Role of synaptic variability in spike-based neuromorphic circuits with unsupervised learning

D. R B Ly, A. Grossi, T. Werner, T. Dalgaty, C. Fenouillet-Beranger, E. Vianello, E. Nowak

Role of synaptic variability in spike-based neuromorphic circuits with unsupervised learning

D. R. B. Ly, A. Grossi, T. Werner, T. Dalgaty, C. Fenouillet-Beranger, E. Vianello, E. Nowak
CEA-Leti, MINATEC, 38054 Grenoble, France
denys.ly@cea.fr
elisa.vianello@cea.fr

Abstract—**Resistive Random Access Memory (RRAM)-based artificial Neural Networks (NNs) have been shown to be intrinsically robust to RRAM variability but no study has been done to clearly explain and quantify this robustness. In this paper, we fully characterize a 4kbit RRAM array under different programming conditions. The impact of the electrical characteristics of RRAM (resistance variability, memory window, endurance performance) on the detection rate of a NN designed for object tracking and trained with a stochastic Spike-Timing Dependent Plasticity (STDP) rule is studied. We introduce a new parameter called the Synaptic Window (SW), defined as the ratio between the arithmetic mean conductance values of the low and high resistance distributions. The network performance was found only to be sensitive to the value of the SW (a SW>100 is required to achieve the maximum NN performance). Moreover, we demonstrate that a high resistance variability increases the SW for a given window margin.**

I. Introduction

Synapses play a central role in two principal tasks of the brain; information processing and information storage. Of many important synaptic properties, it has been proven that synapses are noisy devices [1], [2]. In biology, variability has been observed in the postsynaptic response (if a presynaptic cell is driven repeatedly with identical stimuli, there is trial-to-trial variability in the postsynaptic response). While the purpose of such variability is not completely understood, it is generally accepted that variability might offer distinct advantages (*e.g.* energy-saving [3] or enhance sensitivity to weak signals [4]).

Resistive memory devices (RRAM) or RRAM-based circuits are promising candidates to emulate the behaviour of synapses in artificial Neural Networks (NN), and in particular to reproduce their capability to learn. The synaptic connections (RRAM devices) among neurons are created, modified, and preserved accordingly to a learning model. As a result of a *learning process* it is possible to perform pattern detection and classification. In recent years, a variety of approaches have been considered to implement learning, such as the bio-inspired Spike-Timing Dependent Plasticity (STDP [5]–[7]). STDP facilities the NN to learn the synaptic weights in an unsupervised way - without a labeled dataset or external teacher. It has been demonstrated that RRAM-based NNs are robust to synaptic variability [5], [8]. However, a clear study explaining the origin of this robustness is still missing.

In this work, a visual pattern extraction application based on a fully-connected network of Leaky-Integrate and Fire (LIF) neurons and RRAM-based synapses implementing the STDP learning rule is adopted to demonstrate the impact of RRAM electrical characteristics (variability, Memory Window, endurance) on network performance. We demonstrated that RRAM technologies with a large Memory Window (MW), defined as the ratio between the high and low resistance at 3σ of the resistance distribution, allow for improvement of the network performance. However, an increase of MW is challenging to realise with state of the art RRAM due to the large variability. Synaptic variability is beneficial in the case of NNs since it increases the dynamic range of resistance, thus reducing the constraints on the MW. The results obtained from the experimental characterization of a 4kbit RRAM array have been exploited to give general guidelines for the design of hardware-oriented neuromorphic circuits.

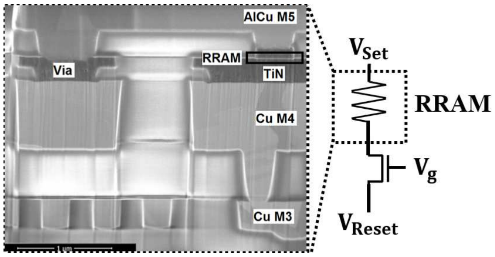
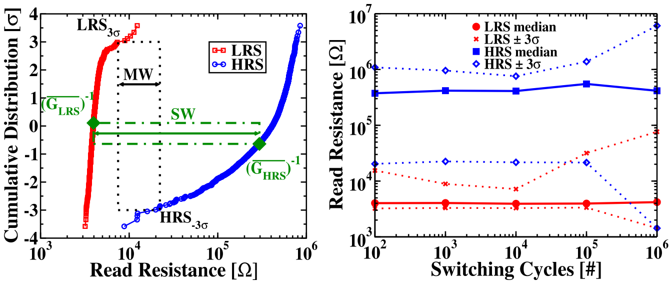


