Proceedings of the 17th
IEEE International Conference on Nanotechnology
Pittsburgh, USA, July 25-28, 2017

3D ReRAM Arrays and Crossbars: Fabrication, Characterization and Applications

Gina C. Adam
National Institute for R&D in Microtechnologies (IMT)
Bucharest, Romania
gina.adam@imt.ro

Bhaswar Chrakrabarti, Hussein Nili, Brian Hoskins, Miguel A. Lastras-Montaño, Advait Madhavan, Melika Payvand, Amirali Ghofrani, Kwang-Ting Cheng, Luke Theogarajan, Dmitri B. Strukov
Department of Electrical and Computer Engineering
University of California, Santa Barbara, USA

***Abstract*—As the rapid progress of memristor technology continues, multi-layer stacking of these crossbars is needed in order to maximize the use of vertical space and achieve the required density for high throughput applications. This work summarizes our efforts of designing and building three-dimensional monolithically integrated memristive arrays and crossbars, both standalone and onto CMOS chips. We discuss the fabrication and electrical characterization details of stand-alone and CMOS integrated ReRAM arrays and crossbars together with their use in experimental demonstrations of digital and analog applications such as three-dimensional stateful logic, hardware security primitives and dot-product operations.**

Keywords— Three-dimensional (3D), crossbar, metal-oxide, analog, material implication, CMOS integration, security primitives

I. Introduction

The 3D "CMOL" (CMOS+molecular devices) architecture [1] implemented using resistive switches is a promising candidate for energy-efficient hardware implementations of neuromorphic circuits and dense non-volatile memories. Several challenges, like the the thermal budget, yield and uniformity, have to be considered during the fabrication of such hybrid multi-stack systems. The memristors should have tight switching variations to achieve high performance circuits. There have been demonstrations of multi-layer crossbar circuits [2-3], but mostly targeted towards digital memories and showing limited characterization statistics. Sidewall vertical integration is a cost-effective stacking solution [4-5], but it has been shown so far only for small linear arrays, not crossbars, and is not entirely suitable for the CMOL architecture.

Important steps have been made towards the hybrid integration of ReRAM technology with CMOS. There have been reports where memristors were fabricated between CMOS metal layers or onto CMOS chips [6-8], but they include limited discussions on the quality of the interface between the CMOS chip and the layers of resistive switches. Recently, in-memory computation capability was shown in 3D vertical resistive memory devices monolithically integrated with FinFET selectors [6]. However, successful demonstrations of CMOL circuits are still missing.

This work summarizes our efforts of designing and building 3D monolithically integrated memristive crossbars, both standalone and onto CMOS chips, and further using them for prototyping promising applications. The paper is organized as following. Firstly, the fabrication details are presented in section II. Section III is dedicated to the electrical characterization. Lastly, in section IV, the applications implemented experimentally with these devices are presented, ranging from stateful logic, to physically unclonable functions (PUFs) and to dot-product operations.

II. Fabrication

Firstly, we reported the fabrication of a small three-dimensional array of four bipolar ReRAM devices (Fig. 1a) [9]. The middle electrode was shared between bottom and top devices. The bottom devices were larger so as to prevent device failure due to misalignment and had an active area of 500 nm × 500 nm, while the the top devices an active area of 300 nm × 500 nm. All photolithography steps were done with an ASML DUV stepper. Three lift-off steps were used to pattern a) the bottom electrode deposited with e-beam metal deposition, b) the first active layer and the middle electrode deposited via sputtering and c) the second active layer and top electrode deposited via sputtering as well. A fourth photolithography was used to dry etch through the sacrificial SiO_2 in order to electrically contact the samples. The thermal budget of this process was 175°C and thus CMOS compatible.

