Impact of RTN on Pattern Recognition Accuracy of RRAM-based Synaptic Neural Network

Zheng Chai, Pedro Freitas, Weidong Zhang, Firas Hatem, Jian Fu Zhang, John Marsland, Bogdan Govoreanu, Ludovic Goux and Gouri Sankar Kar

***Abstract*— Resistive switching memory devices can be categorized into either filamentary or non-filamentary ones depending on the switching mechanisms. Both types have been investigated as novel synaptic devices in hardware neural networks, but there is a lack of comparative study between them, especially in random telegraph noise (RTN) which could induce large resistance fluctuations. In this work, we analyze the amplitude and occurrence rate of RTN in both Ta_2O_5 filamentary and TiO_2/a-Si (a-VMCO) non-filamentary RRAM devices and evaluate its impact on the pattern recognition accuracy of neural networks. It is revealed that the non-filamentary RRAM has a tighter RTN amplitude distribution and much lower RTN occurrence rate than its filamentary counterpart which leads to negligible RTN impact on recognition accuracy, making it a promising candidate in synaptic application.**

***Index Terms*—random telegraph noise, RRAM, pattern recognition, neural network, filamentary, Si, TiO_2, Ta_2O_5, RTN**

INTRODUCTION

Oxide based resistive switching (RS) memory device (RRAM) has emerged as an attractive candidate for synapses in large-scale artificial neural networks (ANNs) due to its natural synaptic response, simple structure, low energy consumption, and CMOS-compatible 3D integration potential [1]. RRAM can be categorized into either filamentary or non-filamentary, for the RS caused by the restore/rupture of a conductive filament (CF) or the areal modulation of defect distribution inside the oxide, respectively. Both types have been intensively studied for novel synaptic applications [2-6], but there is a lack of comparative analysis between them.

Random Telegraph Noise (RTN) is the current fluctuation between discrete levels caused by electron trapping and de-trapping in defects. RTN has become a critical issue in nanoscale semiconductor devices where the impact of a single defect becomes significant [7]. As RRAM devices can be scaled down below 10 nm [8], RTN could significantly reduce its memory window and cause read errors. It is therefore essential to evaluate the impact of RTN disturbance on the performance of RRAM-based synaptic arrays. On the other hand, RTN provides useful information on the responsible defect [9, 10]. The impact of RTN has been analyzed for CF RRAM [11], but there is a lack of comparative studies on the non-CF (NCF) RRAM whose synaptic application has also drawn extensive interests [4, 12].

In this work, we analyze the amplitude distributions and occurrence rate of RTN in both Ta_2O_5 CF RRAM and TiO_2/a-Si (a-VMCO) NCF RRAM devices. Based on the experimental results, a novel RTN disturbance model is developed to simulate its impact on the synapse arrays in a trained artificial neural network. It is revealed that the NCF RRAM has a tighter RTN amplitude distribution and much lower RTN occurrence rate than its filamentary counterpart, leading to negligible RTN impact on recognition accuracy. It proves to be a promising candidate as synapse in neural network applications.

DEVICES AND EXPERIMENTS

Both types of RRAM devices were fabricated in a cross-point structure with the size of 75 nm ×75 nm and show bipolar switching characteristics (Fig. 1(a) and (b)). The Ta_2O_5 device consists of a TiN/4nm stoichiometric Ta_2O_5/20nm nonstoi-chiometric TaOx/10nm TaN/TiN stack (inset of Fig. 1 (a)). The a-VMCO device has a stack of TiN/8nm amorphous-Si/8nm anatase TiO_2/TiN structure (inset of Fig. 1(b)). The detailed process parameters can be found in refs. [13-14]. All electrical tests were done with a Keysight B1500A analyzer. Eight uniformly distributed resistance levels are obtained in both devices, between 25 kΩ and 200 kΩ for Ta_2O_5, and between 1 MΩ and 7.5 MΩ for aVMCO, by incrementing the reset voltages [18]. The read-out is at 0.1V and 3V for Ta_2O_5 and aVMCO devices, respectively, averaged by five consecutive 20-ms read-out tests, followed by a 400-ms delay period. RTN measurement is then carried out at each R level at the read-out voltage, with a sampling time of 2 ms/point and 10,000 sampling point per resistance level for a RTN measurement period of 20 s. A 3-layer ANN was simulated using Matlab [15, 16]. The neural network was trained and tested with the MNIST handwritten digit database [17]. Out of the total 60,000 images, 50,000 were used for training and the remaining 10,000 images unseen during training were used for testing.

