3D RRAMs with Gate-All-Around Stacked Nanosheet Transistors for In-Memory-Computing

S. Barraud[1], M. Ezzadeen[1,2,3], D. Bosch[1], T. Dubreuil[1], N. Castellani[1], V. Meli[1], J.M. Hartmann[1], M. Mouhdach[1], B. Previtali[1], B. Giraud[2], J. P. Noël[2], G. Molas[1], J.M. Portal[3], E. Nowak[1], F. Andrieu[1]

[1] CEA, Leti, Univ. Grenoble Alpes, 38000 Grenoble, France. E-mail: sylvain.barraud@cea.fr

[2] CEA, List, Univ. Grenoble Alpes, 38000 Grenoble, France, [3] Univ. Aix Marseille, CNRS, IM2NP, Marseille, France.

2020 IEEE International Electron Devices Meeting (IEDM) | 978-1-7281-8888-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IEDM13553.2020.9371982

***Abstract*—** This paper explores a novel 3D one transistor / one RRAM (1T1R) memory cube. The proposed architecture integrates HfO_2-based OxRAM with select junctionless (JL) transistors based on low-voltage Gate-All-Around (GAA) stacked NanoSheet (NS) technology. A bitcell size of $23.9\times F^2/N$ is achieved ('N' being the number of stacked-NS) as well as a very high write and read parallelism. Extensive characterization of JL transistors and OxRAMs is performed to show their ability to be co-integrated inside a same 1T1R memory cell. Electrical characterization of 4kbits OxRAM arrays shows a large memory window (HRS/LRS=20) up to 10^4 cycles with a current compliance of 150µA, compatible with the performances of our JL transistors. Then, we experimentally demonstrate scouting logic operations capability with 2 operands, which should be extended to 4 operands thanks to an original two cells/bit "double coding" scheme assessed by SPICE simulations. Finally, we evidenced that this computing scheme is 2 times more energy efficient than a write-verify approach.

I. INTRODUCTION

Today, storage-class memories like high-density 3D crossbar Resistive-Random-Access-Memories (RRAM) are promising for applications requiring a large amount of on-chip memory. RRAM is a leading candidate due to its high density, good scalability, low operating voltage, and easy integration with CMOS devices [1-7]. Another attractive aspect of RRAM is their ability to perform primitive Boolean logic operations for in-memory [8] and neuromorphic computing [9]. However, if the 1T1R design is the most reliable architecture for in-memory-computing (IMC), the cell size remains limited by the conventional access transistor. In this work, our 1T1R architecture benefits from the high density of vertically stacked NS transistors, developed for advanced CMOS, which feature excellent scalability and 3D integration scalability. This novel 1T1R 3D RRAM architecture is competitive from the crossbar density point of view, while getting rid of any sneak path current, enabling large-scale IMC. Some of the RRAM-based architectures for IMC proposed in the literature imply programming operations [10-11]. Here, we promote the "scouting logic" (SCL) approach, which is only based on reading operation [12-13], ensuring high memory endurance. In the following, we first present the design and the process integration of our 3D RRAM architecture. We discuss the performance of JL transistors and 4kbits OxRAM arrays. Next, based on these experiments, SPICE simulations assess the ability of our 3D RRAM cube to perform SCL operations with up to 4 operands. Finally, we discuss the high parallelization and flexibility for operands selection for write and SCL operations.

II. DESIGN FOUNDATIONS

A. Structure topology and bitcell operating conditions

Our 3D 1T1R architecture is derived from our GAA CMOS structure and process flow [14]; the main difference is that each horizontal GAA channel features an independent source connected to a BitLine (BL) and a drain directly connected to a pillar of RRAM memory cells (**Fig. 2**). This imposes a horizontal Wordline (WL) and vertical Bitline/Sourceline (BL/SL) (**Fig. 3**). A bitcell located at $X_0Y_0Z_0$ is thus selected by the activation of the Wordline $(WL)_{Z0}$ in a XY plane, the Sourceline $(SL)_{X0}$ in a YZ plane and the Bitline (BL_N) ('*N*' being the number of stacked channels). Unselected bitcells are under a WL biased at V_{WL}=GND and/or inhibited at $V_{SL}=V_{BL}$. A parallel programming is possible in the selected Z_0 plane without any parasitic SET/RESET. Similarly, parallel read operations are also allowed in different planes (Z_0, Z_0+1, etc.) on different Bitlines for each plane.