Three fabrication details were taken into consideration when designing the processing flow for these devices. Firstly, after extensive material engineering as previously reported in [10], optimized sub-stoichiometric TiO_{2-x} developed via reactive sputtering was chosen as the active material, together with a thin Al_2O_3 barrier film to increase the non-linearity of the device. The TiO_{2-x} film had controlled stoichiometry by controlling the oxygen flow rate during the deposition. By depositing sub-stoichiometric films, it was possible to reduce the forming voltages of these devices as needed for the monolithic integration into working arrays and crossbars.

Secondly, it was observed in previous work [11] that the devices with in-situ top interface performed better and did not need annealing. Deposition of the active film and of the top metal layer in the same vacuum preserved the properties of the sub-stoichiometric material, by avoiding its oxidation in air which would increase the forming voltage and defeat the purpose of having sub-stoichiometric engineered material in the first place.

The third aspect to consider was the necessity of having the second layer of devices fabricated on a flat surface to prevent poor step coverage and low device yield. Therefore, a planarization step consisting of chemical mechanical polishing followed by a controlled dry etch-back of a PECVD SiO_2 layer deposited at 175°C was introduced after the fabrication of the first layer of devices.

It is important to mention that by using lift-off to pattern layers that were sputtered, the devices were prone to having rabbit-ear like formations around those lines. Despite using the smallest pressure of 0.9mTorr in the sputter chamber to reduce the sidewall redeposition of the metal and a thicker undercut layer, occasional rabbit ears formations posed challenges to the fabrication of these devices, creating shorts and affecting the yield. Moreover, the lift-off procedure presents a tradeoff between the undercut layer thickness as permitted by the smallest feature size desired vs. the deposited metal thickness. This method limits the metal lines to small thicknesses, which lead to high line resistances unsuitable for large crossbars. Overall, while successful on the small scale, this method was not deemed suitable for the fabrication of larger three-dimensional crossbars and scaling up of multi-layer stacking.

An alternative fabrication flow based on ion milling can be used to avoid the rabbit-ear formations and produce clean line edges. An additional advantage is that the metal could be much thicker than in the lift-off case discussed above. The ion milling process is a physical dry etch process using inert ions, like Argon (Ar), to bombard and dislocate the material in order to remove it from the areas not protected by the mask. By comparison with the lift-off method where the photoresist is patterned first and the material deposited secondly, in this etching process the material is deposited in blanket first via sputtering, then the mask is patterned. Since the Ar ion milling is purely physical etching, it has low selectivity and it is prone to material redeposition, thus a thin hard mask – like Al_2O_3 – is the most suitable.

Based on this modified process flow, we reported three-dimensional monolithic integration of two 10x10 $Pt/Al_2O_3/TiO_{2-x}/TiN/Pt$ -based passive crossbars with shared middle electrode (Fig. 1b) [12]. Chemical mechanical polishing (CMP) of sacrificial SiO_2 was used to planarize the surface after the fabrication of the bottom layer of devices, similar to the small arrays described above. Through AFM-controlled dry etch-back of the remaining SiO_2, the middle electrode was carefully exposed in order to make electrical contact with the top layer of devices. Wafer-scale SiO_2 thickness non-uniformities of several tens of nanometers introduced by chemical mechanical polishing step led to some of the crossbars on the wafer to not have their middle electrode exposed, thus having a top crossbar not fully contacted. Overall the yield across the wafer was 74 out of 100 crossbar dies with both bottom and top crossbars functional, and 26 dies with only a bottom crossbar. By using chemical mechanical polishing of the metal instead of the sacrificial SiO_2, the yield can be increased and this

