Low Conductance State Drift Characterization and Mitigation in Resistive Switching Memories (RRAM) for Artificial Neural Networks

Andrea Baroni, Artem Glukhov, Eduardo Pérez, Christian Wenger, Daniele Ielmini, *Fellow, IEEE,* Piero Olivo and Cristian Zambelli, *Member, IEEE*

***Abstract*—The crossbar structure of Resistive-switching random access memory (RRAM) arrays enabled the In-Memory Computing circuits paradigm, since they imply the native acceleration of a crucial operations in this scenario, namely the Matrix-Vector-Multiplication (MVM). However, RRAM arrays are affected by several issues materializing in conductance variations that might cause severe performance degradation. A critical one is related to the drift of the low conductance states appearing immediately at the end of program and verify algorithms that are mandatory for an accurate multi-level conductance operation. In this work, we analyze the benefits of a new programming algorithm that embodies Set and Reset switching operations to achieve better conductance control and lower variability. Data retention analysis performed with different temperatures for 168 hours evidence its superior performance with respect to standard programming approach. Finally, we explored the benefits of using our methodology at a higher abstraction level, through the simulation of an Artificial Neural Network for image recognition task (MNIST dataset). The accuracy achieved shows higher performance stability over temperature and time.**

***Index Terms*—RRAM, Neural Networks, Reliability, Low Conductance states, Drift**

I. Introduction

THE last decade exposed applications such as Machine Learning (ML) and Artificial Neural Networks (ANN) to be of paramount importance in many scenarios (i.e., image recognition, biomedical analysis, data analytic, etc.) [1], [2]. In state-of-the-art Von Neumann computing architectures, those tasks are executed by Central Processing Units (CPU) and Graphics Processing Units (GPU), although it is ultimately proved that their performance and energy features are threatened by the constant data shuttling between the information processing and memory units. A revolution in computing architecture then materialized in the In-Memory Computing (IMC) concept, that has risen as one of the most promising candidates for next-generation computing thanks to its high offered throughput, low energy and good scaling [3], [4]. The technology enabler for IMC architectures has been identified in high density crossbar arrays based on non-volatile memory devices (see Fig.1a), among which stands out the resistive-switching non-volatile memory (RRAM) [5]–[7]. Crosspoint arrays of RRAM elements are in fact able to achieve massive parallelism in performing Matrix-Vector-Multiplication (MVM) through the application of the Ohm's and Kirchoff's physical laws in the analog domain [8]–[11].

However, despite the evident attractive properties, these devices have physical limitations that can have a tremendous impact on the performance of many ML and ANN tasks. Among them, the limited tunability of the conductance levels in the RRAM devices is one of the most tedious issues exposed in the accelerators based on this technology. Studies in literature evidenced that the sources are to be found in the Device-to-Device (D2D) and the Cycle-to-cycle (C2C) variations [12], the Random Telegraph Noise (RTN) [13]–[15], and the conductance drift [16]–[18], which impair the Multi-Level Conductance (MLC) capability of the RRAM technology.

An approach to overcome those limitations, relies on the application of program/verify techniques to accurately set the RRAM in a desired conductance state [19], although the stochastic nature of the technology questions their effectiveness. This calls for algorithms optimization at many levels [20]. Our approach proposed in [18], addressed both the short and the long-time scale drift of the low conductance states by exercising either a "refresh"-like technique or a combined Set (the operation to bring the cells to a high conductive state)/Reset (the operation to bring the cells to a low conductive state). The latter approach yielded to significant improvements in the distribution variability control while countering the drift.

In this work, we start from the preliminary analysis performed in [18] and extend the discussion towards the assessment of the benefits in using "drift-safe" programming algorithms at application level. In an attempt to better understand the reliability of the proposed algorithm with respect to temperature, we tracked the behavior of the low conductance states drift during a 168 hours retention experiment performed at different temperatures up to 125 °C. Finally, we project the results of the electrical characterization performed on 4 kbits RRAM arrays in the context of ANN. We will study the

A. Baroni, E.Pérez and C. Wenger are with IHP – Leibniz Institut fur innovative Mikroelektronik, 15236 Frankfurt (Oder), Germany, E-mail: baroni.ihp-microelectronics.com.

C. Wenger is also with BTU Cottbus-Senftenberg, 03046 Cottbus, Germany

A. Glukhov and D. Ielmini are with the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano and IU.NET, 20133, Milano, Italy.

C. Zambelli and P. Olivo are with the Dipartimento di Ingegneria, Universita degli Studi di Ferrara, Ferrara, 44122 Italy.

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement no. 101007321. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Belgium, Czech Republic, Germany, Italy, Sweden, Switzerland, Turkey.

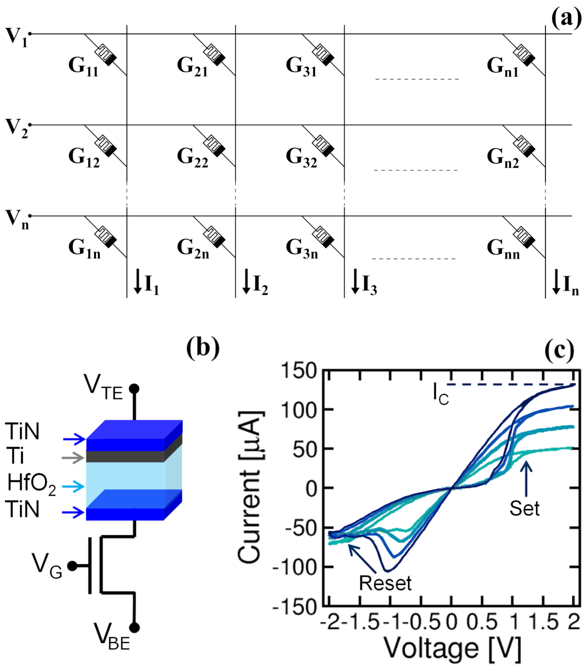

Fig. 1. (a) Representations of the crossbar structure (b) Schematic of a 1T1R RRAM device integrated in the 4 kbits array used in this work. (c) I-V characteristics of a 1T1R RRAM device measured for increasing V_G proving MLC capability.

