Investigation of the potentialities of Vertical Resistive RAM (VRRAM) for neuromorphic applications

G. Piccolboni, G. Molas, M. Portal, R. Coquand, Marc Bocquet, D. Garbin, E. Vianello, C. Carabasse, V. Delaye, C. Pellissier, et al.

Investigation of the potentialities of Vertical Resistive RAM (VRRAM) for neuromorphic applications

G. Piccolboni, G. Molas, J. M. Portal[1], R. Coquand, M. Bocquet[1], D. Garbin, E. Vianello, C. Carabasse, V. Delaye, C. Pellissier, T. Magis, C. Cagli, M. Gely, O. Cueto, D. Deleruyelle[1], G. Ghibaudo[2], B. De Salvo, L. Perniola

CEA, LETI, MINATEC Campus, GRENOBLE, France, giuseppe.piccolboni@cea.fr, (1) IM2NP Marseille, France, (2) IMEP LAHC CNRS Grenoble, France

Introduction

Combining Resistive RAM concept with Vertical NAND technology and design, Vertical RRAM (VRRAM) was recently proposed as a cost-effective and extensible technology for future mass data storage applications [1]. 3D RRAM based neural networks were also proposed to emulate the potentiation and depression of a synapse [2], but more complex circuits were not discussed. In previous works [3-4], various RRAM based neuromorphic circuits were proposed and investigated, using planar devices.

In this paper, we investigate for the 1st time the potentiality of the VRRAM concept for various neuromorphic applications, one synapse being emulated by one VRRAM pillar. First, basic functionality of HfO_2 based VRRAM is presented. 20ns switching time, up to 10^7 cycles and stable 200°C retention were demonstrated. Then specific analyses are made on the resistance and switching voltage variability. We demonstrate that a correlation effect exists between adjacent cycles, meaning the filament keeps a memory of its shape in the previous state, leading to reduced cycle to cycle variability. Based on this preliminary study, using compact model and circuit simulations, VRRAM are proposed for cochlea and convolutional neural network applications, showing good reliability with significant area gain with respect to planar approaches.

Technological details

TEM cross section of our VRRAM is presented in fig.1. A TiN/SiO_2 double layer is deposited on a W plug, the TiN thickness being 10nm. Dry etching is used to pattern a cylinder and reveal a TiN liner, used as a bottom electrode (BE). The etching conditions are particularly critical and were optimized to ensure a vertical profile of the top electrode and reduce the corners effects at the TiN/SiO_2 interfaces. NMOS transistor is used as selector of the memory.

VRRAM characteristics

1. Performances – Typical bipolar IV forming SET and RESET characteristics are showed in fig.2. Low resistance state (LRS) and high resistance state (HRS) resistances versus SET current (I_{SET}) dependence (fig.3) show an increase of both R_{LRS} and R_{HRS} as I_{SET} decreases. Functionality down to 7µA - with ~2.5 decades of mean window margin but higher dispersion - is put in evidence. The higher R_{HRS} at low operating currents results in higher V_{SET} as seen in fig.4. HRS resistance gradually increases with the RESET voltage as illustrated in fig.5. The forming voltage (V_F) is controlled by the Ti top electrode thickness (fig.6), V_F saturating to ~3.3V for Ti>30nm. 200°C retention (fig.7) is insured for both LRS and HRS (I_{SET}=100µA). Switching capability is demonstrated down to 20ns at 2.3V (fig.8). Endurance performances can be adjusted, playing with the SET and RESET conditions, in order to optimize the window margin, the number of cycles or the SET current (fig.9). At least 10^7 cycles could be achieved with 1 decade of window margin (fig.11).