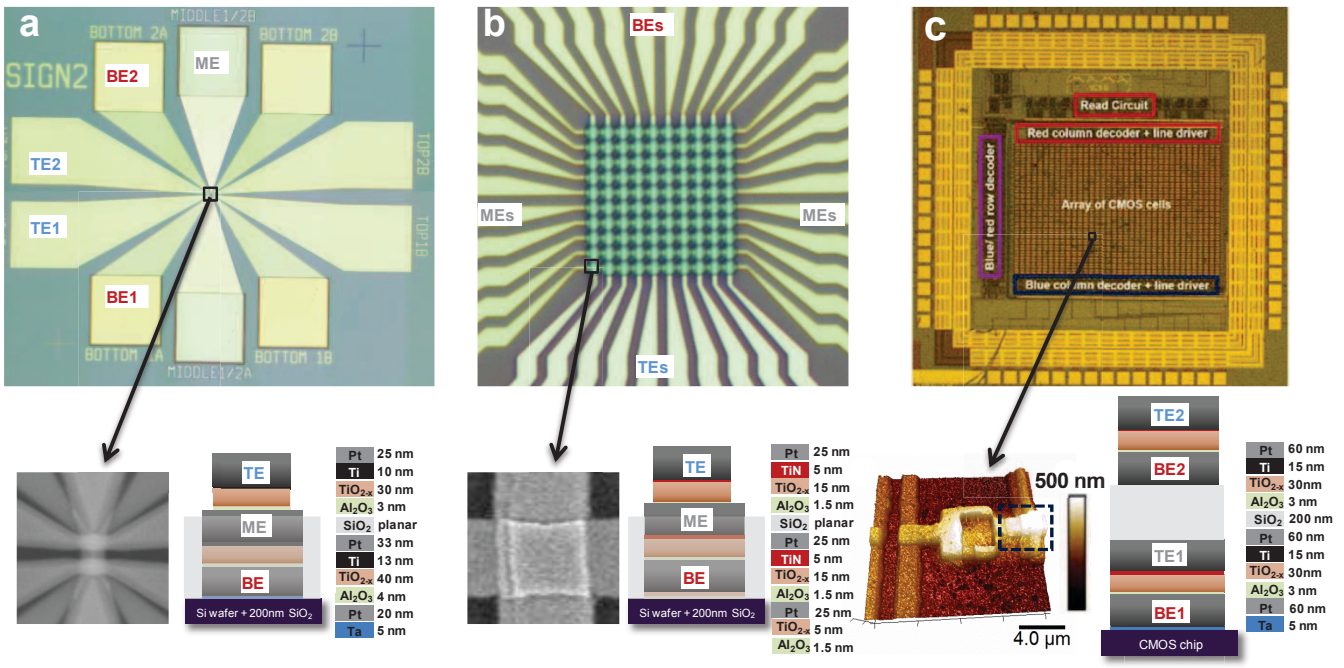

Fig. 1. Three-dimensional ReRAM fabrication and integration. (a) Standalone 2x2 arrays, showing an optical microscope image and a SEM zoom of one array together with a sketch detailing the device layers and thicknesses [9]; (b) Standalone 2x10x10 crossbars showing an optical microscope image and a SEM zoom of one device together with a sketch detailing the device layers and thicknesses [12]; (c) Devices and crossbars integrated monolithically onto CMOS, showing the foundry processed CMOS chip with arrays of CMOS cells used for accessing the two layers of ReRAM devices. A device stack is shown in the AFM detail while the layer thicknesses are presented in the sketch [13].

process could be scaled to multiple layers as required for high density systems.

The fabrication processes developed for these stand-alone three-dimensional arrays and crossbars were developed with a <175°C thermal budget, compatible with the CMOS integration. We reported a hybrid 3D circuit based on the CMOL (CMOS + "Molecular") architecture [13] with two layers of ReRAM crossbars monolithically integrated on a pre-fabricated CMOS substrate (Fig. 1c) [14-15]. The layers of memristive crossbars were added via post processing of the 5mm x 5mm CMOS chip, firstly by planarizing the chip using chemical mechanical polishing to create a suitable smooth surface for the ReRAM stacks, then etching 4µm x 4 µm vias to connect electrically the CMOS chip with the ReRAM access lines. Similarly to the stand-alone crossbars, reactive sputtered Al_2O_3/TiO_{2-x} were also used as active layer for both of the ReRAM stacks. The back-end-of-line process of monolithically integrating devices onto foundry CMOS chips comes with several fabrication challenges, when the available chips have small dimensions and therefore are hard to handle. Such small chips lead to non-uniformity issues, from the edge-beads of the spinned photoresist to challenges during the chemical mechanical polishing, for which we used a 4" carrier wafer with a pre-etched slot. To simplify the already challenging fabrication for this prototype, the two ReRAM device layers did not share an electrode and were separated by an 200nm SiO_2 layer isolating against electrical and thermal crosstalk.