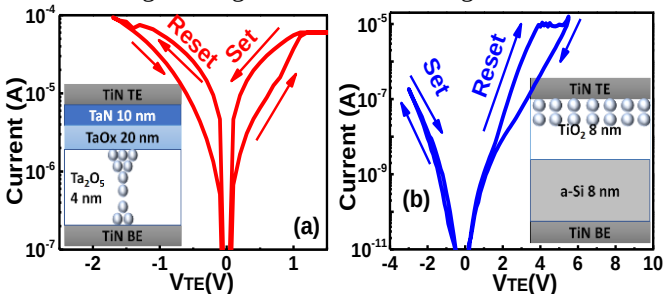

Fig. 1. I-V switching curves of Ta_2O_5 (a) and aVMCO (b) devices; The insets are the schematics of the corresponding structures and the switching mechanism: the restore/rupture of a conductive filament (CF) or the areal modulation of defect distribution inside the oxide (NCF).

Manuscript received xx. This work was supported by the EPSRC of U.K. under Grants EP/M006727/1 and EP/S000259/1. The review of this paper was arranged by Editor Dr. Chandra Mouli.

Z. Chai, W. Zhang, P. Freitas, F. Hatem, J. F. Zhang, J. Marsland are with the Department of Electronics and Electrical Engineering, Liverpool John Moores University, L3 3AF Liverpool, U.K. (e-mail:W.Zhang@ljmu.ac.uk).
B. Govoreanu, L. Goux, and G.S.Kar are with IMEC, 3001 Leuven, Belgium.

RESULTS AND DISCUSSIONS

As shown in Fig. 2(a) and (b), the maximum relative RTN amplitude, $\Delta I/I_{read}$, can be as high as ~300% in Ta_2O_5 device, but only ~10% in the non-filamentary aVMCO. Their CDF distribution plots measured at the 8 resistance levels are shown in Fig. 2 (c) and (d), respectively. RTN amplitude in Ta_2O_5 device spreads widely from 0.1% to 300%, whilst it is only from 1% to 10% in aVMCO. For both devices, the RTN amplitude follows the lognormal distribution [19]. Moreover, RTN in Ta_2O_5 device has a much higher occurrence rate than in aVMCO device, as shown in Fig. 3(a).

This significant difference in RTN amplitude distribution and occurrence rate can be attributed to the different switching mechanisms, as shown in the insets in Fig.1: in the CF Ta_2O_5 device, the resistance switching is caused by the rupture and restoration of a conductive filament. After the reset, there are only a few defects in the constriction of the CF, and each of them is critical in current conduction, so that its trapping/ detrapping leads to large RTN, and hence the higher the resistance level, the larger the RTN amplitude [20]. In the NCF aVMCO device, RS is caused by the uniform modulation of defect distribution [14]. Resistance becomes higher when the "defect-less" region is uniformly widened. A single defect has limited contribution in conduction, hence the much smaller RTN amplitude and much smaller occurrence rate, and the amplitude is also only slightly larger at higher resistance levels.