2. Variability – Resistance and switching voltages variability (including both cell to cell and cycle to cycle contributions) can lead to reliability degradation of RRAM based neural networks, in particular in case of internal switching probability. As expected, the resistance standard deviation increases with the mean value for both HRS and LRS (fig.12) in agreement with [4]. Moreover, an interesting feature concerns the fact that the resistance values during cycling do not follow a pure random process but are correlated from one cycle to the other [5] (fig.13). Correlation (quantified by the correlation coefficient) decreases over cycling after ~50 cycles (fig.14). Cycle to cycle correlation means that, despite no resistance drift is measured (fig.15 bottom), the dispersion of the resistance for a given cycle increases as cycling goes on (fig.15 top). This means that the conductive filament keeps for ~50 cycles a "memory" of the shape and resistance it had in the previous cycles. This resistance correlation leads to a correlation of the switching voltage itself, as seen in fig.16. Again, no switching voltage drift is measured but its standard deviation increases during endurance, and saturates after ~30 cycles (fig.17). *In summary, for applications requiring a limited (~10) number of cycles, correlation between adjacent cycles reduces variability (fig.18).*

Neuromorphic applications

1. Synapse behavior implementation – In our approach, a synapse is composed by stacked VRRAM with one common select transistor (Fig.25 left). This offers significant area gain with respect to neural networks in planar configuration with 1T1R elements in parallel (Fig.25 right) [8]. The OxRAM cells operate in binary mode, only two distinct resistive states (LRS and HRS) per device. The analog-like conductance behaviour is achieved thanks to the parallel of n OxRAM cells. We use the intrinsic variability of RRAM in order to implement progressive on line learning (for instance the stochastic learning rule [3]). The switching probability is governed by the RRAM itself (internal switching probability). *Consequently, VRRAM based neural networks (this work) offer area gain due to (1) stacked VRRAM, (2) 1TNR configuration and (3) no random number generator circuit*. 1st, we use VRRAM to emulate the typical progressive behaviour of synapse response. Fig.19 shows the percentage of switched cells (~50 VRRAM measured) as a function of the applied bias and pulse times. SET and RESET conditions can be identified to control the probability to switch the memory with a given value for each pulse, the probability being imposed by the application and neural network structure. An OXRAM compact model [6] is then used to set a model card based on our experimental results. Fig.19 is fitted with good agreement, while cell-to-cell and cycle-to-cycle variability is well reproduced (fig.20). Then the model is used to simulate a synapse composed by stacked VRRAM with 2 to 24 levels (fig.23). The VRRAM are addressed in the same time (one selector for one VRRAM pillar), and the output signal corresponds to the sum of all the VRRAM in parallel (fig.22). Then we calculated the standard deviation of the number or required pulses to SET or RESET half of the VRRAM stacked in one pillar (fig.23). Fig.24 shows that σ decreases as the number of levels increases. Thus it is possible to identify the required number of levels to reach a given σ and emulate typical analogic synaptic behaviour.

2. Cochlea – In the case of real-time auditory pattern extraction (inspired from a 64-channel silicon cochlea emulator, figs.25-26) [3] at least 3 RRAM are required per synapse [4]. In our case, one synapse is thus composed by 3 stacked VRRAM with one select transistor, offering an integration density improvement of >3 with respect to previous work [4]. Pulse times and voltages are adjusted based on fig.19 in order to target switching percentages of 20% for SET and 5% for RESET [7]. As the number of required cycles in cochlea application is high (>10^5) [7], both cell-to-cell and cycle-to-cycle variability are taken into account and affect the internal switching probability. Nevertheless, the circuit behaviour shows good agreement with respect to the perfect case involving an external random number generator (fig.27). Finally, we also assumed a perfectly controlled VRRAM technology where only cycle-to-cycle variability leads to switching voltage dispersion (σ/3 in fig.18). In this configuration, very similar results to the perfect case are obtained (fig.28).

3. Convolutional Neural Network (CNN) – VRRAM can also be a suitable solution for visual pattern extraction applications (inspired from the processing inside visual cortex). It has been demonstrated that for Convolutional Neural Networks (CNN) a higher VRRAM density per synapse is required (~20) [8]. Based on the design used for the recognition of handwritten digits shown in fig.29 and thoroughly described in [4] it appears (from simulations, fig.30) that given a certain resistance distribution (fig.18 left) the recognition rate depends on the number of stacked VRRAM cells. Thus, at least 12 VRRAM levels are required to reach >98% circuit reliability. As CNN applications require limited number of cycles, correlation reduces RRAM cycle to cycle variability and improves circuit reliability with internal switching probability. Being able to stack several memory cells on a pillar (VRRAM) a considerable gain in total area (x10) is obtained compared to a standard design using planar devices. Further gain could even be envisaged for more complex applications requiring more synaptic levels.