B. Emulated Process-Flow

Fig. 4 shows the process flow used to fabricate the 3D RRAM cube. It is very similar to the GAA CMOS one [14]. We start with an epitaxial growth of (Si:P/SiGe) multilayers and patterning to create the active grid. In-situ phosphorus doped Si layers are used to benefit from doping in the source/drain (S/D) region. Then, a staircase structure is formed at an edge of the memory array. This will enable each BL to be connected independently. Next, the SiO_2/Poly-Si dummy gate is patterned and a SiN spacer fabricated before the anisotropic etching of the (SiGe/Si) fins in the S/D regions. A selective etch of SiGe is then performed prior to the formation of SiN inner spacers. Then, a new module is integrated to reduce the BL resistance, which consists in contacting the Si:P S/D with a lateral metallization, forming a metal spacer. The first option that we had considered was to keep the BL material in n-doped Si (compatible with the JL transistor). Obviously, this configuration would result in too high BL resistivities (**Fig. 5**). The use of metals reduces the BL resistance by a factor of 10^3, enabling longer BL lines. Next, cutting gate poly-Si in a manner that keeps the spacers intact, ensures the integrity of the metal lines under the spacer. Finally, we used a replacement metal gate process to form the WL (gate of selector) and the memory stack is deposited in the SL region with a self-aligned-contact like process. Via0 contacts the Metal1 with the WL, the SL and the different BL stairs. Such a process flow emulation enabled us to define tentative design rules derived from a 28nm CMOS design kit.

C. Bitcell layout

Using this dedicated design kit, we designed a typical bitcell size of ($23.9\times F^2$)/N, 'N' being the number of stacked layers, and 'F' the minimum feature size (45nm in our case) (**Fig. 6**). It should be pointed out that this size is limited by the poly-cut (W_{CT}) width, the space to Via0 distance ($S_{CT\text{-}V0}$), the metal/via pitch (P_{M1}), and the transistor width in the Z direction as well as the gate to Via0 distance and the poly pitch in the X direction. This means that our proposed 3D RRAM technology is competitive with the crossbar memory density (of $4\times F^2$ area) as soon as the number of stacked layers is higher than 6.

III. Selector transistor and RRAM experiments

In the following, we study the performance of GAA JL transistors. First, preliminary TCAD simulations enabled us to optimize the channel doping (N_D) and the channel width to find a good compromise between high performance and good electrostatic control. Typical $I_{DS}(V_{GS})$ curves for W=80nm and different N_D are shown in **Fig. 7**. For a 10nm channel thickness and a NS width of 80nm (yielding a higher I_{ON} current for GAA transistor thanks to a higher effective width), TCAD simulations show that N_d close to 10^{19} cm^{-3} is a good tradeoff, yielding a DIBL=28mV/V (**Fig. 8-9**). In order to confirm this statement, we fabricated nMOS JL transistors with one level of NS (EOT=1nm, T_{Si}=10nm and $N_d \bullet 8\times10^{18} cm^{-3}$) (**Fig. 10**). They exhibit an I_{ON}=120µA at V_{DS}=1.3V, V_{GS}=1.5V, L_G=60nm and W=50nm, which should be translated into a compliance current I_{CC} close to 150µA at W=80nm (maximum width of GAA transistors already demonstrated [14]). We also checked that the local variability is not strongly degraded with JL transistors (**Figs. 11 and 12**) and that they can be cycled up to V_{DD}=2V more than 10^7 times with negligible electrical characteristic modifications. Such a performance is compliant with RRAM programming. Similarly, we characterized 4kbits 1T1R RRAM arrays following a similar integration (TiN bottom electrode/5nm HfO_2/5nm Ti/TiN top electrode integrated in the BEOL) as in refs. [15-17]. Cumulative distributions of LRS and HRS states for various I_{CC} and V_{RESET} are shown in **Fig. 13**. They show that the high (HRS) and low (LRS) resistive states can be well differentiated for high values of V_{RESET} (3.25V) and I_{CC} (150µA). A large memory window (HRS/LRS=20) and an endurance $>10^4$ cycles are obtained (**Fig. 14 to 16**). From those experimental data (**Fig. 13b**), LRS and HRS distributions are fitted with normal and/or lognormal distributions for different compliance currents (75, 100, and 150µA). These resistances are then used in SPICE simulations with a transistor model emulating the performances of JL transistors.