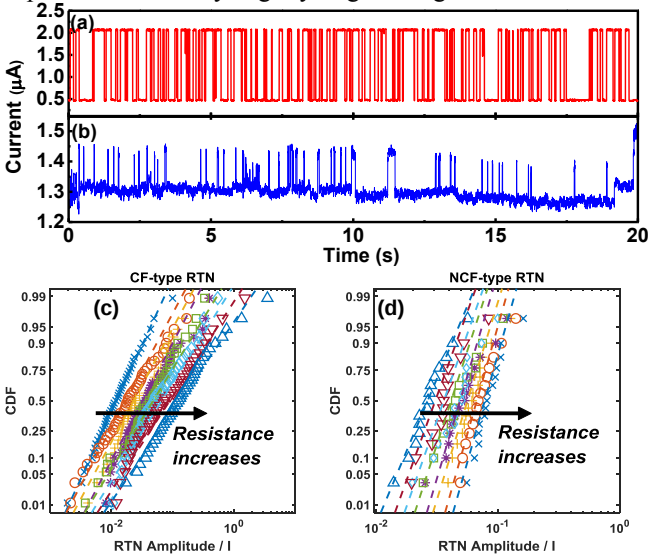

Fig. 2. (a-b) Examples of largest RTN signal in (a) Ta_2O_5 and (b) a-VMCO devices. The relative RTN amplitude can be as high as ~300% for Ta_2O_5 device, but only ~10% for a-VMCO. (c-d) CDF of relative RTN amplitude in Ta_2O_5 (c) and a-VMCO (d) devices, respectively, both following the lognormal distribution.

The probability and cumulative distribution functions (PDF and CDF) of lognormal distribution are described as

$$y = f(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt[2]{2\pi}} e^{\frac{-(lnx-\mu)^2}{2\sigma^2}} \quad \text{Eq.(1)}$$

$$p = F(x|\mu,\sigma) = \frac{1}{\sigma\sqrt[2]{2\pi}} \int_0^x \frac{e^{\frac{-(lnx-\mu)^2}{2\sigma^2}}}{t} dt \quad \text{Eq.(2)}$$

respectively. The mean, μ, and standard deviation, σ, at eight resistance levels for both devices are extracted from the experimental data shown in Fig. 2 (c) and (d) and their values in log scale are plotted in Figs. 3(a-b). The RTN time constants are extracted with the Hidden Markov Model for both devices, and their CDF distribution are shown in Fig. 3(c) and (d), respectively. The distributions of time constants in both devices are very similar, at least within the sampling rate and time used in this work. For the purpose to carry out comparative analysis, we consider that the impact of time constant is similar between the two types of devices, therefore.

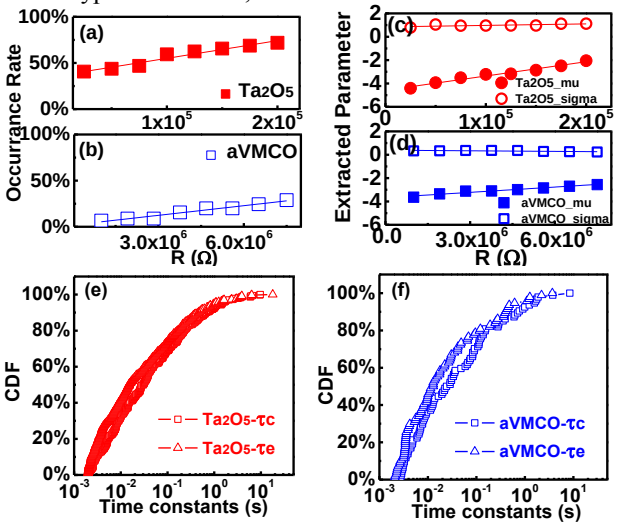

Fig. 3. (a-b) Occurrence rate of RTN signal at 8 resistance levels, in Ta_2O_5 (a) and a-VMCO (b) devices respectively, obtained by dividing the number of RTN occurrence by the total number of RTN tests. (c-d) Extracted parameters of lognormal RTN amplitude distribution in both devices. (e-f) CDF of RTN time constants in Ta_2O_5 (e) and aVMCO (f).

The comparative analysis of RTN's impact on the pattern recognition accuracy of RRAM based synaptic neural network is divided into three steps: First, the neural network is trained with the mini-batch gradient descent backpropagation algorithm for demonstration purpose. As an example, the neural network consists of 3 layers with 30 neurons in the hidden layer, as shown in Fig. 4(a). The trained network is firstly tested for the ideal accuracy without considering the RTN induced disturbance. The recognition accuracy after training reaches ~95%, which is satisfactory [21]. Secondly, RTN induced disturbance is then added to the synaptic array in the simulation. For each synapse, a random RTN amplitude variation is generated using the CDF distributions in Fig. 2(c-d) and the model in Eqs. (1) and (2). Next, the RTN amplitude variations are applied to the synaptic network according to the occurrence rate at its corresponding R level. The μ, σ and occurrence rate are obtained by linear fitting and interpolating the experimental data in Fig.3 (a-b) and mapping the disturbed R value to the trained weights. Thirdly, the trained network consisting of the RTN induced disturbance is tested for both types of devices and the resultant accuracy is compared.