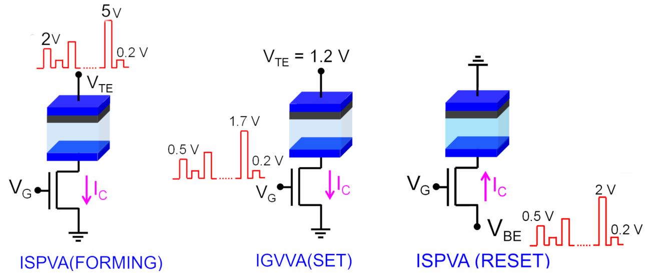

Fig. 2. Depiction of the ISPVA and IGVVA algorithms applied for Forming, Set and Reset operations used in this work [18].

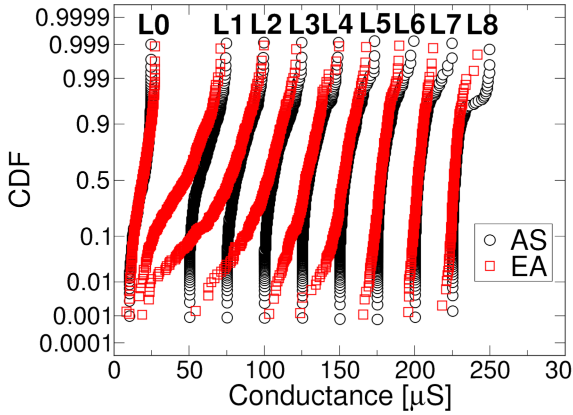

Fig. 3. Evidence of the conductance distributions drift in RRAM arrays. The AS to EA time delay is in the range of ten minutes [18].

drift-induced recognition accuracy degradation proving that our MLC algorithm provides superior results in countering the phenomenon.

II. RRAM DEVICE AND ARRAY CHARACTERISTICS

The RRAM devices considered in this work are based on the 1T1R structure depicted in Fig.1b, consisting of a TiN/Ti/HfO$_2$/TiN stack. The memristive element is formed by a 150 nm TiN top and bottom electrodes deposited by magnetron sputtering, a 7 nm Ti layer (under the TiN top electrode), and an 8 nm HfO$_2$ switching layer grown by atomic layer deposition (ALD) [21]. Every RRAM cell is selected by a n-channel MOS, manufactured in 0.25 μm CMOS technology from IHP Microelectronics. Fig.1c shows the current-voltage (I-V) characteristics of an RRAM device in the array for increasing compliance current (I_C), suggesting a controllable multi-level conductance operation by tuning I_C via gate voltage V_G. The devices are arranged in a 4 kbits crossbar array featuring 64 wordlines and 64 bitlines. All the experiments were performed on quad flat packaged (QFP) devices.

All the RRAM devices in the array are prepared for conductance switching through a Forming operation with the Incremental Step Pulse program and Verify Algorithm (ISPVA) [19]. The gate voltage V_G is set to 1.4 V and the top electrode voltage V_{TE} is gradually increased from 2 V to 5 V in steps of 10 mV. The target conductance for the operation has been chosen as 200 μS to avoid excessive stress on the RRAM cells. After the Forming, we performed a Reset operation to bring all the cells to the lowest conductance state, namely L0 at 25 μS. The Reset use the ISPVA in which the bottom electrode voltage V_{BE} is swept from 0.5 V to 2 V with 100 mV steps. The V_G is set to 2.7 V to ensure a high I_C required to disrupt the conductive filament in the RRAM cell.

III. EXPOSING THE LOW CONDUCTANCE STATES DRIFT

A. Set-based MLC operation

The standard approach used so far to achieve accurate MLC programming of the 4 kbits RRAM array was through a controlled Set operation. The Incremental Gate Voltage and Verify Algorithm (IGVVA) proven superior capabilities in conductance distribution placement [20]. In this work, the gate voltage is gradually incremented from 0.5 V to 1.7 V with 10 mV steps, featuring 1 μs pulse duration. Both the rise and the fall time of the pulses are set to 100 ns. The V_{TE} is chosen to be 1.2 V, granting reliable Set operation. With such approach, we obtained eight linearly spaced conductance levels (L1-L8) between 50 μS and 225 μS. The IGVVA characteristics are depicted in Fig.2 along with the ISPVA counterpart used for Forming and Reset operations.