Conclusions

In this paper, vertical RRAM are investigated and proposed to increase the density of neuromorphic circuits, one 1TnR pillar emulating one synapse. For cochlea application, good agreement with planar binary OXRAM configuration with random generator is obtained, with an area gain of more than a factor 3. Resistance correlation between adjacent cycles improves the reliability for neuromorphic applications requiring low endurance performances. Thus, for convolutional neural network, 10-15 RRAM levels are sufficient to reach a recognition rate of more than 98%. VRRAM based synapses open the path to high density neuromorphic circuits requiring aggressive synaptic levels.

References

[1] I. G. Baek et al., 2011 IEDM Tech. Dig. pp.737-740.
[2] I. T. Wang et al., 2014 IEDM Tech. Dig., pp.665-668.
[3] M. Suri et al., IEDM 2012 Tech. Dig., pp.235-238.
[4] D. Garbin et al., IEDM 2014 Tech. Dig., pp. 661-664.
[5] A. Calderoni et al., proc. of IMW 2014.
[6] M. Bocquet et al., IEEE Trans. on Elec. Dev. 61, 3, 674-681, 2014
[7] M. Suri et al., IEEE Trans. on Elec. Dev. 60, 7, 2402-2409, 2013.
[8] D. Garbin et al., IEEE Trans. on Elec. Dev. , 2015.

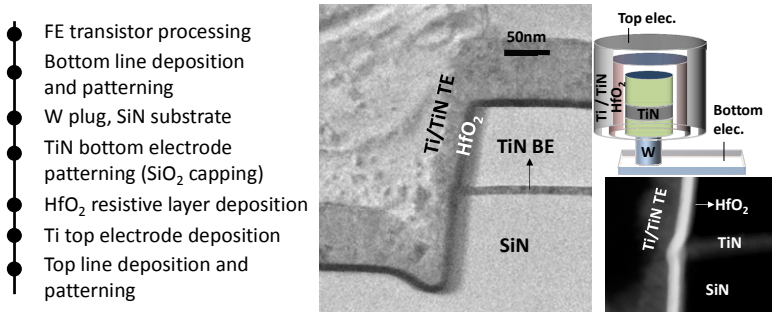


Fig 1. Description of the integration flow, and TEM cross sections (high resolution and dark field) of the TiN/HfO_2/Ti/TiN VRRAM.

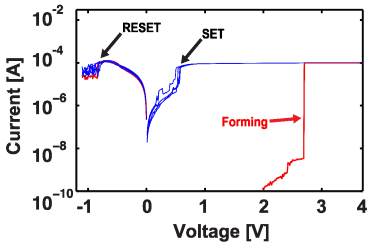


Fig.2 Typical IV forming, SET and RESET characteristics for 100µA of SET Current.

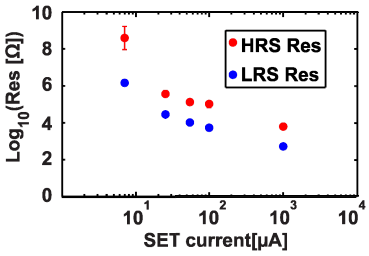


Fig.3 LRS and HRS resistances as a function of the SET current.

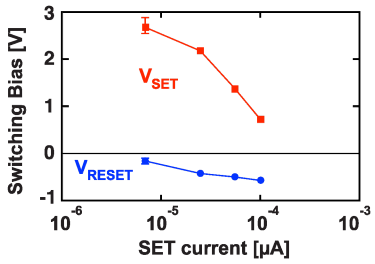


Fig.4 V_{SET} and V_{RESET} as a function of the SET current. Higher V_{SET} is due to higher R_{HRS} (fig.3).

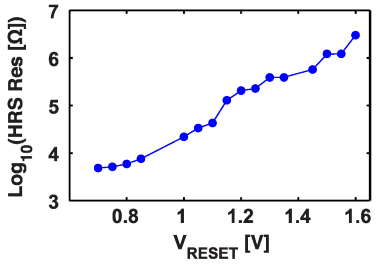


Fig.5 HRS resistance as a function of the bit line RESET voltage.