Fig. 1. (Left) SEM cross section of an integrated TiN/HfO$_2$/Ti/TiN RRAM cell between M4 and M5. (Right) Schematic view of the 1T-1R cell configuration.

II. 4kbit RRAM array fabrication and electrical characterization

A 4kbit RRAM array was fabricated using a 130 nm CMOS front-end process [9]. The RRAM devices are composed of a TiN/HfO$_2$/Ti/TiN stack, where both HfO$_2$ and Ti layers are 10 nm thick and are integrated on the top of the fourth metal layer. The cross section of a 300nm diameter RRAM device is shown in Fig. 1 (Left). The cell configuration is the 1T-1R presented in Fig. 1 (Right). RRAM devices switch between two distinct resistance states, a Low Resistance State (LRS) and a High Resistance State (HRS), when Forming/Set and Reset conditions are applied respectively. Forming and Set operations are performed by applying a voltage pulse on the RRAM top electrode (V_{Set}), whereas Reset is performed by applying a voltage pulse on the drain of the NMOS transistor (V_{Reset}). The compliance current (I_{cc}) is fixed by the voltage applied on the gate of the transistor (V_g).

Fig. 2 (a) shows the Cumulative Distribution Function (CDF) of LRS and HRS measured on the 4kbit array after 10^4 Set/Reset operations, whereas Fig. 2 (b) shows the evolution of LRS and HRS during 10^6 Set/Reset cycles. In memory applications RRAMs are used to store one bit of informa-

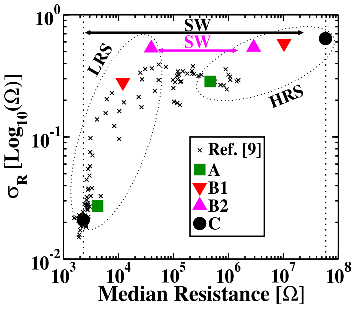


(a) (b)

Fig. 2. (a) CDF of the LRS and HRS measured on the 4kbit array after 10^4 switching cycles and (b) endurance with programming condition A in Fig. 3 (a). Variability is defined as: $\sigma_R = std[log_{10}(R)]$. No smart algorithm is applied.

Condition		A	B1	B2	C
Voltage [V]	V_{Set}	2	2	2	2
	V_{Reset}	2.5	2.5	2.5	2.5
I_{cc} [μA]		200	57	50	600
Energy [pJ/spike]	E_{Set}	40	11.4	10	120
	E_{Reset}	50	14.3	12.5	150
$\sigma_{R,LRS}$ $[\log_{10}(R)]$		0.03	0.28	0.53	0.02
$\sigma_{R,HRS}$ $[\log_{10}(R)]$		0.22	0.58	0.54	0.64
MW [#]		3	1.3	0.014	370
SW [#]		89	374	32	8800
Endurance [#]		10^6	10^4	10^7	10^2

(a)



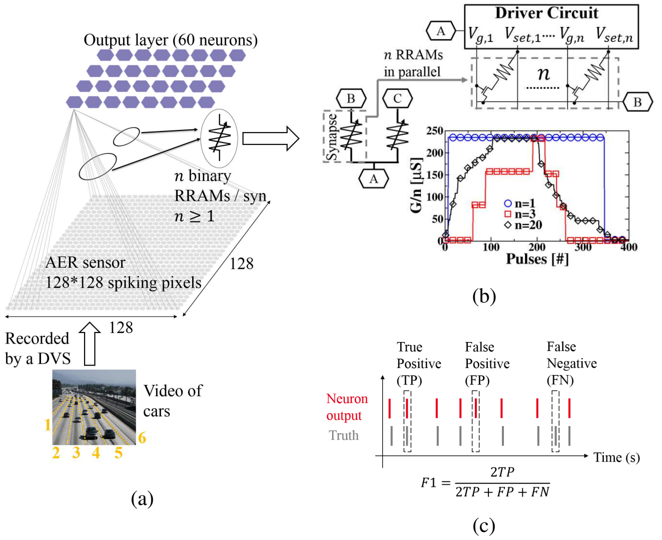
(b)