It should be noted that the weights in the simulation can be both positive and negative. To implement this feature in synaptic hardware array, the synaptic weights are separated into two matrices, one containing all the positive weights and the other containing all the negative weights, and two RRAM crossbar arrays are implemented in parallel and then an analog subtraction circuit is used to process the results of the two cross bar arrays [22-25]. To obtain a statistically reliable result, the training-disturbance procedure is repeated for 50 times and the accuracy after training without and with both CF disturbance

and NCF disturbance are statistically shown in Fig. 4(b). The change of weights in one of these procedures is visualized in Fig. 4(c), in which the weights are shown in (1) without disturbance, (2) after the CF disturbance, and (3) after the NCF disturbance. The weight differences are shown in (4) after CF disturbance and (5) after NCF disturbance.

As shown in Fig. 4(b), after the CF RTN disturbance, the average accuracy drops to ~75% with a wide repeatability distribution and its lowest is less than 50%, while after the NCF disturbance the accuracy drops negligibly only to 94% with a similar repeatability to that without disturbance, as can also be clearly seen in the weight differences shown in Fig. 4(c). This proves that the non-filamentary RRAM device has a strong advantage compared to the conventional filamentary devices in the synaptic application, due to its small RTN amplitude and low RTN occurrence rate. Furthermore, as shown in Fig.4(d), the synaptic network with NCF devices maintains a high accuracy of ~90% even when only 10 neurons are used in the hidden layer, whilst the accuracy drops sharply with the CF devices. NCF synaptic devices allows the ANN to use much less number of neurons and synapses to achieve better accuracy due to its robust RTN resilience, therefore.

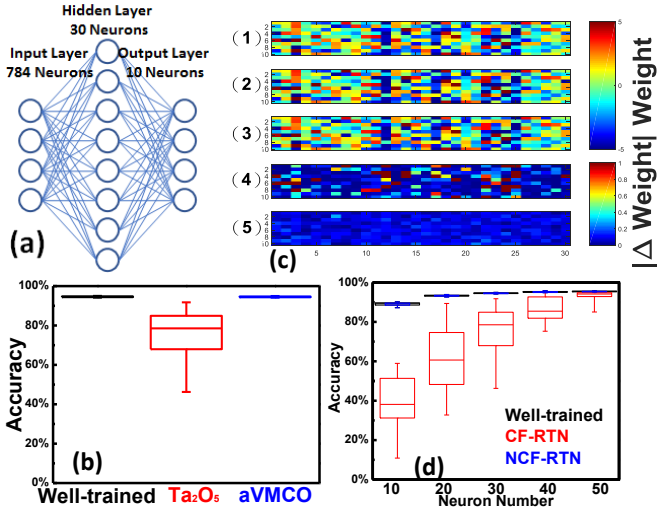

Fig. 4. (a) Schematic of the pattern recognition ANN. (b) Statistical accuracy in 50 training-disturbance procedures: Accuracy is hardly affected with the NCF disturbance, while with the CF disturbance the accuracy is severely deteriorated. (c) Visualization of weights: (1) directly after training; (2) with CF RTN disturbance; (3) with NCF RTN disturbance; (4-5) their differences to case (1), respectively. (d) Accuracy of ANN with different neuron number without and with CF and NCF RTN disturbance. ANN with NCF devices needs fewer neurons/synapse and have better accuracy.

One concern is that the impact of RTN may diminish for neurons with a larger number of inputs, as the "averaging" effect of independent variation sources scales following 1/sqrt(N). To examine this possibility, we investigate the impact of RTN on NN with different number of inputs. As shown in Fig. 5, the accuracy and error rates improved, but only slightly when the input number increase from 28x28 to 56x56, instead of following the 1/sqrt(N) rule. This is probably due to that some synapses play a more critical role in pattern recognition and their weight fluctuation affects the accuracy more than other synapses, making the averaging effect weaker. RTN may remain an issue for larger neural networks, therefore.