Fig.3 shows that aside from the L0 distribution there is a significant drift of the L1-L4 distributions occurring in the time elapsed between the After Switching (AS) point and the End Algorithm (EA) point. We defined the former time as the time in which the target conductance is reached by the IGVVA and the latter one as the moment where the algorithm ends for all the cells programmed in the array (i.e., the last readout of the cells). By considering a population under test of 1024 RRAM

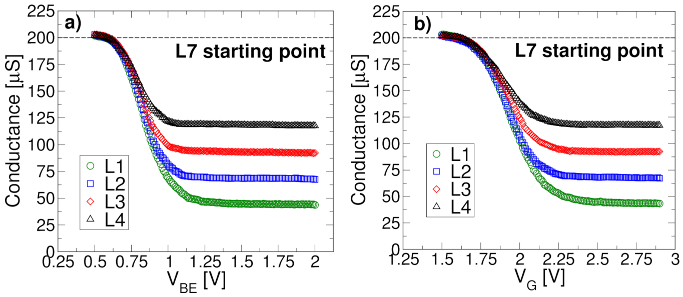

Fig. 4. (a) G-V characteristics measured during the ISPVA Reset. (b) G-V characteristics measured during the IGVVA Reset [18].

cells for each distribution, we experience an AS to EA delay time of about ten minutes. Interestingly, we observe that the L1-L4 conductance levels are the most affected by the drift, exhibiting a large fraction of the cells (≥ 50% in some cases) with their conductance falling well below their desired target G_{trg}.

B. MLC with Mixed algorithm: a ”drift-safer” approach

In our previous work [18], we explored different solutions to cope with the drift issue of the L1-L4 distributions. The first attempt conceived the application of a Refresh-like technique [22]: a selective re-application of the IGVVA algorithm was performed on the cells that show a conductance value falling below their G_{trg}. However, such approach turned out to be poorly effective since after the second IGVVA round all the EA distributions returned almost to their preliminary status. The second attempt explored an alternative algorithm to achieve L1-L4 distributions. Instead of starting from L0 distribution and apply an IGVVA in Set to reach them, we start from the L7 distribution and reach L1-L4 through a controlled Reset operation. We named this approach as Mixed algorithm since it embodies two different switching operations of the RRAM cells in the array. In Fig.4, we compared the switching dynamics from L7 to L1-L4 distributions obtained with an ISPVA Reset approach with respect to that achieved with an IGVVA Reset. The latter shows a smoother trend in reaching the desired G_{trg}. This justifies the choice of the IGVVA Reset approach in the Mixed algorithm. To avoid the over-stress of the device, we performed experiments with a V_{BE} set to 1.2 V and sweeping V_G from 1.5 V to 2.9 V in steps of 10 mV. An argument could arise in the choice of the IGVVA as Reset mechanism, since it would lead to improper results ascribed to the fact that after the operation the transistor might go in triode region since the current becomes very low, potentially damaging the RRAM cell. With our RRAM devices this is unlikely to happen because the V_G = 1.5 V lower bound is high enough. We also explored the possibility of addressing the minor drift in L5 and L6 conductance levels (see Fig.3) with the Mixed method. Unfortunately, the results have discouraged this approach since a higher variability compared with the standard Set method is experienced, as shown in Fig.5.

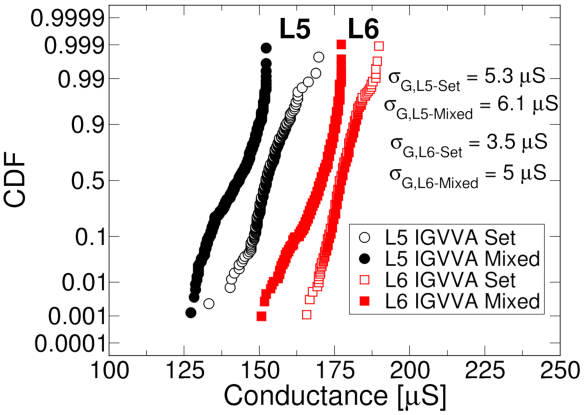

Fig. 5. L5-L6 distributions comparison when either Set MLC or Mixed algorithm is used. Results evidence that for higher conductance levels the latter method leads to higher σ_G.

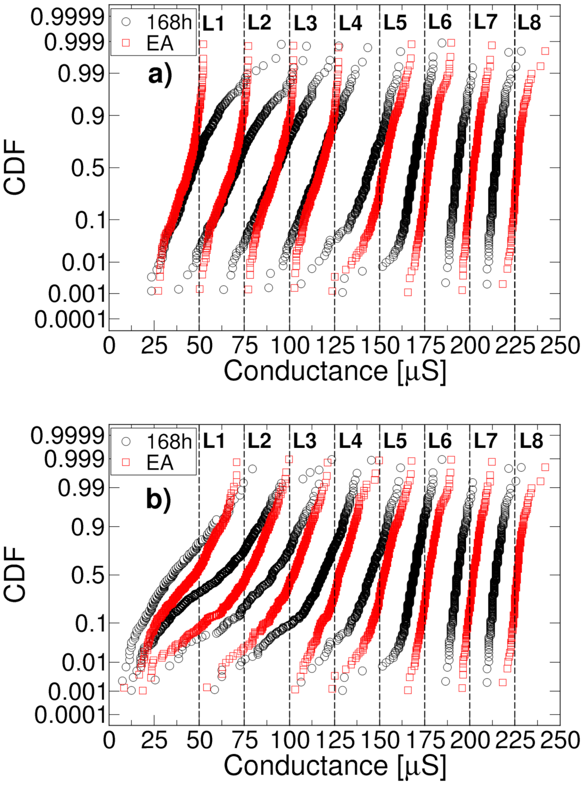

Fig. 6. (a) Drift measured after EA and at the end of a 168 hours room temperature experiment for L1-L4 distributions obtained through Mixed algorithm and L5-L8 with Set MLC. (b) Same measurement but with L1-L8 obtained all with Set MLC [18].