Fig. 3. (a) Programming conditions used in this work, with t_{pulse}=100 ns. (b) Variability as a function of median resistance value for different programming conditions. Experimental measurements used in this work are highlighted.

tion; therefore the most important parameter is the Memory Window (MW). The variability in the high resistance state reduces the MW, hindering the use of this technology for large memory arrays. In neuromorphic applications the RRAM devices define the weight of the connection between two neurons. Therefore they have different requirements than standard memory applications. In the following section, we will study the impact of RRAM electrical characteristics on the performance of a NN for object tracking. We introduce a new parameter, the Synaptic Window (SW), defined as the ratio between the arithmetic mean conductance values of LRS ($\overline{G_{LRS}}$) and HRS ($\overline{G_{HRS}}$). The resistance variability in both LRS and HRS is estimated as the standard deviation of the $\log_{10}$ of the resistance distribution [10].

MW, SW, endurance performance and variability of both LRS and HRS depend on the programming conditions and they cannot be decoupled (see Fig. 3 (a)). Fig. 3 (b) shows the link between resistance variability and median resistance values obtained with different programming conditions. Variability increases with median resistance value and saturates for resistance values higher than 13 kΩ [9]. Moreover, it has been demonstrated that a tradeoff exists between MW and endurance performance: higher MW implies lower endurance [11]. Fig. 3 (a) summarizes the programming conditions used in the following section to study the impact of RRAM performance on a NN by means of simulation:

- A: best compromise between endurance and MW (the most suited condition for standard memory applications);
- B1 and B2: low power consumption, high variability in both LRS and HRS and low MW (cannot be used for memory applications due to the low MW);
- C: highest MW among the 4 conditions, high power consumption and low endurance.



(a) (b) (c)

Fig. 4. (a) FCNN topology for cars tracking application, with stochastic STDP learning rule. (b) Synapses are implemented with n binary RRAMs in parallel to achieve n+1 synaptic levels. (c) Definition of F1-score to assess network performance.

III. Impact of RRAM Programming Conditions on a Neural Network for Object Tracking

A. Network Topology

We performed full system-level simulations of a visual pattern extraction application with our special purpose event-based N2D2 simulator tool [12]. The neuron circuits are modeled with behavioral equations as in [13]. The effects of both device-to-device and cycle-to-cycle variations are captured in the RRAM-based synapse model. Fig. 4 (a) presents the simulated two-layer Fully-Connected Neural Network (FCNN) topology. A video of cars passing on a six-lane wide motorway is recorded in Address Event Representation (AER) format by a Dynamic Vision Sensor (DVS) and it represents the input data [14]. The FCNN is composed of an input layer, corresponding to the DVS with 128*128 spiking pixels and an output layer of 60 neurons [13]. An input pixel generates a spike each time there is a change of luminosity of the input video. Each input pixel is connected with 2 synapses to every output neuron to denote an increase and decrease in illumination respectively. The total number of synapses is 128*128*2*60 = 1966080. A similar network has been implemented in [15] and [16] exploiting multi-level Phase-Change Memory and binary Conductive Bridge RRAM synapses, respectively. In this work, we adopted the RRAM technology presented in Section II. The training is unsupervised with a stochastic STDP rule [16]: for each LTP (LTD) event every synaptic device has a probability p_{LTP} (p_{LTD}) to be set in the LRS (reset in the HRS). After training every output neuron becomes sensitive to a specific lane and spikes whenever a car passes on this lane. The network can then be used for the detection of cars.