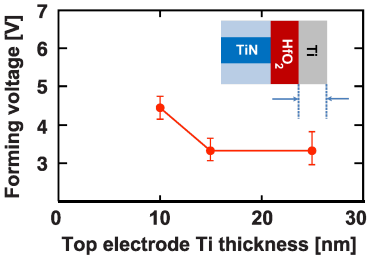


Fig.6 Forming voltage as a function of the top electrode Ti deposited thickness.

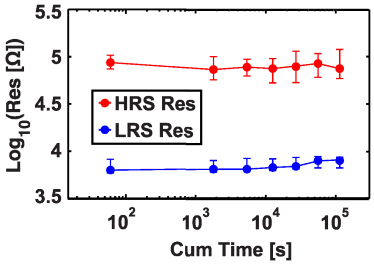


Fig.7 200°C HRS and LRS retention. Median and percentiles (30%-70%) are represented.

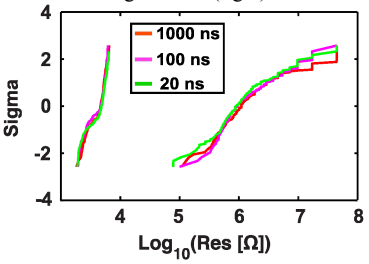


Fig.8 LRS and HRS distributions for various programming pulse widths (20ns - 1µs) for V_{pulse}=2.3V.

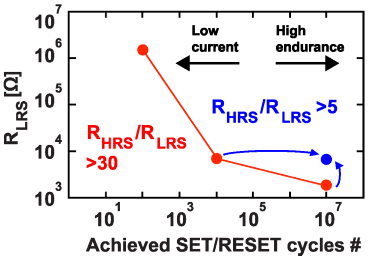


Fig.9 Relation between LRS resistance (related to operating current) and achieved number of cycles during endurance.

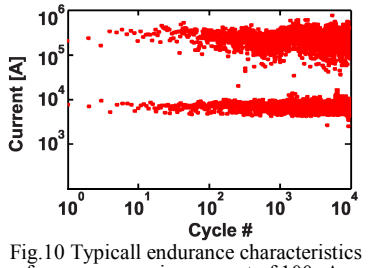


Fig.10 Typicall endurance characteristics for a programming current of 100µA.

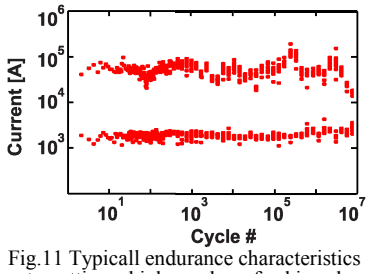


Fig.11 Typicall endurance characteristics targetting a high number of achieved cycle.

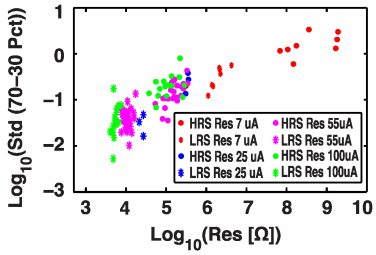


Fig.12 LRS and HRS resistance dispersion as a function of the resistance median.

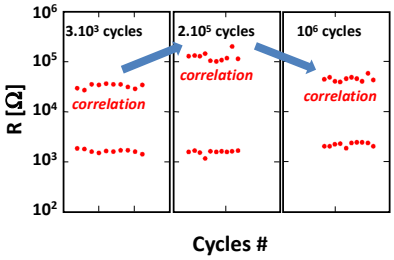


Fig.13 Zoom on cell endurance (fig.11) showing resistance correlation between adjacent cycles.

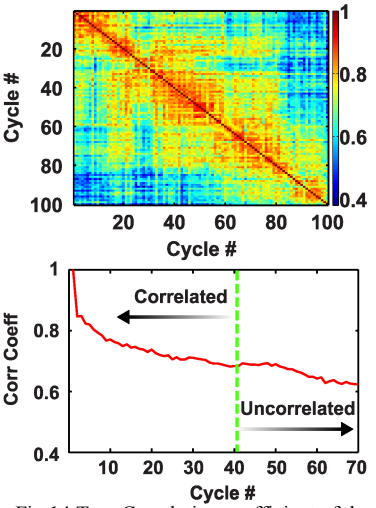


Fig.14 Top: Correlation coefficient of the cell resistance for 50 cells over 100 cycles. Bottom: 1D cut of the curve.