IV. Evaluation of the CUBE for scounting logic by SPICE simulations

In this last part, we discussed the feasibility of SCL, enabling native Boolean operations, with our 3D RRAM cube (**Fig. 3**). Monte Carlo simulations are performed with 1000 runs, and variability is considered up to +/-3σ. To perform SCL operations on *n* operands, *n* 1T1R layers are selected simultaneously by activating the appropriate WL and setting the corresponding BL_i to the read voltage. The resulting total current I_{OUT} belongs to different distributions depending on the number of activated cells at HRS (*i.e.* logic state '0') and LRS (*i.e.* logic state '1') (**Fig. 17a**). By setting a reference current I_{REF} between the different possible current distributions, and comparing I_{OUT} to these references, OR and AND operations can be performed natively (**Fig. 17b**). In **Fig. 18**, we show an experimental demonstration of I_{OUT} distributions, with our 4kbits OxRAM array, based on two operands. The distributions related to the states '00', '01/10' and '11' have large enough memory window (MW) enabling sensing operation. To evaluate the best bitcell configuration and programming conditions, we simulated three cases, for various I_{CC}: (a) "Single Coding" (SC) where a bitcell is represented by a single 1T1R device. (b) "Double Coding" (DC) where a bitcell is represented by two parallel 1T1R devices sharing the same SL and programmed at the same value, and (c) SC with a maximum of five write-verify cycles (SC-WV) to tighten the OxRAMs resistance distributions. Simulations show that SCL can be successfully achieved with up to four activated layers (for I_{CC}=150µA) with DC and SC-WV, since DC and SC-WV increase the MW, facilitating sensing operations compared to SC (**Fig. 19**). As shown in **Fig. 20**, DC and SC-WV seem to provide the highest number of operands (N_{OP}) with positive MW compared to SC. Moreover, DC is the only one to provide proper read operations at the lowest I_{CC} (**Fig. 21**). Moreover, as shown also in **Fig.21,** the DC coding scheme is 2 times more energy efficient than the SC-WV one, when considering one write operation followed by a successful read or SCL operations. Finally, our cube enables high parallelization and flexibility for operands selection. Unlike planar RRAM arrays, SCL can be performed between words of different planes, as long as they share the same row levels in each plane (**Fig. 22**).

V. Conclusion

In this work, we proposed a novel 1T1R 3D RRAM architecture combining two emerging technologies (GAA stacked-NS transistors and RRAMs). We show that a proper engineering of JL GAA transistors can result in compliance currents (I_{CC}=150µA) which are high enough to address OxRAM arrays whose memory window reaches HRS/LRS=20 up to 10^4 cycles. This new architecture, which offers a high write and read parallelism, can be leveraged for IMC. Actually, SPICE simulations show that under optimal bitcell configuration and programming conditions, SCL operations can be successfully achieved with up to four operands.

Acknowledgment

This work was partly funded by the European Research Council through MYCUBE project (grant N° 688101) and by the French Public Authorities through the NANO 2022 program.

References

[1] I.G. Baek et al., IEDM (2011), [2] W.C. Chien et al., VLSI Technology (2012), [3] H.-Y. Chen et al., IEDM (2012), [4] Q. Luo et al., IEDM (2017), [5] G. Piccolboni et al. IEDM 2015, [6] S. Qin et al., IEEE TED, **66**, p.5139 (2019), [7] A. Bricalli et al. IEEE TED, **65**, p.122 (2018), [8] J. Yu et al., Nanoarch symp. (2019), [9] A. Valentian et al., IEDM (2019) [10] S. Kvatinsky et al., IEEE Trans. on VLSI Systems, **22**, p.2054 (2014), [11] S. Kvatinsky et al., IEEE Trans. on circuits and Systems, **61**, p. 895 (2014), [12] S. Li et al., Proc. Design Automation Conf., pp.1-6 (2016), [13] L. Xie et al., IEEE Comp. Soc. Annual Symp., pp. 176-181 (2017), [14] S. Barraud et al., VLSI technology (2020), [15] Grossi et al., IEDM (2016), [16] A. Grossi et al., IEEE EDL, **39**, p.27 (2018), [17] J. Sandrini et al., IEDM (2019).