Fig. 4 (c) sketches the spiking activity of one output neuron (red) and the actual traffic (a grey spike corresponds to a car passing on the lane). If the neuron detects a car, we have a True Positive (TP) event. If it spikes with no car passing, we have a False Positive (FP) event. If it misses a car, we have a False Negative (FN) event. We use the F1-score as a metric to assess network performance:

$$F1 = \frac{2TP}{2TP + FN + FP} \qquad (1)$$

F1 ranges from 0 to 1, with F1 = 1 being the best performance. Each output neuron becomes sensitive to one lane. Since there are 60 output neurons and only 6 lanes, several neurons become sensitive to the same lane. As more cars pass on the lanes 4 and 5, more neurons are sensitive to these lanes than to the lane 6, the least active lane. To assess network performance, only the most sensitive neuron for each lane is considered.

The RRAM cells presented in Section II are intrinsically binary devices: they switch between two distinct states, LRS and HRS. The use of only two resistance levels per synapse, with respect to the multi-level approach, can be insufficient to achieve good performance in neuromorphic systems designed for complex applications [17]. In [10] we proposed *n* binary RRAMs operating in parallel as artificial synapse (Fig. 4 (b)). Since parallel conductances add up, the equivalent synaptic weight spreads from the sum of n condutances in HRS to n conductances in LRS, with n+1 distinct conductance levels. This allows for the implementation of an analog synapse with binary devices: when we have an LTP (LTD) event, every device has a probability p_{LTP} (p_{LTD}) to switch to the LRS (HRS). The switching probability can be governed by the RRAM itself (internal switching probability): Set and Reset conditions can be tuned to control the probability to switch the memory as shown in [18]. Another possibility, that allows finer tuning of the switching probability, at the expense of increasing the circuit complexity, consists of using stronger programming conditions (*i.e.* internal switching probability is equal to one) and extrinsic stochasticity. Extrinsic stochasticity is obtained using an external Pseudo Random Number Generator circuit block, which provides tunable switching probabilities. We used extrinsic stochasticity, with $p_{\mathrm{LTP}} = 0.13$ and $p_{\mathrm{LTD}} = 0.2$.

In the following, we will present the impact of the characteristics of RRAM-based synapses (number of synaptic levels, memory window, variability, endurance) on the network performance. In order to account for the RRAM variability, each point has been averaged over 20 simulations.

B. Impact of the RRAM-based synapse characteristics (number of synaptic levels, MW, variability) on the NN performance

First, we investigated the impact of the number of synaptic levels and the RRAM memory window on the NN performance. Fig. 5 (a) shows F1-score as a function of the MW at 3σ for different number of RRAMs per synapse. We used an artificial log normal LRS and HRS distributions with $\sigma_{\mathrm{R,LRS}} = \sigma_{\mathrm{R,HRS}} = 0.05$, and different LRS median values to vary the MW. The NN performance is independent of the number of devices per synapse. The same result was achieved with different resistance variability values (not shown). Therefore, in the following simulations we implemented every synapse with one binary RRAM. The essential parameter to improve the network performance is the MW. F1 increases with the MW and it saturates at F1≈0.96 for MW larger than 50.

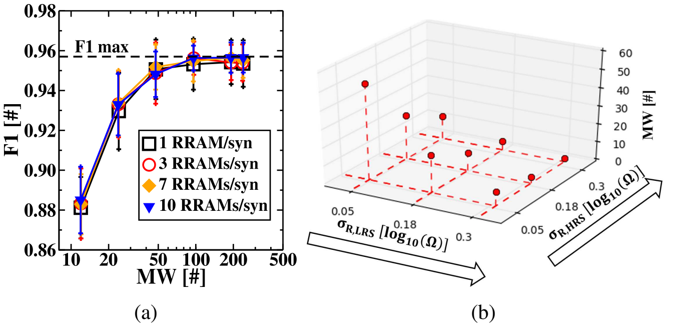
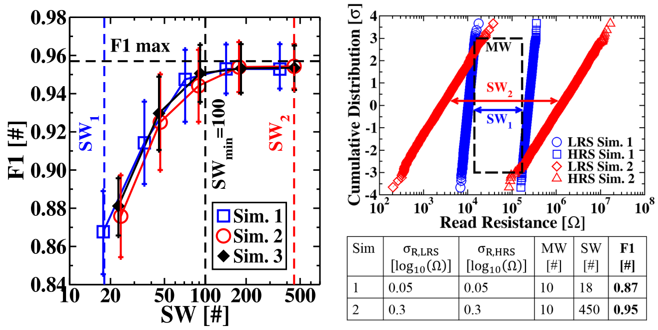


Fig. 5. (a) NN performance (F1) as a function of the MW for different number of RRAMs per synapse. (b) Minimal MW required to achieve the maximum NN performance (F1 = 0.96) as a function of LRS and HRS variability.