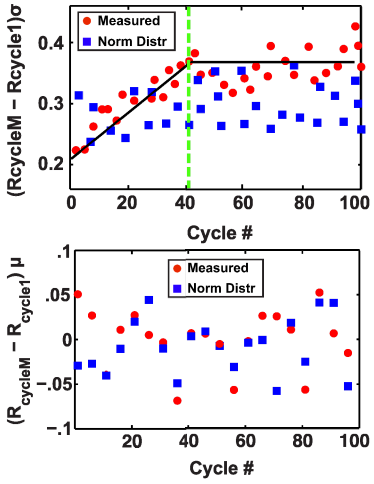


Fig.15 Top: Dispersion of the cell resistance shift during cycling. Bottom: Cell resistance drift during cycling.

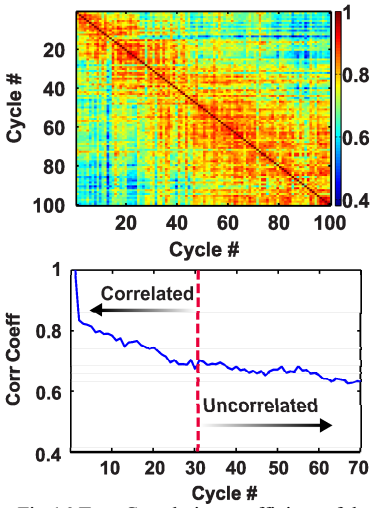


Fig.16 Top: Correlation coefficient of the cell switching voltage for 50 cells over 100 cycles. Bottom: 1D cut of the curve.

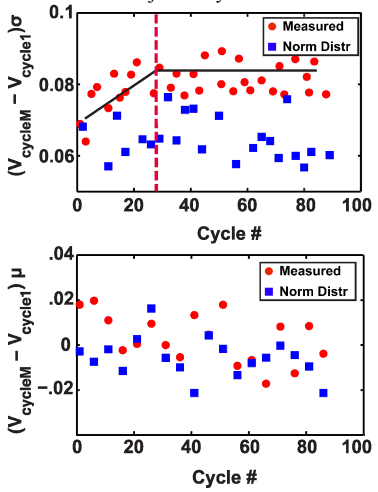
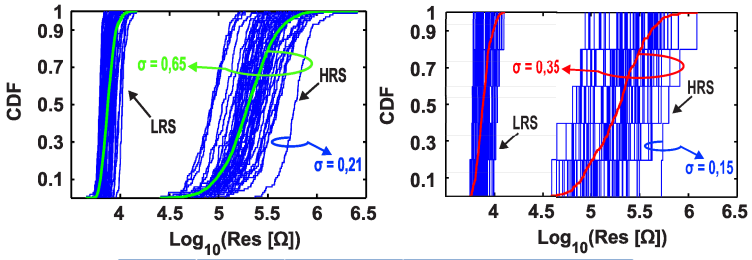


Fig.17 Top: Dispersion of the switching voltage shift during cycling. Bottom: Switching voltage drift during cycling.



Cycles	σ(R_HRS) cycle to cycle	σ(R_HRS) cycle to cycle & cell to cell	Variability
5	0,15 x2,5 → 0,35		A correlation exists between adjacent cycles: cycle to cycle variability is limited
>50	0,21 x3 → 0,65		Cycle to cycle correlation is lost: cell to cell and cycle to cycle variability come into play

Fig.18 SET voltage distribution for 50 cells for 5 (right) and 100 (left) cycles. For a low number of cycles, correlation reduces the dispersion.