III. Characterization

All stand-alone devices were characterized using an Agilent B1500 Semiconductor Device Parameter Analyzer system controlled by a computer via custom Visual C++ code. An Agilent B1530A Waveform Generator/Fast Measurement Unit was used to generate voltage and current pulses used to test to test the digital and analog tuning capabilities of the devices. The stand-alone crossbars were wirebonded on a custom PCB board and controlled via a low-leakage Agilent E5250A Switch Matrix to selectively bias the lines and characterize all the devices. The ReRAM devices monolithically integrated onto CMOS chips, were controlled entirely through the CMOS circuitry. The hybrid chip was wire-bonded and controlled via a custom made PCB board and a custom interface that allowed access to each ReRAM device. Two access modes were available: a) via a 12 bit address to select a cross-point device within one multi-cell in the CMOS array and b) via a 8 bit address to select 8 multi-cell columns in parallel.

All the devices show similar behavior in both bottom and top layers (Fig. 2a-c) and reproducible operation range below

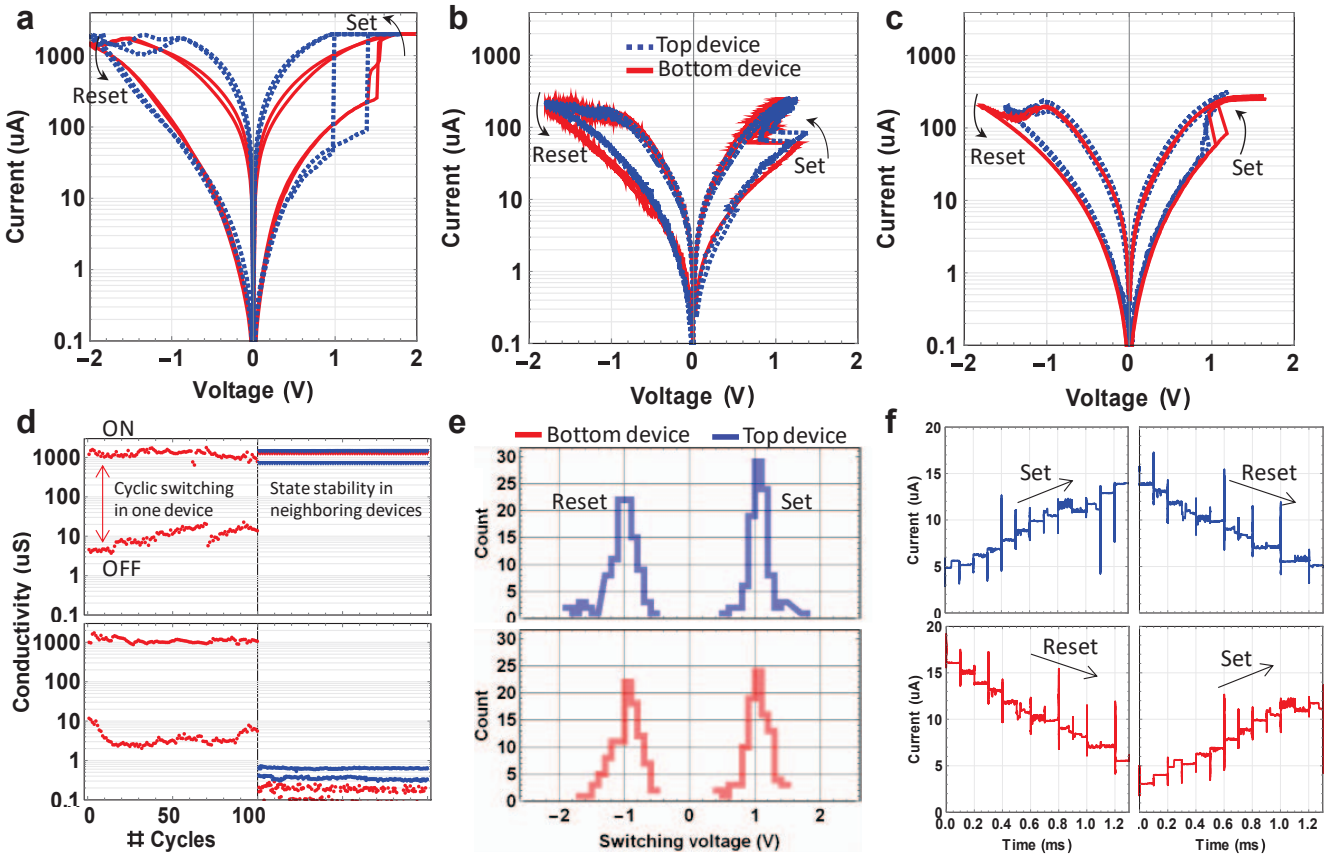

Fig. 2. Three-dimensional ReRAM electrical characterization (a) Typical I-V characteristics of the bottom vs. top devices in the 3D 2x2 arrays [9]; (b) Typical I-V characteristics of the bottom (red solid) vs. top devices (blue dashed) in the 3D 2x10x10 crossbars [12]; (c) Typical I-V characteristics of the bottom vs. top devices monolithically integrated and entirely accessed through the CMOS chip [13]; (d) No thermal cross-talk between the bottom and top devices, as example the devices in the array are used. A bottom device is cycled twice 100 times between ON and OFF state while the state stability of the neighboring bottom and top devices is measured for both ON and OFF states, respecively. (e) Device variation in the set and reset threshold voltages for both bottom and top devices, as exemplified on standalone crossbar devcices. (g) Analog tuning characteristics for devices monolithically integrated with CMOS. The tuning is exemplified for both devices in both layers and for set and reset.

