Extended Abstracts of the 2016 International Conference on Solid State Devices and Materials, Tsukuba, 2016, pp121-122

B-7-05

Exploitation of RRAM variability to improve on-line unsupervised learning in small-scale Spiking Neural Networks

Thilo Werner[1], Elisa Vianello[1], Olivier Bichler[2], Blaise Yvert[3], Barbara DeSalvo[1] and Luca Perniola[1]
[1] CEA, LETI, Minatec Campus, Grenoble, France [2] CEA, LIST, F-91191 Gif-sur-Yvette, France [3] INSERM, Grenoble, France
E-mail: thilo.werner@cea.fr

Abstract

The impact of LRS and HRS RRAM variability on the classification rates of fully unsupervised Spiking Neural Networks is studied. An increased performance of 5-15% has been achieved assuming an increased LRS variability for two independent applications. HRS variability is found to slightly reduce network performances.

1. Introduction

Several studies among the last decade have demonstrated the potential of non-von-Neumann computing paradigms for the analysis of vast amounts of complex data. Therefore, so-called neuromorphic (brain-inspired) networks are designed specifically for certain applications and mostly simulated using conventional computers. Physical implementations of spiking neural networks (SNN) are urgently needed to achieve the next level of computing in the Big Data era. Synapse implementations will play a key role in this regard as they typically outnumber the neurons of an SNN. Requirements such as high integration density, CMOS process compatibility, low power and high lifetime caused resistive RAM (RRAM) to be one of the main candidates. An inherent drawback of RRAM is its variability. We previously demonstrated that RRAM based SNN's based on off-line supervised learning using a back propagation algorithm are strongly tolerant to resistance variability [1].

In this paper, we study the impact of RRAM variability on SNN's using on-line unsupervised learning, where the RRAM resistance status (synaptic weight) is tuned in-situ using probabilistic Spike-Timing-Dependent-Plasticity (STDP) [2]. STDP allows learning the synaptic weight in an unsupervised way and it is particularly useful when the input data is not known a priori. General guidelines for the design of hardware oriented neuromorphic circuits are extracted.

2. Experimental

1T-1R RRAM [3] devices, integrated in standard 65nm CMOS technology, were used for this study (Fig.1). Fig.2 shows the experimental distributions of Low Resistive State (LRS) and High Resistive State (HRS) varying the current compliance during the forming and set operations (CC). As CC is reduced, for both LRS and HRS distributions, the variability increases, while the mean value is shifted to higher values. The advantage of a low CC is the reduced power consumption (PC) during programming (Fig.3) and reading (higher LRS and HRS). Moreover, it allows for low current design. The drawback is the reduced resistance window (RW) between LRS and HRS due to enhanced variability. HRS can be slightly shifted up by higher reset voltages (V_R), rising the RW [4], however inducing degradation in the oxide leading to much higher failure rates with respect to lower V_R (Fig.4). Up to 1G cycles can be achieved using optimized V_R (Fig.5). Good endurance is a fundamental requirement for networks using STDP learning, especially when the input data is not known a priori and therefore the network is in permanent learning mode, i.e. rapidly increasing the number of set and reset events to adapt the synaptic weights.

3. System-level SNN simulations

Two SNN's featuring simplified probabilistic STDP and RRAM based synapses were used for a systematic study of the impact of RRAM variability on the classification rates. Simulations are performed using a special purpose event-driven simulator 'Xnet' [5].

(i) Fig.6 shows the SNN simulated to process temporally encoded video data, recorded directly from an artificial silicon retina [6]. A video of cars passing on a freeway recorded in Address Event Representation (AER) format is presented to a two-layered SNN. In each layer, every input is connected to every output by a single RRAM synapse [7]. To study the impact of the LRS and HRS variability, the synapses are modeled assuming different RRAM test cases (Fig.7): high LRS variability, high HRS variability (C1), low LRS variability, high HRS variability (C2), no LRS variability, no HRS variability (ideal binary device, C3) and high LRS variability, low HRS variability (C4). The test case C4 allows to achieve the highest overall Recognition Rate (RR) (Fig.8(a)) which can be attributed to an improved RR of lane 1 and 6 of approx. 10% (Fig.8(b)). Since line 1 and 6 are at the edge of the AER sensor, they typically have a low RR due to weak input activity [7]. The LRS variability helps us to improve the RR at these lanes. Fig.9 represents the contour plot of the RR as a function of LRS and HRS variability. Whereas the first slightly decreases the RR, the latter strongly improves the RR.