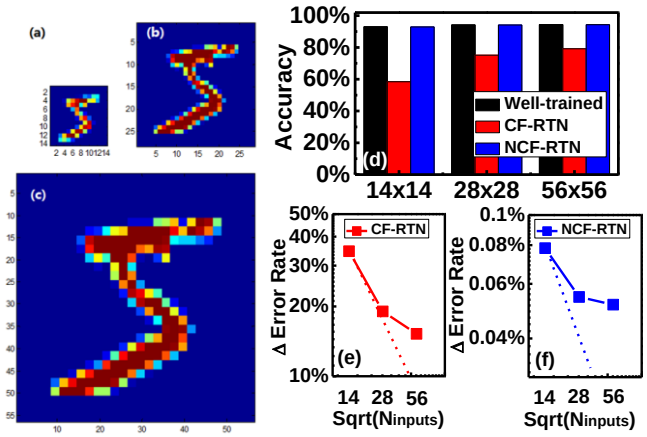

Fig. 5 Pattern recognition accuracy with MNIST images of different resolutions. (a-c) Example of the rescaled MNIST image, with (a) 14x14-pixel, (b) original 28x28-pixel and (c) 56x56-pixel. (d) Pattern recognition accuracy after training, with CF-type RTN and with NCF-type RTN, respectively. (e-f) log-log plot of relative error rate ($Accuracy_{well\text{-}trained} - Accuracy_{CF\text{-}\ or\ NCF\text{-}RTN}$) against square root of the total number of input neurons. The straight dash lines are guide to the eye for the scaling rule of 1/sqrt(Number of inputs).

It should be noted that noise in RRAM has been shown to have complicated structures, including RTN, 1/f noise [26], and random walk (RW) [27]. We found that the 1/f noise and RW in both of our CF and NCF devices has much lower occurrence and/or smaller amplitude when compared with RTN, and has limited impact on the pattern recognition accuracy investigated in this work, therefore. In order to focus on analyzing the impact of RTN signals in this work, we have carefully examined the RTN signals used in this work to exclude any significant interference from time-dependent RTN variations and other noise sources. Moreover, in this work, the RTN is only considered during the testing after the neural network has been well-trained. The impact of RTN during the training process is an interesting issue for RRAM-based neural networks, especially for unsupervised learning. It is a more complex issue involving RRAM reliability such as retention and endurance, which is out of the scope of this letter.

CONCLUSIONS

In this letter, we analyzed the amplitude and occurrence rate of RTN signals at multiple resistance levels in both Ta_2O_5–based CF and TiO_2/a-Si (a-VMCO) based NCF RRAM devices. Based on the statistical experimental results, an RTN-based disturbance model is developed and applied to the trained synapses arrays to simulate its impact on the accuracy of neural networks. It is revealed that the NCF RRAM devices show smaller RTN amplitude, tighter RTN distribution, and lower RTN occurrence rate compared with its filamentary counterpart. The neural network with NCF synapses has much better pattern recognition accuracy due to negligible impact of its RTN induced disturbance, and the ANN using NCF devices needs fewer neurons and synapses to achieve better accuracy. The non-filamentary RRAM device has significant advantage as synaptic devices over the filamentary RRAM thanks to its better RTN resilience, making it a promising candidate in nanoscale neuromorphic synaptic applications.

REFERENCE

[1] D. Ielmini, "Brain-inspired computing with resistive switching memory (RRAM): Devices, synapses and neural networks," *Microelectron. Eng.*, vol. 190, pp. 44-53, Jan. 2018. DOI: 10.1016/j.mee.2018.01.009

[2] J. Woo, A. Padovani, K. Moon, M. Kwak, L. Larcher and H. Hwang, "Linking Conductive Filament Properties and Evolution to Synaptic Behavior of RRAM Devices for Neuromorphic Applications," *IEEE Electron Device Lett.*, vol. 38, no. 9, pp. 1220-1223, July 2017. DOI: 10.1109/LED.2017.2731859