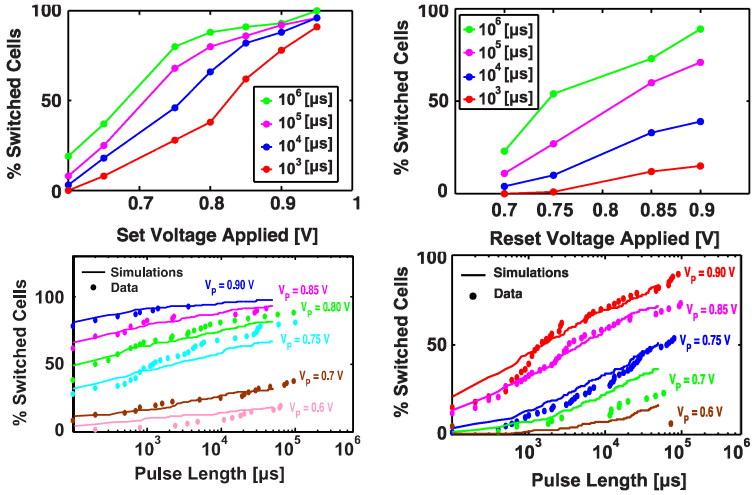


Fig.19 Percentage of SET (left) and RESET (right) cells as function of the pulse time (bottom) or of the pulse voltage(top).

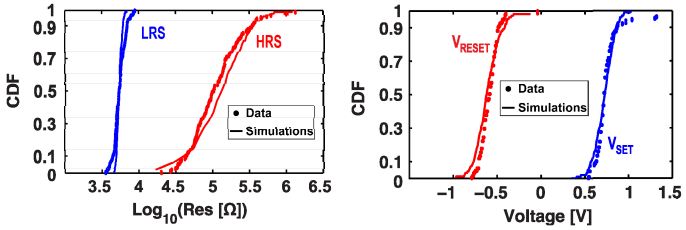


Fig.20 Resistance (left) and voltage (right) distributions (cell to cell and cycle to cycle) by RRAM compact model calibrated on our experimental data.

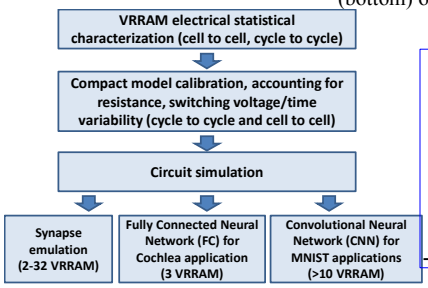


Fig.21 Simulation framework for the 3 neuromorphic applications targetted in this work.

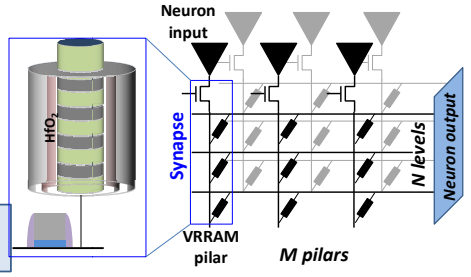


Fig.22 Simulated neuromorphic network. A synapse is composed by N VRRAM cells in a pilar addressed in parallel. The output neuron collects the contributions of all synapses.

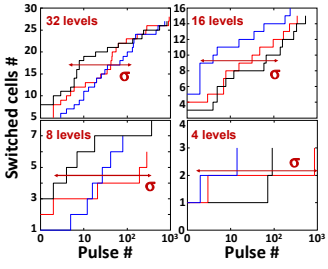


Fig.23 Simulated number of switched cells (for various number of levels) based on the experimental dispersion.

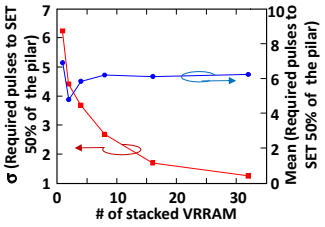


Fig.24 Mean value and standard deviation of the number of cycles to SET 50% of the cells in a pillar as function of the number of stacked VRRAM.

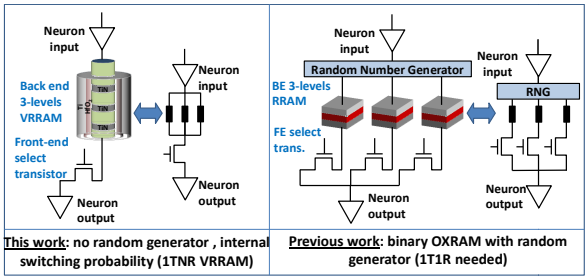


Fig.25 Synapse design comparison between (left) our approach (one 1TnR VRRAM pilar for one synapse) and (right) previous work (planar RRAM [4], select transistor required for every memory cell, with random number generator circuit.

