Design-Technology Co-Optimization for OxRRAM-based synaptic processing unit

A. Mallik[1], D. Garbin[1], A. Fantini[1], D. Rodopoulos[1], R. Degraeve[1], J. Stuijt[2], A. K. Das[2], S. Schaafsma[2], P. Debacker[1], G. Donadio[1], H. Hody[1], L. Goux[1], G. S. Kar[1], A. Furnemont[1], A. Mocuta[1], P. Raghavan[1]

[1]imec-BE, [2]imec-NL, Contact Email: Arindam.Mallik@imec.be

Abstract – In this paper, we present a design-technology tradeoff analysis to implement a fully connected neural network using non-volatile OxRRAM cells. The requirement of a high number of distinct levels in synaptic weight has been established as a primary bottleneck for using a single NVM as a synaptic unit. We propose a mixed-radix encoding system for a multi-device synaptic unit achieving high classification accuracy (94%) including device variability. To our knowledge, this is the first paper to discuss the tradeoff between single and multi-device synaptic weight in terms of design and technology using silicon data. We have demonstrated that high level of variability can be handled by the neuromorphic algorithm. The results presented in the paper has been obtained from 1Mb array.

Introduction – OxRRAM device as a neuromorphic synaptic unit has established itself as a feasible candidate thanks to its ability to have multi-level states, simple metal-insulator-metal structure. Previous research [1] proposed the implementation of a dot-product engine based on single multilevel OxRRAM devices, however using a large compliance current in the range of 1 mA. In a single device large compliances might be required to counter the intrinsic device variability [2-3]. In practical implementation however, this solution might be inconvenient due to the large area required by the access transistor's driving capability. *This paper presents a trade-off analysis between using single and multiple NVM elements for a synaptic process unit.*

Array test vehicle – A fully functional 1Mb OxRRAM array was fabricated using a 3.3V, 65nm CMOS front-end (**Fig.1**). Memory Element (ME) integration starts on top of the third metal layer (Cu) using a "pillar" integration scheme as shown in **Fig.2**. The active stack, is composed by 3nm-thick TaOx-based oxide sandwiched between thin PVD BE and PVD Hf(5nm)\TiN bottom and top electrodes respectively. The array is logically organized in 1024 rows (WL) and 1024 columns (BL). Cell configuration (**Fig.3a**) is 4T1R. This configuration, devised for characterization, allows maximum flexibility in cell biasing and precise measurement of conductance at the expense of electrical cell area.

OxRRAM properties – Tunable conductance levels are obtained by controlling the gate voltage WAN of the forcing pass gate in **Fig.3a.** In the chosen operating region (between 50µA and 200µA) this results in a nearly linear relationship between voltage, operating current I_{op} and conductance (**Fig.3b). Fig 3c** depicts the conductance distribution of 32 level allocated within the same range. Clearly significant symbol overlapping are present. **Fig 4** demonstrates the reprogramming stability of the highest and lowest conductance level. For a 1kb population no failures are reported with substantially unaltered states distribution through 1 million cycles. Results show that it is possible to obtain multiple distinguished conductance levels, albeit with a tail overlap between consecutive distributions resulting in an error. While this overlap is too large for conventional multibit storage memory applications, it is suitable for neuromorphic applications as shown in the results presented in this study.

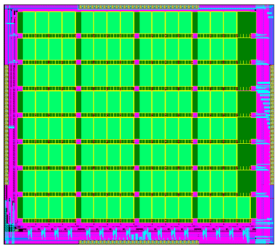
MLP Algorithm Simulation – To demonstrate the effectiveness of the approach, we have used a 2 layer Multi-Layer Perceptron (MLP) algorithm [4] as demonstrated in **Fig.5.** A simulation framework is developed using MATLAB to implement the algorithm with a synaptic model based on the OxRRAM array data. As a representative application benchmark, the MNIST dataset of handwritten digits [5] has been used. In the software version of the MLP, the synaptic weights are stored as floating point values. To implement these weights using an OxRRAM device, we have to convert these floating point weights into finite levels. **Fig.6** summarizes the impact of quantization on the accuracy of the MLP algorithm. Under the current set of assumptions, we don't see much benefit of accuracy going beyond 64 levels.