[3] C. Sung, S. Lim, H. Kim, T. Kim, K. Moon, J. Song, J. -J. Kim and H. Hwang, "Effect of conductance linearity and multilevel cell characteristics of TaOx-based synapse device on pattern recognition accuracy of neuromorphic system," *Nanotechnology* vol. 29, p. 115203, Feb. 2018. DOI: 10.1088/1361-6528/aaa733

[4] C. -C. Chang, P. -C. Chen, T. Chou, I. -T. Wang, B. Hudec, C. -C. Chang, C. -M. Tsai, T. -S. Chang, and T. -H. Hou, "Mitigating Asymmetric Nonlinear Weight Update Effects in Hardware Neural Network Based on Analog Resistive Synapse," *IEEE J. Emerg. Sel. Topic Circuits Syst.*, vol. 8, no. 1, pp. 116-124, Nov. 2018. DOI: 10.1109/JETCAS.2017.2771529

[5] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. L. Gallo, K. Moon, J. Woo, H. Hwang and Y. Leblebici, "Neuromorphic computing using non-volatile memory," *Adv. Phys.: X*, vol. 2, no. 1, pp. 89–124, Nov. 2016. DOI: 10.1080/23746149.2016.1259585

[6] S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z. -Q. Wang, A. Calderoni, N. Ramaswamy and D. Ielmini, "Neuromorphic Learning and Recognition With One-Transistor-One-Resistor Synapses and Bistable Metal Oxide RRAM," *IEEE Trans. Electron Devices*,vol. 63, no. 4, 2016 DOI: 10.1109/TED.2016.2526647

[7] M. J. Kirton and M. J. Uren, "Noise in solid-state microstructures: A new perspective on individual defects, interface state and low frequency (1/f) noise," *Adv. Phys.*, vol. 38, no. 4, pp. 367-468, 1989. DOI: 10.1080/00018738900101122

[8] B. Govoreanu, G. S. Kar, Y. Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. P. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altimime, D. J. Wouters, J. A. Kittl and M. Jurczak, "10×10nm^2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation," in *IEDM Tech. Dig.*, 2011. DOI: 10.1109/IEDM.2011.6131652

[9] Z. Chai, J. Ma, W. D. Zhang, B. Govoreanu, J. F. Zhang, Z. Ji, M. Jurczak, "Probing the Critical Region of Conductive Filament in Nanoscale HfO_2 Resistive-Switching Device by Random Telegraph Signals," *IEEE Trans. Electron Devices*, vol. 64, no. 10, pp. 4099 - 4105, 2017. DOI: 10.1109/TED.2017.2742578

[10] J. Ma, Z. Chai, W. Zhang, J. Zhang, Z. Ji, B. Benbakhti, B. Govoreanu, E. Simoen, L. Goux, A. Belmonte, R. Degraeve, G. Kar and M. Jurczak, "Investigation of pre-existing and generated defects in non-filamentary a-Si/TiO_2 RRAM and their impacts on RTN amplitude distribution," *IEEE Trans. Electron Devices*, vol. 65, no. 3, pp. 970-977, Jan. 2018. DOI: 10.1109/TED.2018.2792221

[11] J. Kang, Z. Yu, L. Wu, Y. Fang, Z. Wang, Y. Cai, Z. Ji, J. Zhang, R. Wang, Y. Yang and R. Huang, "Time-dependent variability in RRAM-based analog neuromorphic system for pattern recognition," in *IEDM Tech. Dig.*, 2017. DOI: 10.1109/IEDM.2017.8268340

[12] K. Moon, A. Fumarola, S. Sidler, J. Jang, P. Narayanan, R. M. Shelby, G. W. Burr and H. Hwang, "Bidirectional Non-Filamentary RRAM as an Analog Neuromorphic Synapse, Part I: Al/Mo/$Pr0.7Ca_{0.3}MnO_3$ Material Improvements and Device Measurements," *IEEE J. Electron Devices Soc.*, vol. 6, pp. 146-155, 2017. DOI: 10.1109/JEDS.2017.2780275