±2V and ±2mA. The crossbar and CMOS integrated devices showed lower operating voltages and currents due to the scaling of the device, which is advantageous for energy efficient operation. The ON/OFF ratio was around 10 to 100. Given the thin films utilized for the device fabrication, thermal crosstalk was a potential concern [16], but experimental results (Fig. 2d) shows good state stability of the neighboring devices while a device is being switched.

Device variation, both device-to-device and cycle-to-cycle, is of concern for both digital and analog applications. The devices in the lift-off based arrays showed a fairly significant standard deviation ~0.14V in the set voltage for bottom and ~0.19V for the top device layer. For the devices in the ion-milled stand alone crossbars, the device variation was equally significant, with standard deviations for set of 0.2V and 0.17V respectively (Fig. 2e), which hints at the need for more material engineering to control the variability in the future. Both the stand-alone and CMOS integrated devices in the crossbars could be easily tuned using a modified version of the high precision tuning algorithm [17] from 8 to 16 states (Fig. 2f and [12]). It is worth noticing that device endurance to full ON/OFF swing is typically lower than the analog tuning in small steps, but the devices were functional for thousands of cycles as required for the implementation of applications. Long pulses of 500 µs were used for tuning, since the purpose of this work was not to investigate the device switching time.

IV. Applications

A. 3D Implication logic

ReRAM devices have been proposed for ultra-dense memory technologies due to their promising non-volatility, fast switching and endurance [18]. The tunable resistance of ReRAM is also suitable for stateful logic, combining the state storing with the state processing in one hardware system [19]. However the device variations pose challenges for conditional switching over many switching cycles. Three dimensional stacking, while beneficial to achieve ultra-dense memory systems, complicates the fabrication process thus potentially increasing device variability.