Design Optimization accounting for variability – With the objective of encoding the desired 64 discrete synaptic weight levels into 1 or more OxRRAM devices, the concept of Synaptic Weight Distortion (SWD) is introduced in **Fig.7a**. This figure of merit evaluates the weighted deviation between the target and the actual synaptic level, due to device variability. Encoding the synaptic weight into 1 OxRRAM with 64 conductance levels (notation: [64]) results in a relatively tight distribution (SWD=0.08). The use of a quaternary base with 3 OxRRAMs [4,4,4] with a constant power budget ($1/3 \cdot I_{op}$ for each device, **Fig.7b**) introduces significant secondary distortion peaks due to errors occurring at the most significant Unit of Information (UoI) (SWD=0.16). In order to reduce the amplitude of the distortion peaks, the total power budget is increased by 3× (SWD=0.03, **Fig.7c**). In **Fig.8** multiple encoding options are compared in terms of SWD. It can be observed that, by increasing the number of OxRRAM devices used for the encoding, it is possible to reduce the SWD. This comes with a power and area penalty, an increase by a factor *N*, where *N* is the number of OxRRAMs per synapse. It is worth noticing that the choice of 3 devices with a mixed radix [2,4,8] improves the SWD compared to the quaternary base [4,4,4]. The obtained SWD approaches the optimal SWD offered by the binary encoding, granting a 50% smaller power and area budget. We ascribe the SWD improvement to the role that variability plays in the least significant UoI versus the most significant UoI. **Fig. 9** summarizes the impact of different encoding schemes at an algorithm level. We observe that the algorithm is unable to withstand the OxRRAM device variability when I_{op} is below 200 µA, in agreement with the SWD metric. The encoding of 64 levels in a single device reduces the classification accuracy by 17%. The use of 2 or more devices helps to recover the loss in accuracy. We can achieve the same accuracy as an ideal device with zero variability by using 6 devices with binary encoding. The impact of variability on the encoded synaptic weight is reduced, in agreement with the SWD. Implementing the synaptic functionality with multiple OxRRAM devices comes at the price of additional area and power as shown in **Fig. 10**. The proposed idea of smart encoding with a mixed radix system of [2,4,8] appears to be a good trade-off amongst all configurations with the same area, assuming that the sensing complexity can be handled at circuit level.

Conclusion – We have developed a framework based on measurement data of a 1Mb OxRRAM array to analyze the trade-offs in multilevel synaptic encoding for MLP pattern recognition. We have introduced a new metric (SWD) to quantify the impact of variability on the encoding error. We demonstrated that the proposed metric is able to correctly predict the trends in recognition accuracy. Taking into account the possibility of using multiple devices, results suggest that, for a given area, the choice of a mixed radix encoding is beneficial in terms of accuracy.

References:

[1] M. Hu *et al.*, DAC (2016)
[2] S. Yu *et al.*, IEEE T-ED 58 (2011) p.2729.
[3] D. Ielmini *et al.* IEEE T-ED 58(2011) p.4309
[4] K. Hornik *et al.,* Neural networks 2.5 (1989): 359-366
[5] Y. LeCun *et al.,* The MNIST Database of Handwritten Digits. http://yann.lecun.com/exdb/mnist/



- **1 Mbit RRAM Array**
- 65nm, 3.3V CMOS process
- 4 Cu Metal interconnect
- Active stack: *BE\TaOx-based\Ta\TiN*
- Physical ME CD: 60nm
- Physical ME Pitch: 200nm
- Cell configuration: 4T1R
- Cell size(*): 4um x 4um

Fig 1: Layout view of the designed 1Mb OxRRAM array and relevant technological parameters.

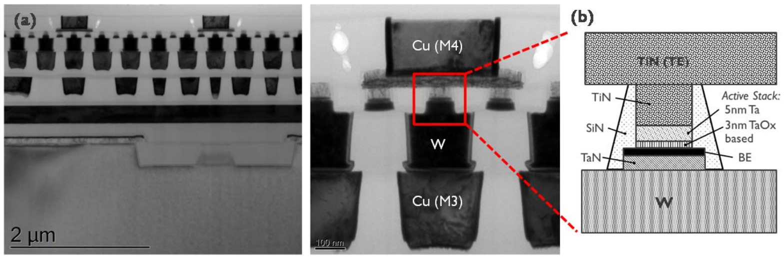


Fig 2: (a) TEM cross section of an array cell showing the 60nm active memory element and the electrically inactive dummy at 200nm pitch. **(b)** Details of the TaOx based memory stack

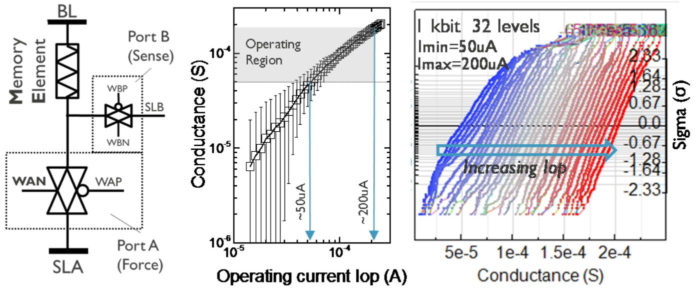


Fig 3: (a) Schematic view of the 4T1R cell configuration **(b)** LRS conductance as function of compliance current **(c)** Distribution of 32 conductance level placed between 50µA and 200µA.