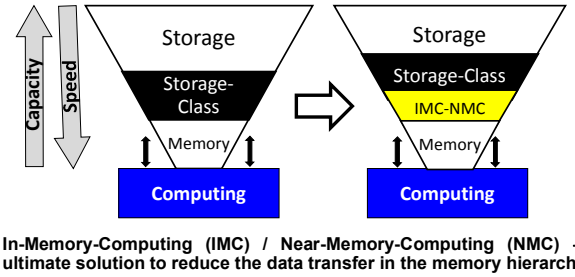


Fig. 1: Evolution of memory hierarchy in computing system including In-Memory-Computing (IMC) in the storage class memory to compensate the gap between the storage media and the main memory.

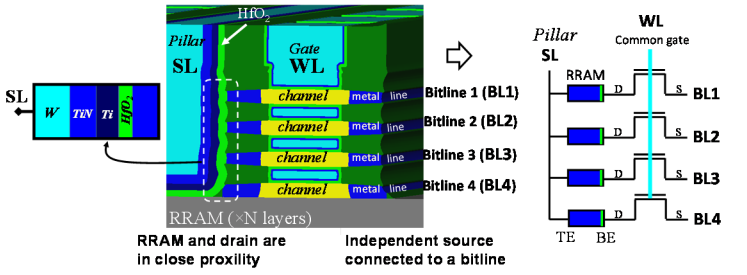


Fig. 2: Cross-sectional schematic of our 1T1R OxRAM memory cell based on GAA nanosheet transistors with an independent **BitLine (BL) source** and a common gate for **WordLine (WL)**. A 3D pillar is used to connect the **SourceLine (SL)**.

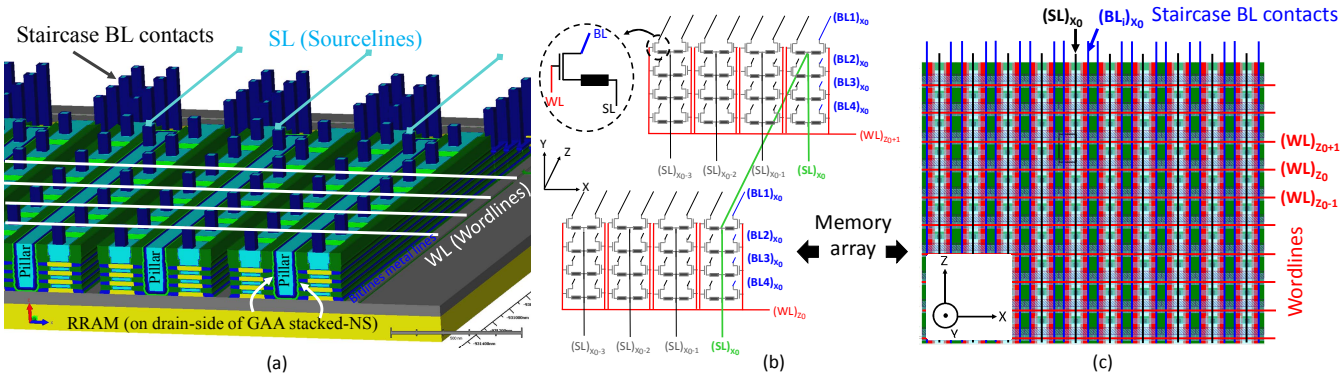


Fig. 3: 3D overview, schematics and layout of our 3D RRAM 1T1R architecture. Staircase BL contacts are defined at an edge of the memory array. The WordLine (WL) are connected along the X direction, the SourceLine (SL) and BitLine (BL) along the Z direction.