C. Preliminary characterization of drift in RRAM arrays

To understand whether the Mixed MLC algorithm for L1-L4 can be beneficial also for long term reliability, we performed a room temperature data retention test where we progressively monitored the conductance distributions of the RRAM arrays in a 168 hours test. The readout times have been fixed to 1, 2, 5, 9, 24, 48, 72, 96 and 168 hours after EA. We added the L5-L8 distributions to the study reminding that those are obtained with the standard Set methodology. Fig.6a shows that at the end of the data retention period the L1-L4 distributions

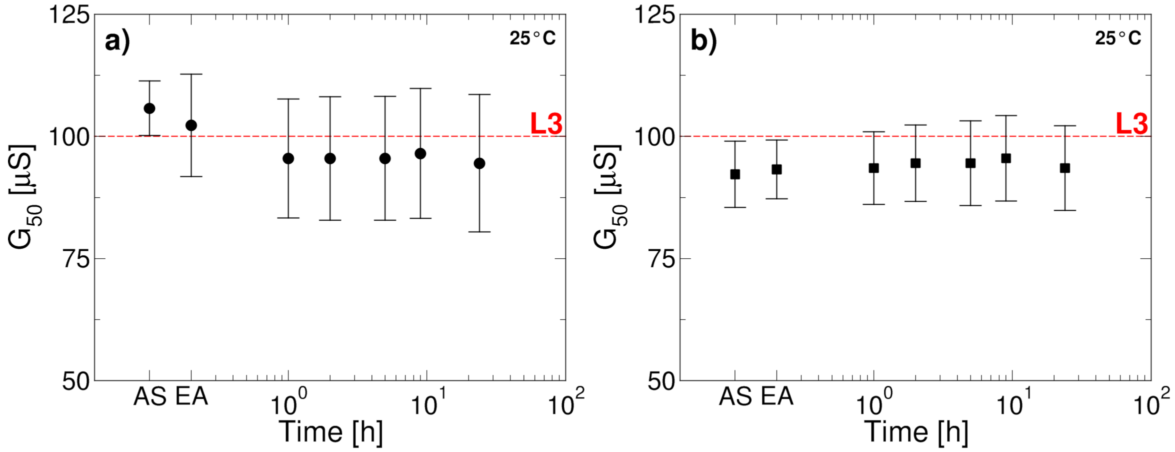

Fig. 7. (a) Evolution of the L3 level distribution (error bars indicate the standard deviation σ_G) obtained with Set MLC during room temperature experiment in the first 24h. (b) Same measurement but with L3 obtained with the Mixed algorithm.

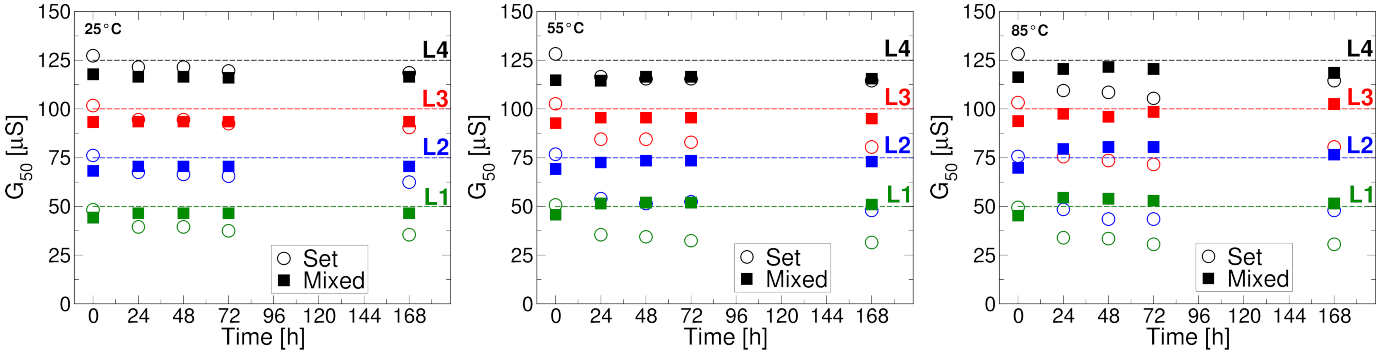

Fig. 8. Evolution of the G_{50} parameter of L1-L4 distributions for both MLC approaches. From the left to the right, we can appreciate the behavior of the distribution at 25°C, 55°C and 85°C. The Mixed approach shows enhanced stability compared to the standard Set MLC, both in terms of time and temperature.

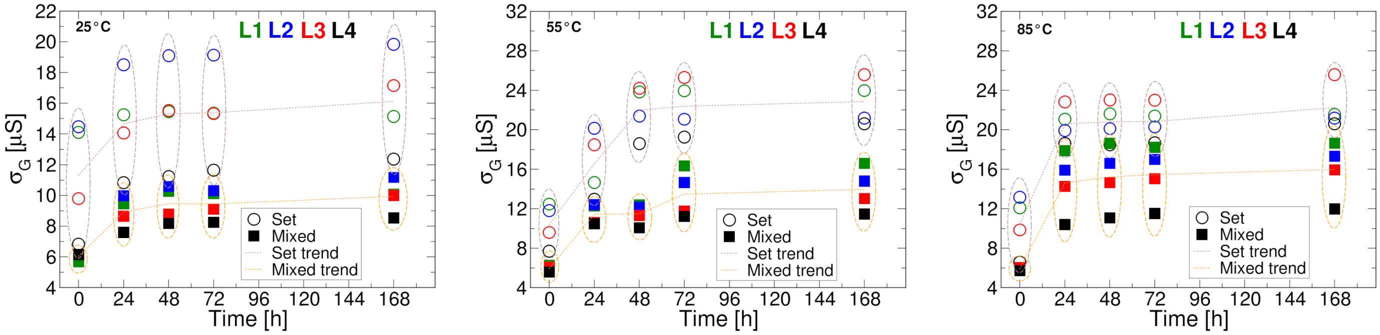

Fig. 9. Evolution of the σ_G parameter of L1-L4 distributions for both MLC approaches. With ”trend”, we indicate the mean value of the σ_G. From the left to the right, we can appreciate their behavior at 25°C, 55°C and 85°C. The Mixed approach show lower variations compare to the Set method, in both Time and Temperature.