[13] Y. -S. Fan, L. Zhang, D. Crotti, T. Witters, M. Jurczak and B. Govoreanu, "Direct Evidence of the Overshoot Suppression in Ta_2O_5-Based Resistive Switching Memory With an Integrated Access Resistor," *IEEE Electron Device Lett.*, vol. 36, no. 10, pp. 1027-1029, Aug. 2015. DOI: 10.1109/LED.2015.2470081

[14] B. Govoreanu, D. Crotti, S. Subhechha, L. Zhang, Y. Y. Chen, S. Clima, V. Paraschiv, H. Hody, C. Adelmann, M. Popovici, O. Richard and M. Jurczak, "a-VMCO: a novel forming-free, self-rectifying, analog memory cell with low-current operation, nonfilamentary switching and excellent variability," in *VLSI Symp. Tech. Dig.*, 2015. DOI: 10.1109/VLSIT.2015.7223717

[15] C. M. Bishop, "Pattern Recognition and Machine Learning," New York, Springer, 2006

[16] M. Nielsen, "Neural Networks and Deep Learning," Determination Press, 2015

[17] Y. Lecun, L. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998. DOI: 10.1109/5.726791

[18] Z. Chai, W. Zhang, P. Freitas, F. Hatem, J. F. Zhang, J. Marsland, B. Govoreanu, L. Goux, G. S. Kar, S. Hall, P. Chalker and J. Robertson, "The over-reset phenomenon in Ta_2O_5 RRAM device investigated by the RTN-based defect probing technique," *IEEE Electron Device Lett.*, vol. 39, no. 7, July 2018, DOI: 10.1109/LED.2018.2833149

[19] F. M. Puglisi, N. Zagni, L Larcher and P. Pavan, "A New Verilog-A Compact Model of Random Telegraph Noise in Oxide-Based RRAM for Advanced Circuit Design," in *47th European Solid-State Device Research Conference (ESSDERC), 2017*, DOI: 10.1109/ESSDERC.2017.8066627

[20] S. Balatti, S. Ambrogio, A. Cubeta, A. Calderoni, N. Ramaswamy and D. Ielmini, "Voltage-dependent random telegraph noise (RTN) in HfOx resistive RAM," in *IEEE International Reliability Physics Symposium (IRPS)*, Waikoloa, 2014, DOI: 10.1109/IRPS.2014.6861159

[21] G. W. Burr, R. M. Shelby, S. Sidler, C. Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi and H. Hwang, "Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3498-3507, Nov. 2015, DOI: 10.1109/TED.2015.2439635

[22] P. Chi, S. Li, Z. Qi, P. Gu, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, Y. Xie, "Processing-in-Memory in ReRAM-based Main Memory," SEAL-lab Technical Report, No. 2015-001, 2015

[23] S. Yu, "Neuro-Inspired Computing With Emerging Nonvolatile Memorys," *Proc. IEEE*, vol. 106, no. 2 pp. 260-285, Feb. 2018, DOI: 10.1109/JPROC.2018.2790840

[24] C. Lindsey and T. Lindblad, "Review of hardware neural networks: a user's perspective," *Proceedings of Third Workshop on Neural Networks: From Biology to High Energy Physics, Isola d'Elba, Italy, Sept.*, 1994

[25] F. M. Puglisi, A. Padovani, L. Larcher and P. Pavan, "Random Telegraph Noise: Measurement, Data Analysis, and Interpretation," in 24th IEEE International Symposium on *Physical and Failure Analysis of Integrated Circuits (IPFA)*, Chengdu, 2017. DOI: 10.1109/IPFA.2017.8060057

[26] S. Ambrogio, S. Balatti, V. McCaffrey, D. C. Wang and D. Ielmini, " Noise-Induced Resistance Broadening in Resistive Switching Memory—Part I: Intrinsic Cell Behavior," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3805-3811, 2015, DOI: 10.1109/TED.2015.2475598

[27] S. Ambrogio, S. Balatti, V. McCaffrey, D. C. Wang and D. Ielmini, " Noise-Induced Resistance Broadening in Resistive Switching Memory—Part II: Array Statistics," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3812 - 3819, 2015, DOI: 10.1109/TED.2015.2477135