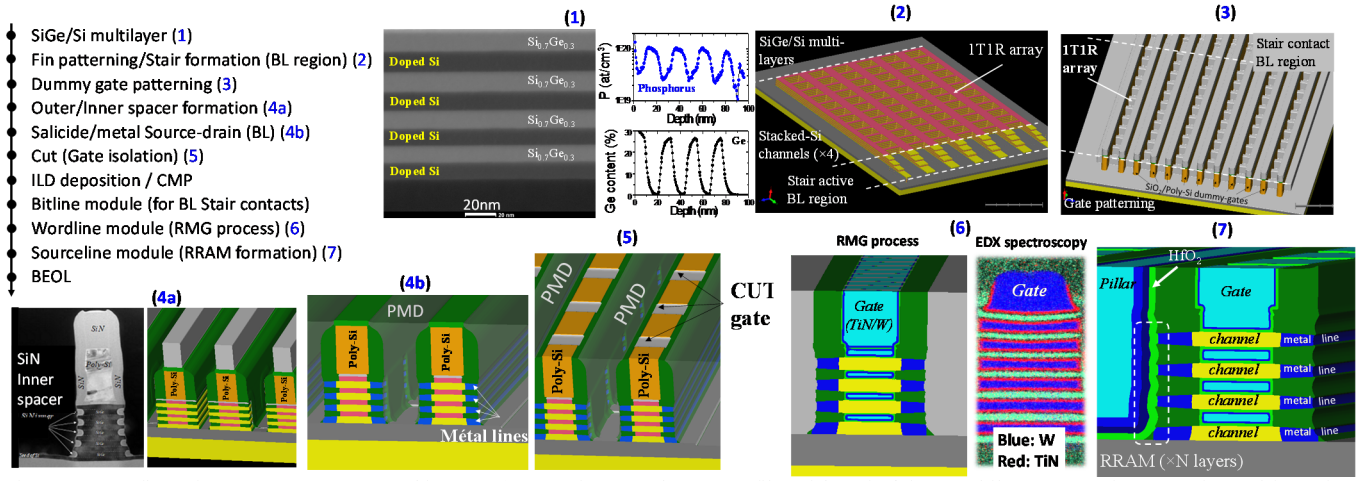


Fig. 4: Process-flow of our 3D RRAM 1T1R architecture. (1) TEM image and SIMS profiles of the (Si:P/SiGe) multilayers, (2) active patterning and formation of stairs at an edge of the memory array, (3) dummy-gate patterning, (4a) TEM image of the SiN inner spacers, (4b) formation of metal lines to reduce the BL access resistance, (5) gate isolation along the Z direction, (6) RMG module for WordLine, (7) deposition of OxRAM stack in the SL pillar and planarization.

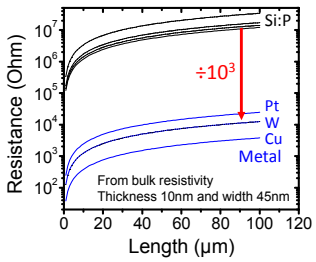


Fig. 5: (a) R_{LINE} *vs* L for Si:P and metal lines. R_{LINE} is reduced by a factor of 10^3 reducing the R_{ACCESS}.

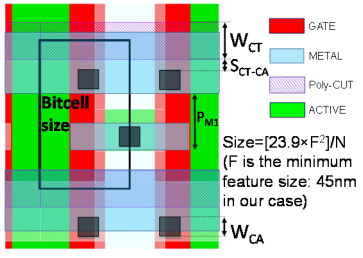


Fig. 6: Evaluation of our cell size (S=[23.9×F^2]/N). 'N' represents the number of stacked layers.

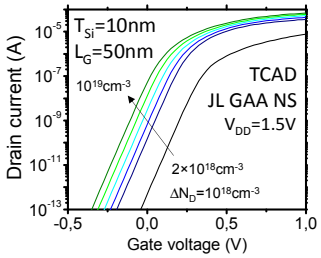


Fig. 7: $I_{DS}(V_{GS})$ electrical characteristics of JL GAA NS FETs for various channel dopings.

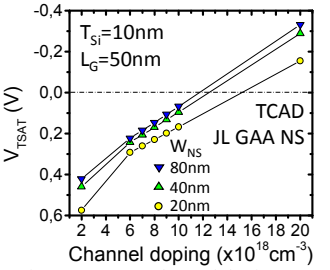


Fig. 8: V_{TSAT} *vs* channel doping (N_D) for different W_{NS}. N_D must be $\leq 10^{19}cm^{-3}$ to keep positive V_{TSAT}.