(ii) Fig.10 presents the SNN used to classify spike waveforms from neurological data [8]. Biological data is filtered and encoded using 32 band-pass filters (2nd order Butterworth). The processed data is then presented to a 2- layer SNN with 160 RRAM based synapses. This more complex problem of classification of overlapping input patterns requires analog synapses. To this purpose we adopted multiple RRAM cells to form one synapse (Fig.11) [9]. The number of devices-per-synapse (N) was varied from 1 to 100 for three RRAM test conditions (Fig.12). The recognition rate strongly increases up to N=20 and increases only slightly for N > 20 (Fig.13). Fig.14 represents the contour plot of the RR as a function of LRS and HRS variability for N=50. As in the previous example, the RR seems to decrease with HRS and increase with LRS variability.

4. Conclusions

We presented the impact of the RRAM variability on two fully unsupervised neuromorphic systems. RRAM devices are programmed as binary synapses and a stochastic-STDP learning rule is adopted. Two different applications were demonstrated: (i) visual pattern extraction and (ii) classification of biological data. We demonstrated that LRS variability allows to improve the recognition rate up to 15%, while HRS variability slightly degrades the network performances (few percent) in both applications. High LRS variability can be achieved using low programming currents (few μA). Moreover, operating at low compliance current values reduces the power consumption during learning and reading (thanks to higher LRS values). Improving HRS variability is more complex and it requires engineering the RRAM cell.

Acknowledgements

The authors would like to thank STMicroelectronics for providing the OxRAM devices for this study. This work has been partially supported by the PANACHE project.

References

[1] E. Vianello, et al., ECS Transactions 69(3), 3-10, 2015.
[2] M. Suri et al., IEEE T-ED, 60(7), 2402-2409, 2013.
[3] H.S.P. Wong, et al., Proceedings of the IEEE 100(6), 1951-1970, 2012.
[4] E. Vianello, et al., IEDM, 2014.
[5] O. Bichler et al., NANOARCH, 2013.
[6] P. Lichtsteiner et al., IEEE Solid-State Circuits, 43(2), 566-576, 2008.
[7] D. Garbin et al., IEEE-NANO, 2013.
[8] T. Werner et al., ISCAS, 2016.
[9] D. Garbin et al., IEEE T-ED, 62(8), 2494-2501, 2015.

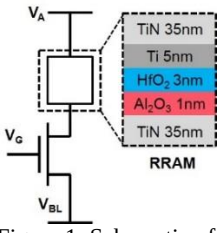

Figure 1: Schematic of 1-Transistor-1-Resistor (1T1R) co-integration used for this study.

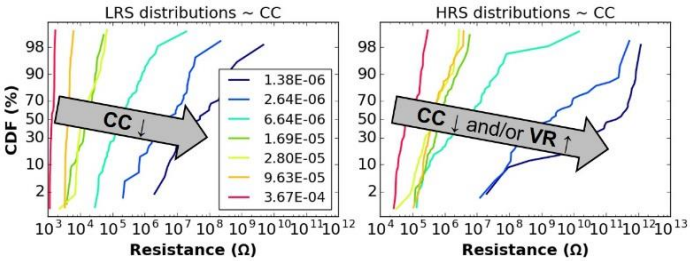

Figure 2: Cumulative distributions of Low Resistive State (LRS) and High Resistive State (HRS) as function of the current compliance (CC) for RRAM of Fig.1. Note the shift and widening of both LRS and HRS CDF for reduced CC.

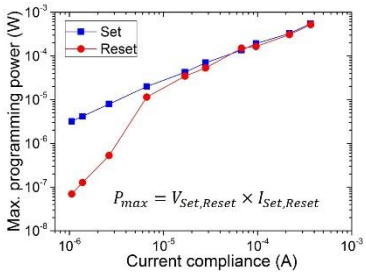

Figure 3: Estimation of maximum programming power of RRAM as a function of the current compliance.

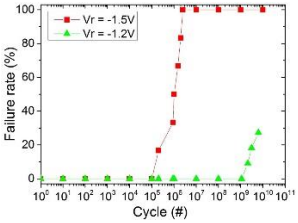

Figure 4: Endurance failure rate of RRAM as a function of the reset voltage V_R. Early HRS failure rate is induced by high V_R.

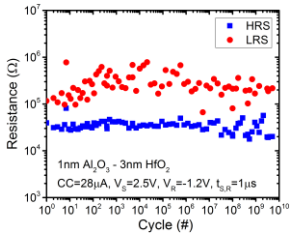

Figure 5: Endurance performance of RRAM. Note the extended endurance thanks to low V_R.