obtained with the Mixed algorithm interestingly drifts towards $G > G_{trg}$ whereas the L5-L8 distributions obtained with standard Set drift in the opposite direction (i.e., $G < G_{trg}$). On the other hand, Fig.6b shows that if L1-L8 are homogeneously achieved through Set MLC we have always a drift in the direction of a $G < G_{trg}$. A deeper investigation for the L1-L4 conductance levels has been performed by analyzing the evolution of the distributions during the readout times. As demonstrated in [18], the largest drift is experienced within 1 hour after EA and then progresses for the consecutive readout times. The largest drift usually occurs between AS and EA points for both MLC methods, although the Mixed algorithm lies shows a slight advantage in this. Further, the Mixed methods shows a reduced progression of the drift over 168 hours, as shown in [18].

IV. Time and temperature evolution of drift

As we can see in Fig.3, the largest drift usually occurs between AS and EA points. This ”fast” phenomenon proves to be critical than the drift observed over the short or long term,

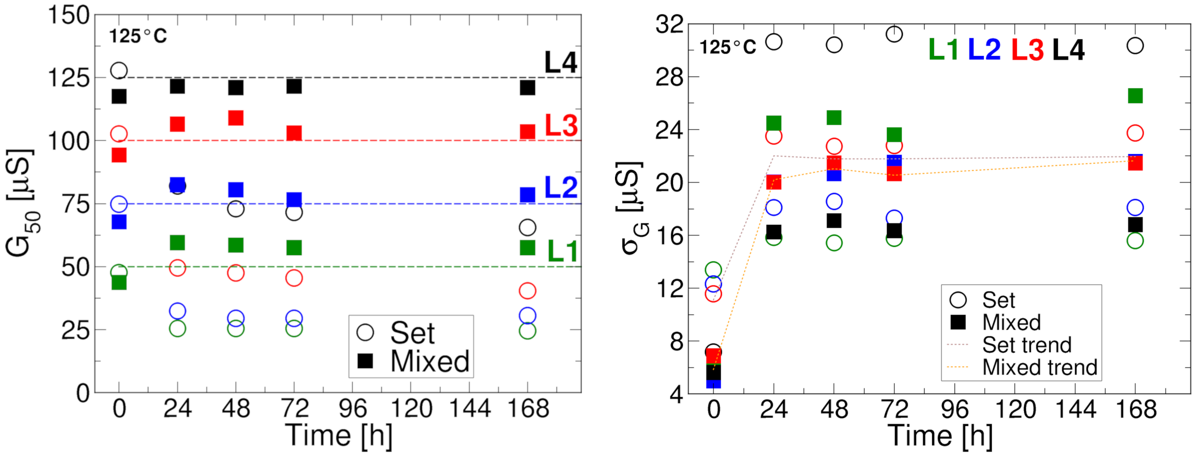

Fig. 10. Evolution of the G_{50} (left) and σ_G (right) parameters of L1-L4 distributions for both MLC approaches at 125°C. Although the G_{50} behaves in a manner consistent with what is observed in lower temperatures and there is a clear more stable behavior shown by the Mixed method, it is not immediate to observe the same evolution for σ_G.

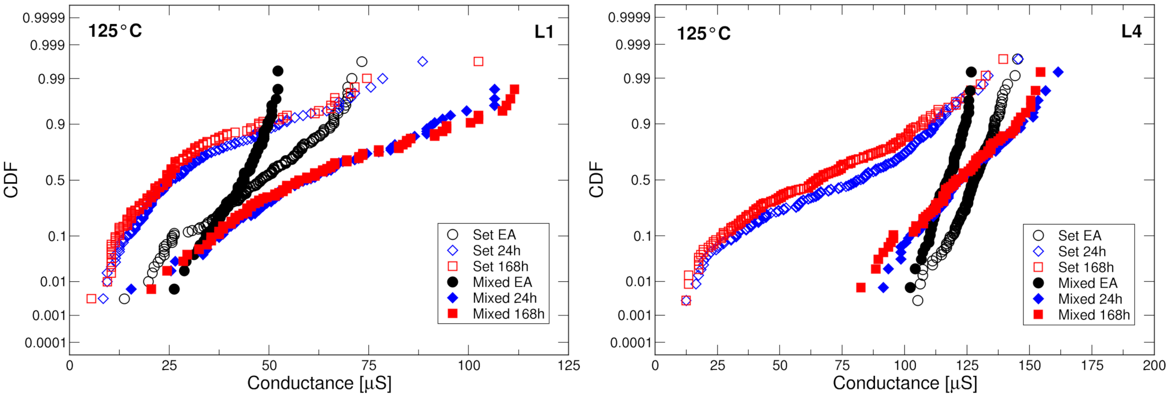

Fig. 11. Readout of the L1 (left) and L4 (right) conductance distribution obtained with both approaches during temperature experiment at 125°C.

and we currently have no way to mitigate it. In Fig.7 we can see how both methods experience the greatest drift between AS and EA, showing σ_G values from 1 μS to 5 μS while for the following hours it gradually increments. We can also notice that, for the Mixed method, the maximum increment of σ_G is more stable in time. Although our goal is not the characterization of the drift in such a short time scale, this observation led us to consider in the following studies the trend of the low conductance state distributions at the EA time, 24 hours and at the end of the 168 hours experiment. This will reflect the behavior of the phenomenon immediately after After Switching (End Algorithm point), in the short term (from 1 to 24 hours), and in the long term by assessing the retention capabilities (up to 168h).