Second, we studied the impact of the synaptic variability. We simulated the proposed application with nine different combinations of LRS and HRS variability. Fig. 5 (b) plots the minimal MW required to reach the maximum NN performance (F1≈0.96) for the nine different cases. A smaller MW is required for the highest LRS and HRS variability values. Improving the RRAM synaptic variability is a way to relax the constraints on the minimal MW required. A high resistance variability increases the synaptic dynamic range, *i.e.* the range of synaptic values that can be reached during the training phase. The improvement of the NN performance for a large synaptic dynamic range is due to the fact that, in order to achieve high performance after the training phase, the majority of the synaptic weights have to be weak (RRAM in HRS), with a tail of stronger connections (RRAM in LRS). In order to reach the maximum NN performance (F1≈0.96), the number of RRAM cells with resistance higher than 50 kΩ, n_{OFF}, has to be 50 times more numerous than the number of synapses with resistance lower than 10 kΩ, n_{ON}.

To quantify this result we studied the impact of the SW, defined in Section II as the ratio between the arithmetic mean conductance values of LRS and HRS, on the NN performance. SW increases with both MW and variability. Fig. 6 (a) reports F1 as a function of the SW for the nine combinations of resistance variability presented in Fig. 5 (b). For the sake of clarity, only three combinations are shown, the result is the same with the nine combinations. The network performance is independent of the synaptic variability, F1 is defined by the SW. As expected, F1 increases with the dynamic range and saturates at 0.96 for SW higher than 100. For a given MW=10, an increase in both LRS and HRS variability (red curve in Fig.



Sim	$\sigma_{R,LRS}$ $[\log_{10}(\Omega)]$	$\sigma_{R,HRS}$ $[\log_{10}(\Omega)]$	MW [#]	SW [#]	**F1 [#]**
1	0.05	0.05	10	18	**0.87**
2	0.3	0.3	10	450	**0.95**

(a)

(b)

Fig. 6. (a) NN performance (F1) as a function of the SW for 3 different LRS and HRS variability combinations of Fig. 5 (b). Sim. 1: $\sigma_{R,LRS,1}$ = $\sigma_{R,HRS,1}$ = 0.05 ; Sim. 2: $\sigma_{R,LRS,2}$ = $\sigma_{R,HRS,2}$ = 0.3 ; Sim. 3: $\sigma_{R,LRS,3}$ = 0.05 and $\sigma_{R,HRS,3}$ = 0.3. F1 only depends on the SW. (b) CDF of LRS and HRS used for the lowest (Sim. 1, blue) and highest (Sim. 2, red) resistance variability combinations, for MW=10. Variability increases the SW and the corresponding NN performance (F1).

6 (b)) allows to increase the SW and consequently the F1-score of about 10%, with respect to the case with low variability (blue curve in Fig. 6 (b)).

Finally, we linked the experimental data measured on the 4kbit array in Section I with the simulation results. Fig. 7 (Top) reports F1 for the four studied programming conditions (Fig. 3). It is worth noting that even with weak programming conditions (B1, F1=0.953) we have a score as good as with strong ones (C, F1=0.959). Condition B1 works well for neuromorphic applications whereas it cannot be used in a memory application due to its high LRS variability. However, for condition B2, RRAM works neither for memory nor neuromorphic applications. A decrease in F1 is observed with standard programming conditions (A) but is still acceptable (F1=0.946) if we can tolerate a loss of performance with an increase in endurance. These results confirm that the most important requirement for the RRAM-based synapses is a large dynamic range while resistance variability is less critical.