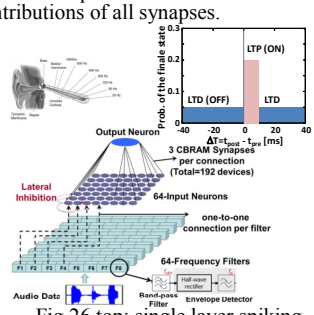


Fig.26 top: single layer spiking neural networksimulated for auditory processing. Bottom: probabilistic STDP learning rule (audio application) [3].

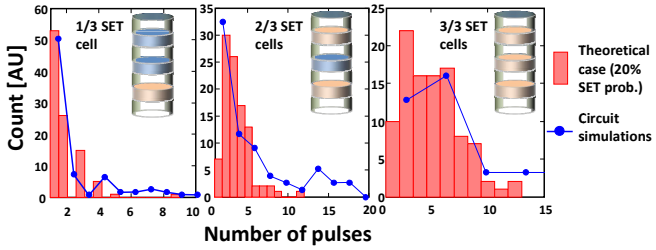


Fig.27 Cochlea application: circuit simulated with synapses composed of 3 stacked VVRAM. SET conditions are fixed to target 20% of occurrence probability (fig. 19). Distributions of required # of pulses to SET respectively 1, 2 or 3 cells per pilar are represented and compared with a theoretical case.

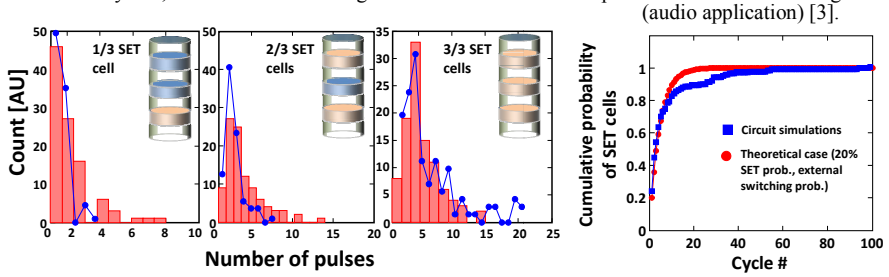


Fig.28 Left: similar simulations from fig.27, but assuming for the VRRAM switching time distributions 1/3 of the measured standard deviation (⇔ perfectly controled VRRAM technology, only cycle to cycle variability taken into account: see fig.18). Right: cumulative distribution of the total number of switched VRRAM showing the good circuit reliabiblity.

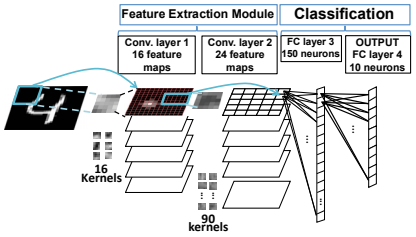


Fig.29 CNN architecture for handwritten digits recognition.

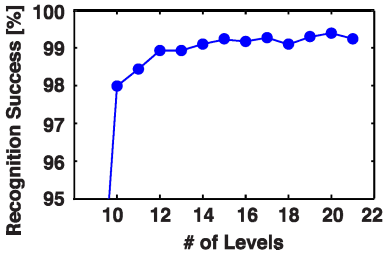


Fig.30 Recognition rate as a function of the number of VRRAM levels (R distributions from fig.18 left) for network of fig.29.

Application	VRRAM #levels (sigma)	Cycles	Area gain	I_{prog} [A]	Impact of RRAM variability
Cochlea	3	10^5	>x3	~10^{-4}	DtD and CtC
CNN	10	10	>x10	~10^{-4}	CtC (correlation between cycles)

Fig.31 Summary of the potentialities of VRRAM for neuromoprhic applications. Area gain is evaluated with respect to planar RRAM with external switching probability.