In [18], we explored the evolution of the median conductance G_{50} and of the σ_G for L1-L4 levels only at room temperature (25°C). In both methods, the G_{50} stayed almost constant throughout the entire experiment, evidencing that the drift for L1-L4 is not a rigid shift, but rather a departure of some tail cells in the distribution. A different result stood out from the σ_G analysis. In general, the distributions obtained with Mixed algorithm featured a lower σ_G with respect to those of the Set MLC, exposing a maximum variability of 11.2 μS with respect to the 19.8 μS of the latter approach.

Encouraged by that, we deepened the study of both algorithms at different temperatures, namely 55°C, 85°C, and 125°C. Figs.8 and 9 show the G_{50} and σ_G parameters of L1-L4 distributions for both methods at different temperatures. We can easily see how the trend over time of these parameters reflects what we already experienced at room temperature. As we can see, although the σ_G values of the two different methods tend to have the same trend as the temperature increases, at 85°C they remain very distinct from each other. In addition, it can be noted that as the temperature increases, the behavior of the G_{50} remains almost stable for the Mixed method, while it continues to worsen for the Set MLC method. This is due to a better control of the G_{trg} distribution. This behavior can also be found for measurements at 125°C. Fig.10 shows that G_{50} remains almost stable for the Mixed method while, for the Set MLC method, we experienced decay of its value up to two conductance levels below the target one. As for the σ_G, the same distinction found in previous experiments is no longer observable between the two different methods. We also investigated the evolution of the distribution at 125°C at each time check to understand this peculiar behavior. It was possible to observe how the two proposed methods behave radically different due to their programming history. Both programming methods show a tendency to return to the last state they were before their programming. The Fig.11 include the two extreme cases, L1 and L4, for both methods. It is easy to notice how the distributions obtained through Set MLC tend to return to the L0 level (25 μS), that is the state obtained with a Reset procedure before reprogramming the cells to their targets with this method. Conversely, the distributions obtained with the Mixed approach tend to return to the L7 level (200 μS), although it remains the one with more stable G_{50} value, lower σ_G, and a more accurate distribution. In the worst case (L1 for the Mixed method and L4 for the Set), the evolution show a similar behavior in terms of σ_G, but with a value of G_{50} that continues to be favorable for the Mixed method (see Fig.10). Although the trend of the G_{50} remains favorable for our approach at any temperature, the gradual approach of the σ_G of the two methods can cast doubts on the validity of our solution when the application environment exceed 85°C. To confirm that the Mixed approach is still well suitable in a relevant scenario, we decided to validate it in a practical simulation environment.

V. Assessing the Impact of Drift Reduction on Neural Network Performance

To better understand the implications of the conductance drift caused by the time relaxation and by the temperature, we simulated an implementation of an ANN. The ANN in our case study is a two layer fully connected neural network (FC-NN) trained to classify images of handwritten digits from the MNIST dataset [23]. Each image of the dataset is reduced in both color-depth and size, resulting in a black and white, 14×14 pixels image. The neural network has 196 neurons in the input layer, 20 neurons in the hidden layer, and 10 neurons in the output layer, each representing a digit between 0 and 9. The total number of synaptic weights is 3943, and a schematic representation is depicted in Fig.12a. Each synaptic weight can be mapped as a conductance value into a 1T1R RRAM device.

Unfortunately, RRAM device can only be programmed with a small number of discrete positive conductance values, while the synaptic weights in the traditional neural networks typically require both positive and negative values, and a numerical precision in the order of 32 or even 64 bits. The first limitation can be overcome first by splitting each weight W into two separate positive values, such as G^+ and G^-, and then by mapping the two values into two separate RRAM cells. Finally, by subtracting the current of the two devices in the analog domain we obtain the desired value $W = G^+ - G^-$, as shown in Fig.12b.

To reduce the numerical precision of the synaptic weights without drastically decreasing the network ability to classify input images, a quantization algorithm must be applied. After training the network with full floating-point precision, we implemented the iterative training algorithm of Incremental Network Quantization [24], that allowed us to optimize the neural network to operate with a reduced number of discrete levels. The objective of the experiment was to study the reliability of the newly proposed Mixed algorithm, and how the increased retention performs in real applications compared to the traditional Set MLC algorithm, therefore we simulated a neural network employing only the lowest four programmable LRS levels (i.e., L1-L4) and the L0 level. By using the differential approach described earlier, a total of 9 discrete conductance levels can be obtained, from -100 μS to +100 μS. We simulated the inference operation by randomly selecting conductance values from the distributions obtained by the characterizations performed at different temperatures.