C. Impact of the RRAM-based synapse characteristics on the learning time

We studied the impact of the RRAM programming conditions on the learning time. The learning time has been defined as the time at which F1 reaches its maximal value ±1% for a given RRAM programming condition. Fig. 7 (Bottom) reports the learning time for the four studied programming conditions. Learning time is degraded only for the condition B2 with reduced SW.

D. Impact of the RRAM aging with endurance on the NN performance

We extracted from the simulations the number of switching events during the learning phase. For A, B1 and C a synapse undergoes a maximum of 20 Set and 40 Reset pulses in the worst case. With an endurance ranging from 10^2 cycles (C) up to 10^6 (A) we could truly consider using these programming conditions for a real hardware neural network.

Another potential problem is the evolution of the MW and variability during endurance. We extracted the resistance distribution during cycling for the condition A (Fig. 2 (b)) and we used these data to evaluate the impact of RRAM aging on F1. The results are shown in Fig. 8. We can maintain a constant F1-score of 0.95 until 10^5 cycles after which F1 plummets. The degradation of F1 at 10^6 cycles is not due to the increase in resistance variability and decrease of MW but to the dead cells (broken cells stuck in LRS).

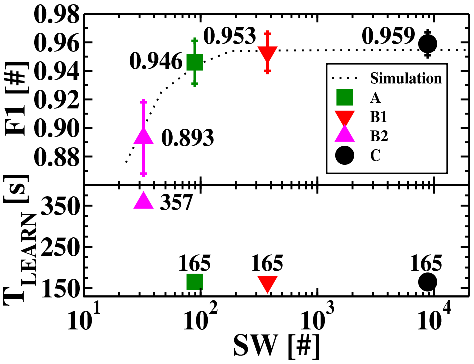


Fig. 7. (Top) F1 and (Bottom) learning speed as a function of the SW for the four programming conditions of Fig. 3.

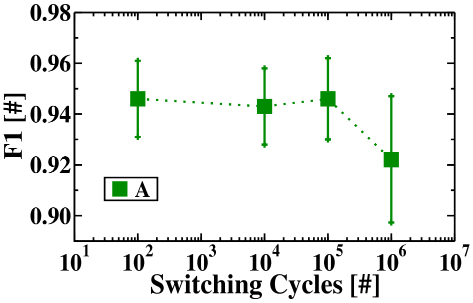


Fig. 8. Impact of the RRAM aging during endurance on F1. Simulations have been calibrated using the data of Fig. 2 (b).

IV. Conclusion

In this paper we provide guidelines to program RRAM-based synapses in a NN for object tracking applications. Multilevel conductance is not necessary, *i.e.* binary synapses are sufficient. We clarified the role played by synaptic variability and the robustness to variability. To achieve high performance after the training phase, the majority of the synaptic weights have to be weak (RRAM in a HRS), with a tail of stronger connections (RRAM in LRS). Consequently, a large RRAM dynamic range is required. Resistance variability increases the dynamic range for a given MW. We introduced a new parameter, the Synaptic Window, which takes into account both MW and variability. The network performance was found only to be sensitive to the value of the SW. A SW > 100 is required to achieve the maximum NN performance.

Acknowledgments

This work has been partially supported by the h2020 NeuRAM^3 project.

References

[1] C. Allen and C. F. Stevens, "An evaluation of causes for unreliability of synaptic transmission." *Proc Natl Acad Sci U S A*, vol. 91, no. 22, pp. 10 380–10 383, 1994.

[2] N. A. Hessler, A. M. Shirke, and R. Malinow, "The probability of transmitter release at a mammalian central synapse," *Nature*, vol. 366, no. 6455, pp. 569–572, 1993.

[3] M. S. Goldman, "Enhancement of information transmission efficiency by synaptic failures," *Neural Computation*, vol. 16, no. 6, pp. 1137–1162, 2004.

[4] M. D. McDonnell and D. Abbott, "What is stochastic resonance? definitions, misconceptions, debates, and its relevance to biology," *PLOS Computational Biology*, vol. 5, no. 5, pp. 1–9, 05 2009.