We demonstrated an optimized circuit configuration (Fig. 3a) [9] and proved its reliability by experimentally showing a multi-cycle multi-gate material implication logic operation (Fig. 3b and c) within the 3D passive array of monolithically integrated memristors with no selectors. Three dimensional data manipulation could enable compact and high-throughput in-memory computing, but issues of device endurance and incomplete conditional switching would have to be solved first.

B. Physically unclonable functions

While the device variability poses a significant problem for memory technology and stateful logic, it can be used as an advantage for hardware security primitives. However, the variability has to be enough to tune the devices to the desired distribution of conductances. The stand-alone stacked ReRAM crossbars were used as compact and highly non-linear programmable analog instances to implement PUF security primitives (Fig. 4) [20]. By comparison with previous work [21], the demonstrated PUF can be reconfigurable and exhibited almost ideal randomness successfully passing the NIST randomness test.

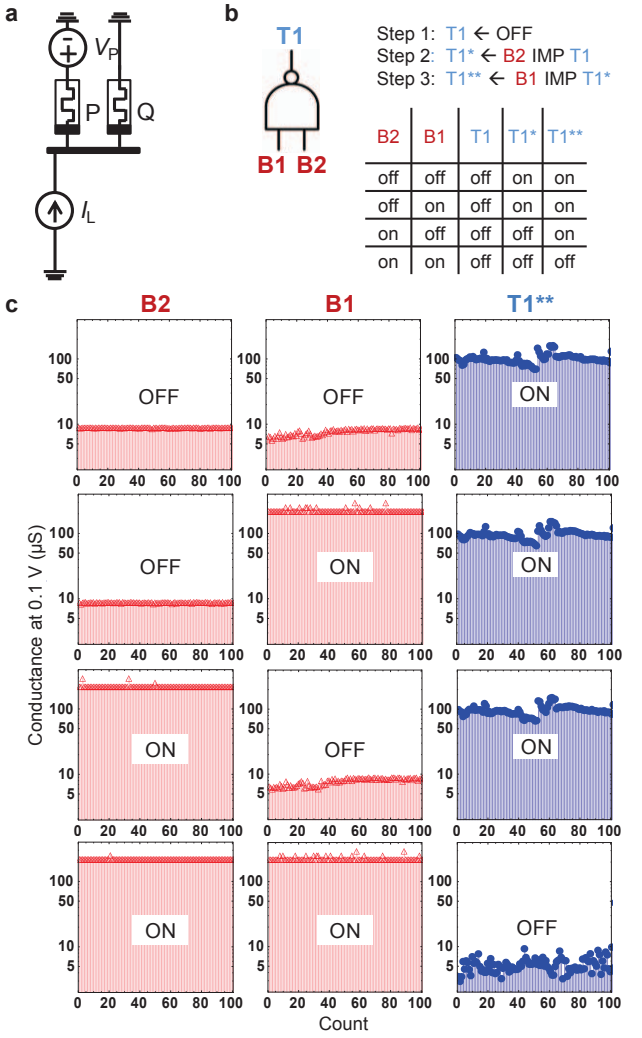

Fig. 3. Multi-cycle operation of three dimensional NAND gate implemented using memristive-based implication logic (a) Proposed optimized circuit for implication logic (b) Three dimensional NAND gate symbol and implementation steps (c) Experimental data showing 100 cycles of operation.