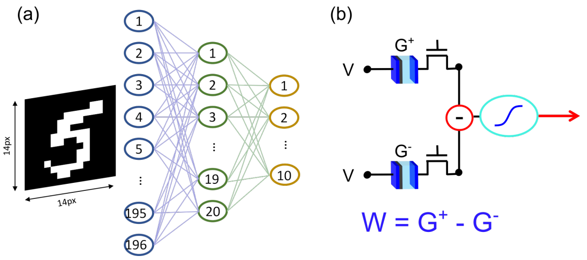

Fig. 12. (a) A 2-layer fully connected NN for recognition of MNIST characters. (b) Differential configuration of 1T1R RRAM cells for synaptic weights implementation.

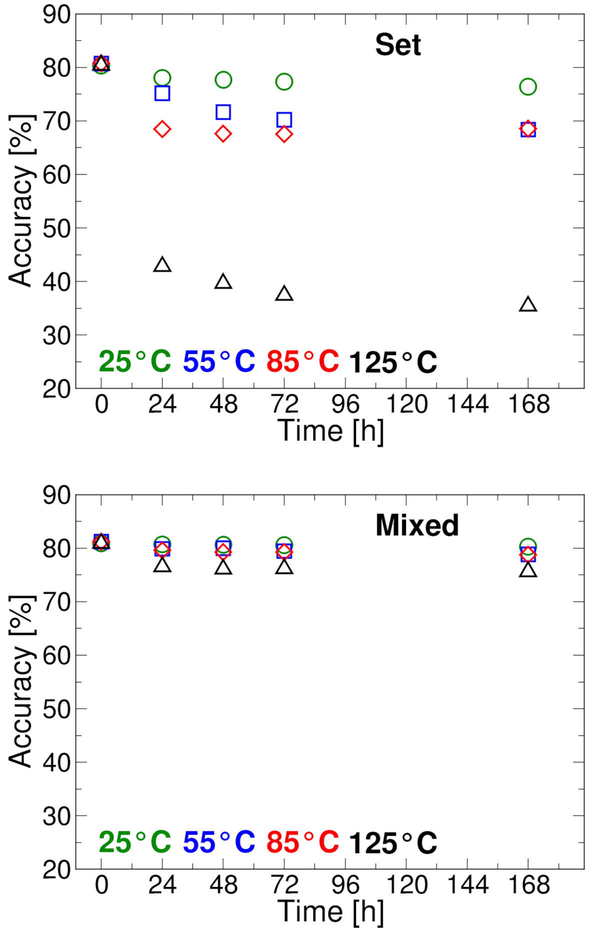

Fig. 13. Evolution of the ANN Accuracy for Set (above) and Mixed (below) MLC algorithms at 25°C, 55°C, 85°C and 125°C.

Fig.13 shows the results of the inference accuracy averaged over 100 simulations, demonstrating that the network implemented with the Mixed algorithm the accuracy it's not significantly impacted by time and is more robust against temperature-induced drift than the network implemented with the traditional Set MLC algorithm.

VI. Conclusions

In this work, we performed a in-depth study of Mixed programming algorithm to reduce the drift affecting the low conductance states in MLC RRAM devices. We have performed a thorough temperature characterization from 25°C to 125°C, for the duration of 168 hours. During all experiments, the proposed method allowed the achievement of a better variability and reliability control, opening the road for a more stable and accurate MLC operation. This new concept has been experimentally validated in 4 kbits RRAM arrays manufactured in IHP 0.25μm technology and compared against state-of-the-art Set MLC. Finally, we validated our approach using the experimental distributions to map the weight of a Neural Network for image recognition allowing us to achieve almost 40% higher accuracy compared to the standard programming methods.

References

[1] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials*, vol. 5, no. 1, pp. 1–36, 2019, DOI. 10.1038/s41524-019-0221-0.

[2] A. Allegra, A. Tonacci, R. Sciaccotta, S. Genovese, C. Musolino, G. Pioggia, and S. Gangemi, "Machine Learning and Deep Learning Applications in Multiple Myeloma Diagnosis, Prognosis, and Treatment Selection," *Cancers*, vol. 14, no. 3, pp. 1–16, 2022, DOI. 10.3390/cancers14030606.

[3] F. Zahoor, T. Z. Azni Zulkifli, and F. A. Khanday, "Resistive Random Access Memory (RRAM): an Overview of Materials, Switching Mechanism, Performance, Multilevel Cell (mlc) Storage, Modeling, and Applications," *Nanoscale Research Letters*, vol. 15, no. 1, pp. 1–26, 2020, DOI. 10.1186/s11671-020-03299-9.

[4] P. Mannocci, A. Baroni, E. Melacarne, C. Zambelli, P. Olivo, E. Perez, C. Wenger, and D. Ielmini, "In-memory principal component analysis by crosspoint array of resistive switching memory: A new hardware approach for energy-efficient data analysis in edge computing." *IEEE Nanotechnology Magazine*, pp. 2–11, 2022, DOI. 10.1109/MNANO.2022.3141515.

[5] G. W. Burr, R. M. Shelby, S. Sidler, C. di 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 Transactions on Electron Devices*, vol. 62, no. 11, pp. 3498–3507, 2015, DOI. 10.1109/TED.2015.2439635.

[6] M. A. Zidan, J. P. Strachan, and W. D. Lu, "The future of electronics based on memristive systems," *Nature Electronics*, vol. 1, no. 1, pp. 22–29, Jan. 2018, DOI. 10.1038/s41928-017-0006-8.

