectures," *IEEE Design & Test,* vol. 36, no. 3, pp. 31-38, 2018.

Hongyu An received a M.S. degree in Electrical Engineering from Missouri University of Science and Technology, USA. He is currently pursuing his doctoral degree in Department of Electrical and Computer Engineering, Virginia Tech. His research interests include neuromorphic computing, nano-electronic devices and circuits for energy-efficient neuromorphic systems.

Mohammad Shah Al-Mamun received his M.S. degree and Ph.D. degree in The Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University in 2015 and 2019, respectively. After graduation, he is with the Intel Labs memory team and works on 7 nm transistor development.

Marius Orlowski (F'97) received an M.S. degree in experimental physics and a Ph.D. degree in theoretical nuclear physics from the University of Tübingen, Tübingen, Germany, in 1976 and 1979, respectively. He has been a Professor with the Department of Electrical and Computer Engineering, Virginia Tech, since 2008. He holds over 80 U.S. patents. Dr. Orlowski has received several awards, including Motorola Master Innovator, Distinguished Innovator awards, and Fulbright Fellow 2014 awards.

Lingjia Liu (SM'15) is an Associate Professor in the Bradley Department of Electrical and Computer Engineering at Virginia Tech. Dr. Liu's research interests mainly lie in emerging technologies for 5G cellular networks and Beyond including machine learning for wireless networks, massive MIMO, massive machine type communications (MTC), and Internet of Everything (IoE).



Yang Yi (M'09-SM16) is an Associate Professor in the Bradley Department of Electrical Engineering and Computer engineering at Virginia Tech. She received the B.S. and M.S. degrees in electronic engineering at Shanghai Jiao Tong University, and the Ph.D. degree in electrical and computer engineering at Texas A&M University. Her research interests include very large scale integrated (VLSI) circuits and systems, neuromorphic architecture for brain-inspired computing systems, and low-power circuits design with advanced nanotechnologies for high speed wireless systems.

0278-0070 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on January 04,2021 at 19:52:21 UTC from IEEE Xplore. Restrictions apply.