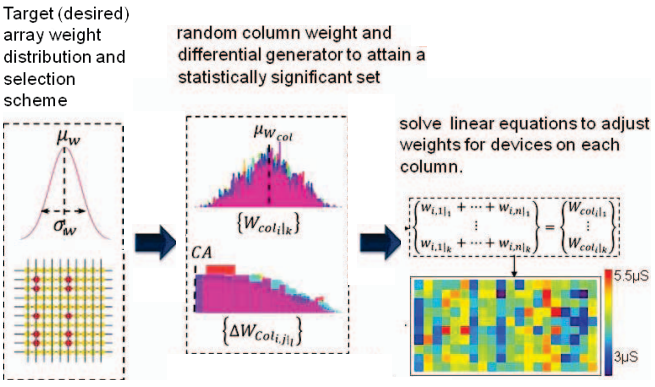

Fig. 4. Proposed ReRAM-based physically unclonable function (PUF) using the state variability and the non-linearity of the device as a source of unpredictability [20]. For the experimental demonstration, a 2x10x10 three-dimensional crossbar was used for compactness. This implementation has the advantage of reconfigurability.

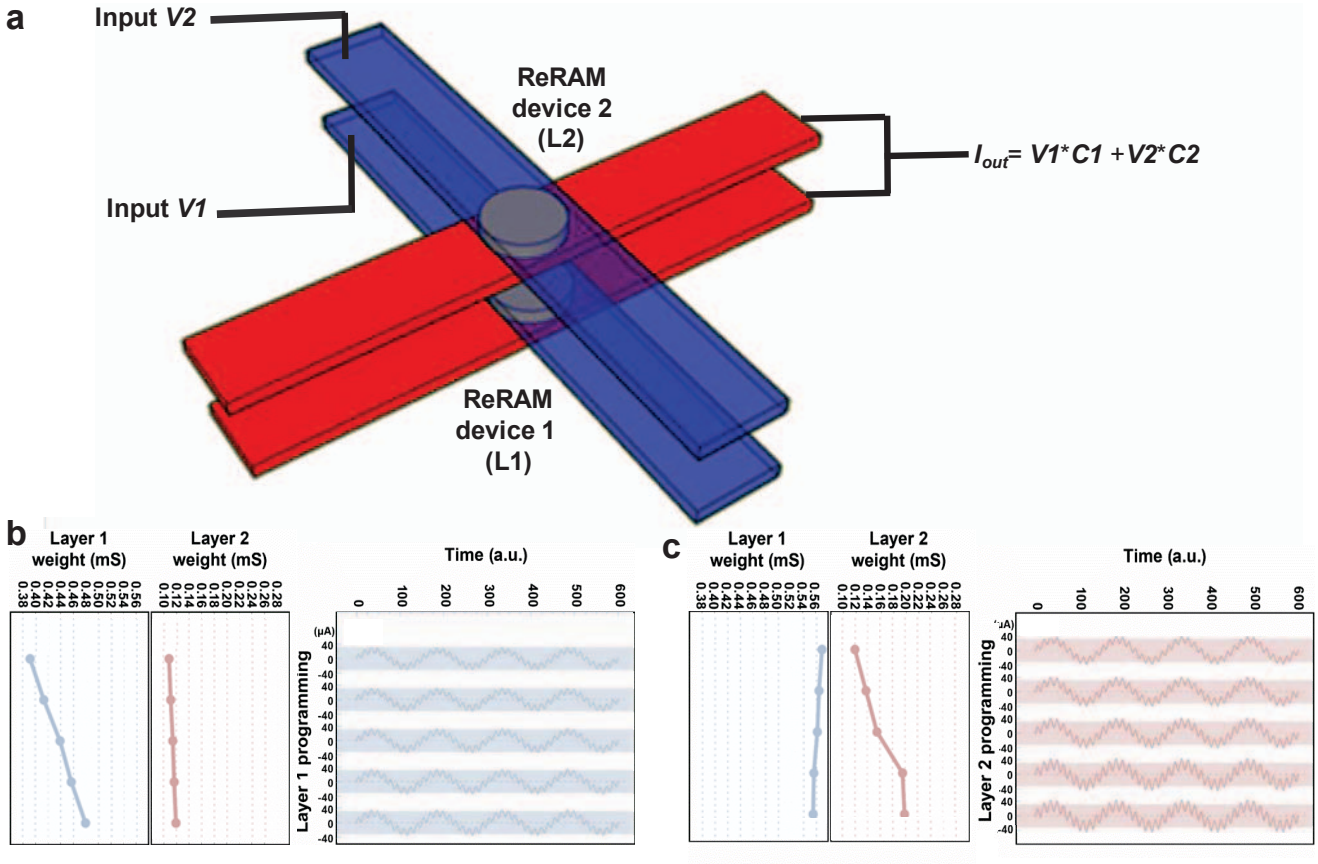

Fig. 5. Dot-product operation implemented with the three-dimensional hybrid CMOS/ReRAM chip (a) Sketch describing the simplest operation, with two sinusoidal voltage imputs V1 and V2 of the same amplitude (=300mV), but with V2 having 10 times a higher frequency than V1. The ReRAM device conductances are tuned according to the desired output. The output is represented by the weighted current, which implements at the physical level the dot product between the inputs and the device conductances; (b) The device in layer 1 is tuned to have increasing conductance, while the device in layer 2 has a constant conductance while in (c) the device 1's state is constant while the device 2 has an increasing conductance as observed in the output waveform.

C. *Dot-product*

We reported the first experimental implementation of a functional three-dimensional CMOL circuit performing dot-product operation (Fig 5) [13]. Dot-product operations were performed to demonstrate the applicability of 3D CMOL circuits as a multiply-add engine. For simplicity, two devices, one in each layer, were used for the implementation. As inputs were applied two sinusoidal voltages V1 and V2 of amplitude 300mV, but of different frequencies (V2 has 10x the frequency of V1). Firstly, the device in layer 1 was programmed to lower conductance, while the one in layer 2 kept its conductance constant. Secondly, the device state in layer 1 was kept constant while the one in layer 2 change its state. The output was the weighted current summation of the inputs and the device states.

V. Conclusions