[7] A. Glukhov, V. Milo, A. Baroni, N. Lepri, C. Zambelli, P. Olivo, E. Pérez, C. Wenger, and D. Ielmini, "Statistical model of program/verify algorithms in resistive-switching memories for in-memory neural network accelerators," 2022, in press.

[8] M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, Q. Xia, and J. P. Strachan, "Memristor-Based Analog Computation and Neural Network Classification with a Dot Product Engine," *Advanced Materials*, vol. 30, no. 9, p. 1705914, 2018, DOI. 10.1002/adma.201705914.

[9] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proc. of the IEEE*, vol. 106, no. 2, pp. 260–285, 2018, DOI. 10.1109/JPROC.2018.2790840.

[10] W. Ma, M. A. Zidan, and W. D. Lu, "Neuromorphic computing with memristive devices," *Science China Information Sciences*, vol. 61, no. 6, pp. 1–9, 2018, DOI. 10.1007/s11432-017-9424-y.

[11] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature Electronics*, vol. 1, no. 6, pp. 333–343, 2018, DOI. 10.1038/s41928-018-0092-2.

[12] A. Fantini, L. Goux, R. Degraeve, D. Wouters, N. Raghavan, G. Kar, A. Belmonte, Y.-Y. Chen, B. Govoreanu, and M. Jurczak, "Intrinsic switching variability in HfO2 RRAM," in *IEEE Int. Mem. Workshop (IMW)*, 2013, pp. 30–33, DOI. 10.1109/IMW.2013.6582090.

[13] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Statistical Fluctuations in HfOx Resistive-Switching Memory: Part II—Random Telegraph Noise," *IEEE Trans. on Electron Devices*, vol. 61, no. 8, pp. 2920–2927, 2014, DOI. 10.1109/TED.2014.2330202.

[14] Z. Chai, P. Freitas, W. Zhang, F. Hatem, J. F. Zhang, J. Marsland, B. Govoreanu, L. Goux, and G. S. Kar, "Impact of RTN on Pattern Recognition Accuracy of RRAM-Based Synaptic Neural Network," *IEEE Electron Device Letters*, vol. 39, no. 11, pp. 1652–1655, 2018, DOI. 10.1109/LED.2018.2869072.

[15] Y. Du, L. Jing, H. Fang, H. Chen, Y. Cai, R. Wang, J. Zhang, and Z. Ji, "Exploring the Impact of Random Telegraph Noise-Induced Accuracy Loss on Resistive RAM-Based Deep Neural Network," *IEEE Transactions on Electron Devices*, vol. 67, no. 8, pp. 3335–3340, 2020, DOI. 10.1109/TED.2020.3002736.

[16] 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 *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 6.4.1–6.4.4, DOI. 10.1109/IEDM.2017.8268340.

[17] Y.-H. Lin, C.-H. Wang, M.-H. Lee, D.-Y. Lee, Y.-Y. Lin, F.-M. Lee, H.-L. Lung, K.-C. Wang, T.-Y. Tseng, and C.-Y. Lu, "Performance Impacts of Analog ReRAM Non-ideality on Neuromorphic Computing," *IEEE Trans. on Electron Devices*, vol. 66, no. 3, pp. 1289–1295, 2019, DOI. 10.1109/TED.2019.2894273.

[18] A. Baroni, C. Zambelli, P. Olivo, E. Pérez, C. Wenger, and D. Ielmini, "Tackling the Low Conductance State Drift through Incremental Reset and Verify in RRAM arrays," in *2021 IEEE International Integrated Reliability Workshop (IIRW)*, 2021, pp. 1–5, DOI. 10.1109/IIRW53245.2021.9635613.

[19] E. Pérez, C. Zambelli, M. K. Mahadevaiah, P. Olivo, and C. Wenger, "Toward Reliable Multi-Level Operation in RRAM Arrays: Improving Post-Algorithm Stability and Assessing Endurance/Data Retention," *IEEE J. of the Elec. Devices Society*, vol. 7, pp. 740–747, 2019, DOI. 10.1109/JEDS.2019.2931769.

[20] V. Milo, A. Glukhov, E. Pérez, C. Zambelli, N. Lepri, M. K. Mahadevaiah, E. P.-B. Quesada, P. Olivo, C. Wenger, and D. Ielmini, "Accurate Program/Verify Schemes of Resistive Switching Memory (RRAM) for In-Memory Neural Network Circuits," *IEEE Trans. on Electron Devices*, vol. 68, no. 8, pp. 3832–3837, 2021, DOI. 10.1109/TED.2021.3089995.

[21] A. Grossi, E. Perez, C. Zambelli, P. Olivo, E. Miranda, R. Roelofs, J. Woodruff, P. Raisanen, W. Li, M. Givens, I. Costina, M. A. Schubert, and C. Wenger, "Impact of the precursor chemistry and process conditions on the cell-to-cell variability in 1T-1R based HfO2 RRAM devices," *Scientific Reports*, vol. 8, no. 1, pp. 1–11, 2018, DOI. 10.1038/s41598-018-29548-7.

[22] A. M. Tosson, M. Anis, and L. Wei, "RRAM Refresh Circuit: A Proposed Solution To Resolve The Soft-Error Failures For HfO2/Hf 1T1R RRAM Memory Cell," in *Proceedings of the 26th Edition on Great Lakes Symposium on VLSI*, 2016, p. 227–232, DOI. 10.1145/2902961.2903017.

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, DOI.10.1109/5.726791.

[24] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen, "Incremental network quantization: Towards lossless cnns with low-precision weights," *arXiv:1702.03044 [cs]*, 2017.