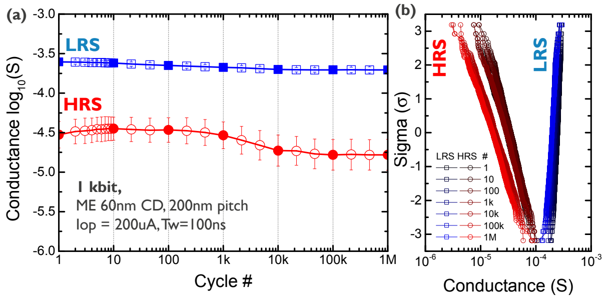
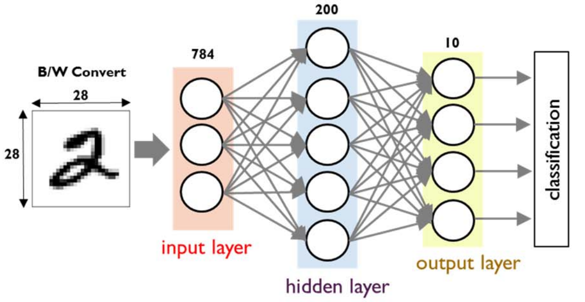


Fig 4: (a) Endurance **(b)** Conductance distributions at different cycling points for the max I_{op}=200µA. Only a slight HRS drift is observed without *any* cell failures. Read margin is constant.



```
Initialization
    $w_1$ = rand(0~1)
    $w_2$ = rand(0~1)
Feed forward
    Hidden layer: $a_1$ = sigmoid($\sum$(input*$w_1$))
    Output layer: $a_2$ = sigmoid($\sum$($a_1$*$w_2$))
Back propagation
    Pre-calculation:
        $s_2 = -2 * a_2 * (1 - a_2) * (output - a_2)$
        $s_1 = \sum(a_1 * (1 - a_1) * w_2 * s_2)$
    Output layer: $w_2 \leftarrow w_2 - \eta * s_2 * a_1$
    Hidden layer: $w_1 \leftarrow w_1 - \eta * s_1 * Input$
Testing: select the largest $a_2$
```

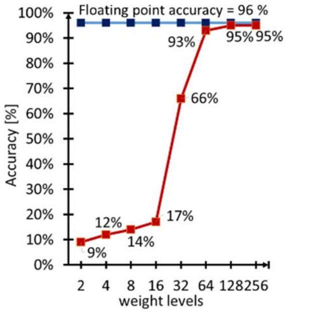


Fig 5: Network topology and Pseudocode of a 2-layer MLP Program (Training with 60K and validation with 10 K MNIST images).

Fig 6: Impact of quantization on the accuracy of the MLP algorithm.

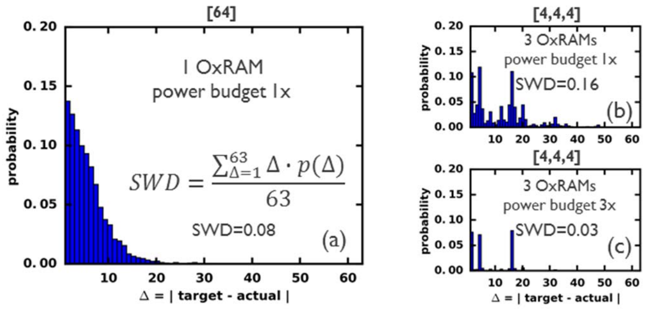


Fig 7: SWD due to device variability when encoding 64 levels in - (a) 1 OxRRAM, (b) 3 OxRRAMs using the same total power budget and (c) 3 OxRRAMs with a 3x power budget.

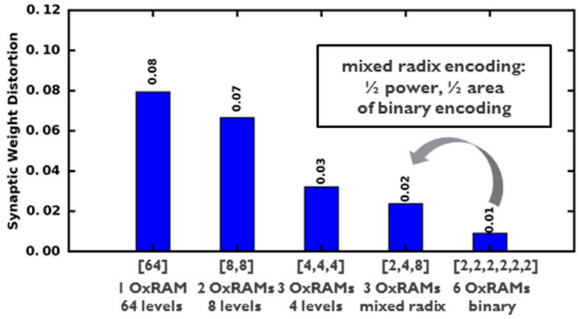


Fig 8: Comparison of SWD introduced by OxRRAM variability corresponding to different encoding options.

Compliance Current (µA)	Maximum Conductance (S)/Device	Classification Accuracy (%) for different Synaptic Weight implementation				
		[64]	[8 8]	[4 4 4]	[2 4 8]	[2 2 2 2 2 2]
70	8.86E-05	-	16	24	24	13
100	1.07E-04	-	33	42	44	45
200	1.95E-04	76	83	89	94	92

Fig. 9: Classification accuracy with MLP benchmark.

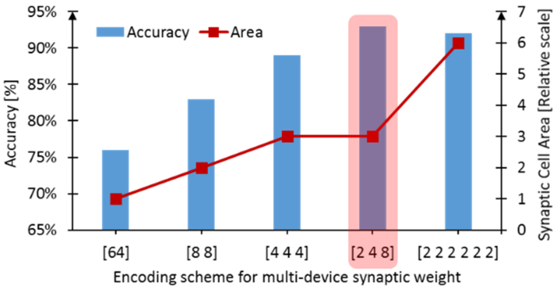


Fig. 10: Accuracy and synaptic cell area trade-off.