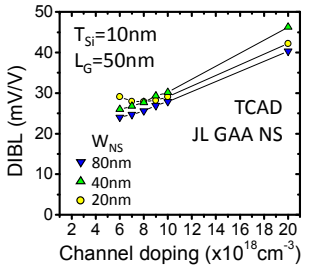


Fig. 9: DIBL *vs* channel doping of JL GAA NS FET for different W_{NS}.

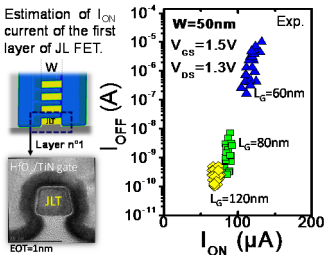


Fig. 10: I_{OFF}-I_{ON} current of the first layer of JL FET. Here, W=50nm.

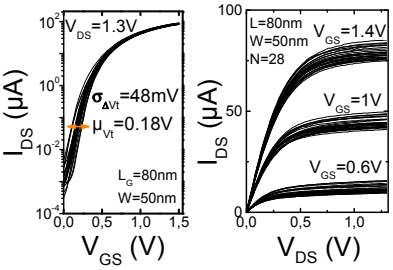


Fig. 11: $I_{DS}(V_{GS})$ and $I_{DS}(V_{DS})$ curves of JL FET for W=50nm and L_G=80nm.

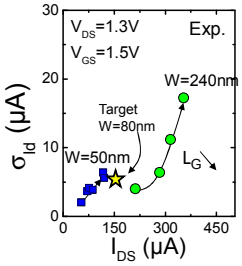
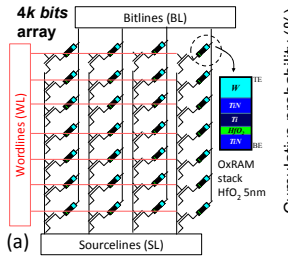


Fig. 12: $\sigma(I_{DS})$ *vs* I_{DS} of JL FET for W=50 and 240nm.



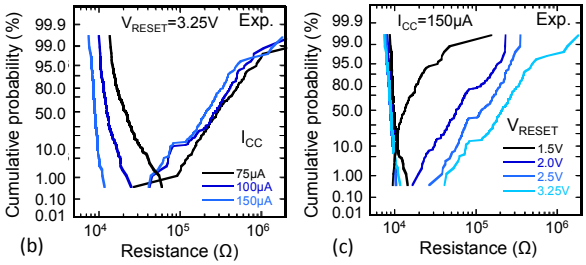


Fig. 13: (a) Schematics of our 4kbits 1T1R memory array integrating TiN/HfO_2 5nm/Ti 5nm/TiN OxRAMs. LRS and HRS distributions of OxRAM memory cells for (b) different I_{CC} and (c) V_{RESET}. High values of I_{CC} and V_{RESET} are preferable to ensure a large memory window (MW) between HRS and LRS states.

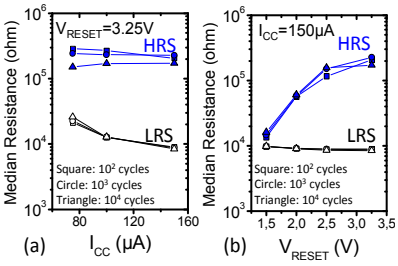


Fig. 14: Median R of LRS and HRS distributions for various I_{CC} (a) and V_{RESET} (b) up to 10^4 cycles.

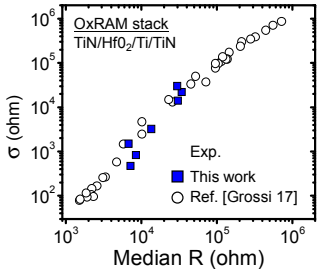


Fig. 15: Standard dev. σ *vs* median R during SET for various I_{CC}.

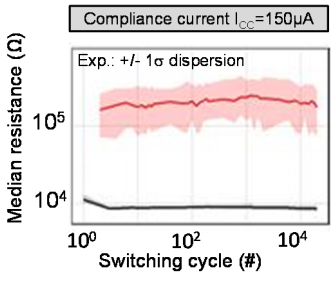
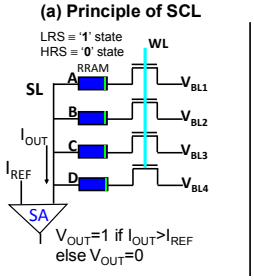


Fig. 16: Median R of HRS and LRS distributions *vs* switching cycle for I_{CC}=150μA. Here, V_{RESET}=3.25V.