[5] E. Vianello, T. Werner, O. Bichler, A. Valentian, G. Molas, B. Yvert, B. D. Salvo, and L. Perniola, "Resistive memories for spike-based neuromorphic circuits," in *2017 IEEE International Memory Workshop*, May 2017, pp. 1–6.

[6] D. Querlioz, O. Bichler, A. F. Vincent, and C. Gamrat, "Bioinspired programming of memory devices for implementing an inference engine," *Proceedings of the IEEE*, vol. 103, no. 8, pp. 1398–1416, Aug 2015.

[7] M. R. Azghadi, B. Linares-Barranco, D. Abbott, and P. H. W. Leong, "A hybrid cmos-memristor neuromorphic synapse," *IEEE Transactions on Biomedical Circuits and Systems*, pp. 434–445, 2017.

[8] D. Garbin, E. Vianello, O. Bichler, M. Azzaz, Q. Rafhay, P. Candelier, C. Gamrat, G. Ghibaudo, B. DeSalvo, and L. Perniola, "On the impact of oxram-based synapses variability on convolutional neural networks performance," in *Proceedings of the 2015 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH15)*, 2015, pp. 193–198.

[9] A. Grossi, E. Nowak, C. Zambelli, C. Pellissier, S. Bernasconi, G. Cibrario, K. E. Hajjam, R. Crochemore, J. F. Nodin, P. Olivo, and L. Perniola, "Fundamental variability limits of filament-based rram," in *2016 IEEE International Electron Devices Meeting*, 2016, pp. 4.7.1–4.7.4.

[10] D. Garbin, E. Vianello, O. Bichler, Q. Rafhay, C. Gamrat, G. Ghibaudo, B. DeSalvo, and L. Perniola, "Hfo2-based oxram devices as synapses for convolutional neural networks," *IEEE Transactions on Electron Devices*, vol. 62, no. 8, pp. 2494–2501, 2015.

[11] E. Vianello, O. Thomas, G. Molas, O. Turkyilmaz, N. Jovanovi, D. Garbin, G. Palma, M. Alayan, C. Nguyen, J. Coignus, B. Giraud, T. Benoist, M. Reyboz, A. Toffoli, C. Charpin, F. Clermidy, and L. Perniola, "Resistive memories for ultra-low-power embedded computing design," in *2014 IEEE International Electron Devices Meeting*, 2014, pp. 6.3.1–6.3.4.

[12] O. Bichler, D. Roclin, C. Gamrat, and D. Querlioz, "Design exploration methodology for memristor-based spiking neuromorphic architectures with the xnet event-driven simulator," in *2013 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, 2013, pp. 7–12.

[13] O. Bichler, D. Querlioz, S. J. Thorpe, J. P. Bourgoin, and C. Gamrat, "Unsupervised features extraction from asynchronous silicon retina through spike-timing-dependent plasticity," in *The 2011 International Joint Conference on Neural Networks*, 2011, pp. 859–866.

[14] T. Delbruck, "Dvs128 dynamic vision sensor silicon retina data," 2008. [Online]. Available: http://sensors.ini.uzh.ch/databases.html

[15] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction," in *2011 International Electron Devices Meeting*, Dec 2011, pp. 4.4.1–4.4.4.

[16] M. Suri, O. Bichler, D. Querlioz, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Cbram devices as binary synapses for low-power stochastic neuromorphic systems: Auditory (cochlea) and visual (retina) cognitive processing applications," in *2012 International Electron Devices Meeting*, 2012, pp. 10.3.1–10.3.4.

[17] J. Bill and R. Legenstein, "A compound memristive synapse model for statistical learning through stdp in spiking neural networks," *Front Neurosci*, vol. 8, p. 412, 2014.

[18] T. Werner, E. Vianello, O. Bichler, A. Grossi, E. Nowak, J. F. Nodin, B. Yvert, B. DeSalvo, and L. Perniola, "Experimental demonstration of short and long term synaptic plasticity using oxram multi k-bit arrays for reliable detection in highly noisy input data," in *2016 IEEE International Electron Devices Meeting*, 2016, pp. 16.6.1–16.6.4.