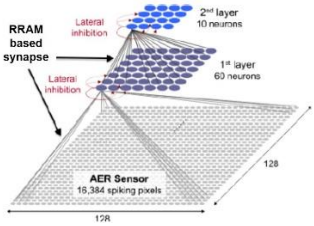

Figure 6: Spiking Neural Network (SNN) used to detect cars on a highway video sequence and automatically extract lane trajectories.

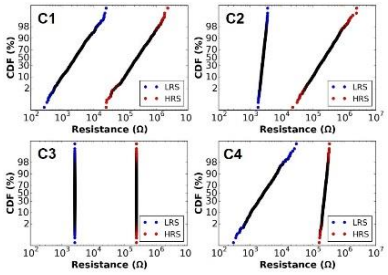

Figure 7: LRS and HRS distributions of test conditions for SNN of Fig.6.

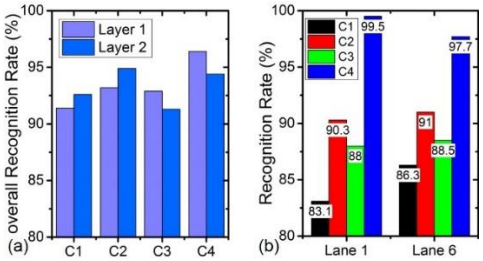

Figure 8: (a) Overall Recognition Rate (RR) for the test conditions C1-C4 of Fig.7. The RR has been computed for the first (light blue) and the second (dark blue) layers (b) RR for lanes 1 and 6. Note the high RR for C4.

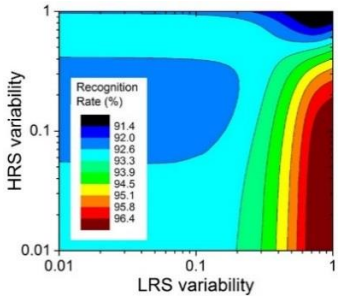

Figure 9: Recognition rate of car detection SNN (Fig.6) as a function of LRS and HRS variability.

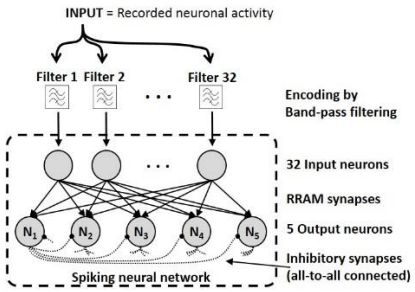

Figure 10: Spiking Neural Network (SNN) used to detect and classify neural spikes (=action potentials). Synapses are based on a multi-cell concept in order to achieve multiple synaptic weights.

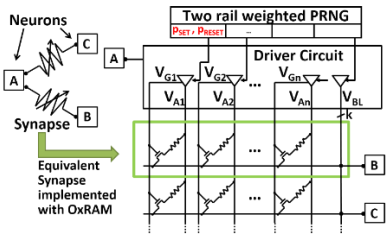

Figure 11: Multi-cell synapse concept. Each equivalent synapse consists of a series of RRAM devices, i.e. the corresponding synaptic weight is the sum of device conductances. A pseudo random number generator (PRNG) is used to enable gradual tuning overcoming the typical abrupt switching characteristic of RRAM.

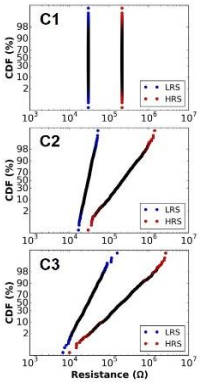

Figure 12: LRS and HRS distributions of test conditions for SNN of Fig.10.

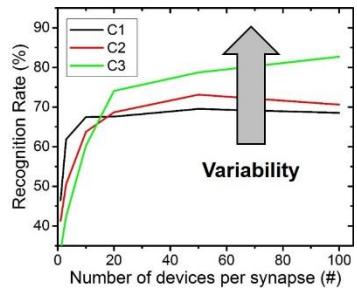

Figure 13: Overall Recognition Rate of spike sorting SNN as a function of number of devices per synapses and for different conditions C1-C3.

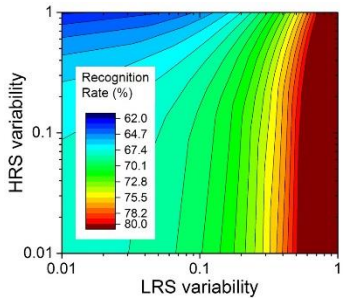

Figure 14: Recognition rate of SNN used for neural spike classification as a function of LRS and HRS variability.