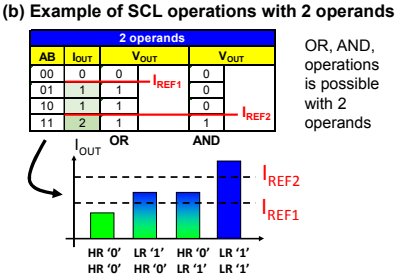


Fig. 17: (a) Scouting logic principle: depending on the RRAM states, I_{OUT} exhibits different values comparable to I_{REF} of various logic operations. (b) Examples of OR and AND operations performed with two operands.

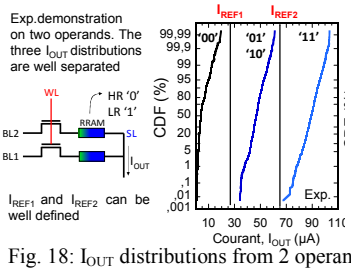


Fig. 18: I_{OUT} distributions from 2 operands. The three states are not overlapped.

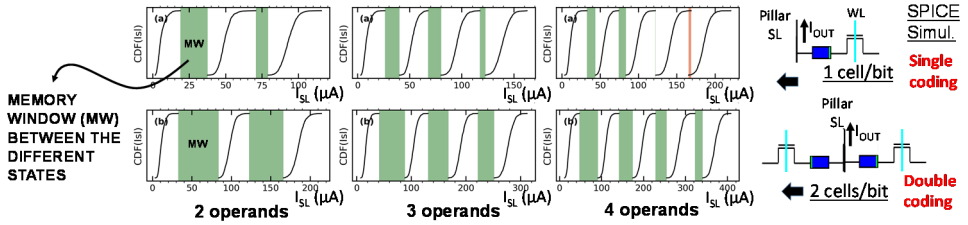


Fig. 19: SPICE simulation results of scouting logic with I_{CC}=150μA on 2, 3 and 4 operands corresponding to a pillar of our 3D RRAM 1T1R architecture with 1cell/bit (single coding) and 2 cells/bit (double coding).

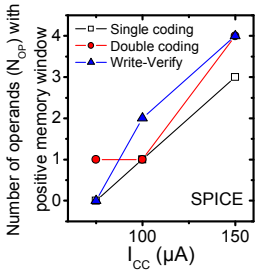


Fig. 20: Max. number of operands (N_{OP}) with positive MW enabling scouting logic (SPICE simulation).

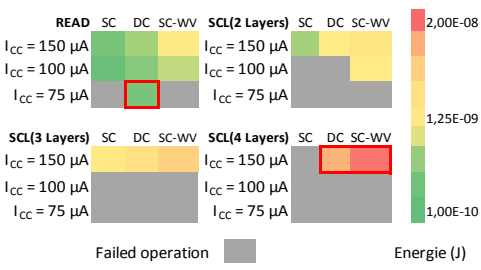


Fig. 21: Heatmap of the mean energy cost per operation for read and SCL operations (from 2 to 4 operands). Different compliance currents (from 75 to 150μA) are considered. (SPICE simulation).

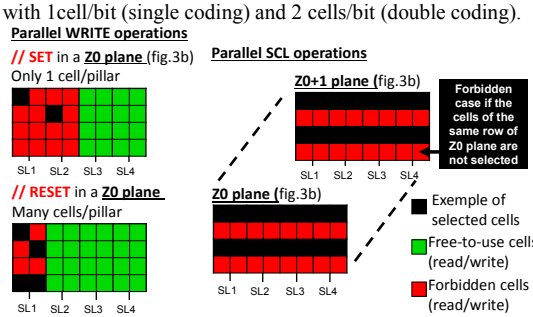


Fig. 22: Illustration of parallel write and SCL operations of our 3D RRAM architecture. Each word can be adressed in a row of the cube. The parallel SCL operation is performed on the same row of the different 'Z_i' planes of the 3D RRAM cube.