In summary, we have presented our experimental demonstrations of monolithically integrated passive resistive arrays and crossbars, standalone and onto CMOS circuitry. Planarization based on chemical mechanical polishing and dry etch-back was used to provide a smooth surface for stacking resistive devices on top of each other or onto foundry CMOS chips. An optimized circuit for implication logic implemented using stacked memristive arrays showed reliable multi-cycle operation despite the device switching variability. The stand-alone 3D crossbars showed good uniformity and were used for the implementation of PUF security primitives with almost ideal randomness. The 3D memristive crossbars monolithically integrated onto CMOS circuitry implemented a multiply-add engine as the first functional 3D CMOL hybrid circuit demonstration.

As future work, the device-to-device and cycle-to-cycle variability needs to be improved through material engineering. Another important aspect is the necessity of a two-terminal passive selector device monolithically integrated with the ReRAM device. By increasing the number of layers, especially if the electrodes are shared between crossbars, the equivalent dimension of the crossbar is increased [22] so the sneak currents can become significant unless controlled via a high non-linear selector device.

REFERENCES

[1] K. K. Likharev, "Hybrid CMOS/nanoelectronic circuits: Opportunities and challenges", J. Nanoelectronics and Optoelectronics, vol. 3, pp. 203-230, 2008.

[2] T.-Y. Liu et al., "A 130.7 mm2 2-Layer 32-Gb ReRAM memory device in 24-nm technology", IEEE Journal Solid-State Circuits, 49 (1), pp. 140-153, 2014.

[3] M.J. Lee et al. "2-stack 1D-1R cross-point structure with oxide diodes as switch elements for high density resistance RAM applications" IEDM, pp.771-774, 2007.

[4] S. Yu et al., "HfOx-based vertical resistive switching random access memory suitable for bit-cost-effective three-dimensional cross-point architecture", ACS Nano, vol. 7, pp. 2320-2325, 2013.

[5] P.Y. Chen, Z. Li, S. Yu, "Design tradeoffs of vertical RRAM-based 3-D cross-point array", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, PP (99), 2016.

[6] H. Li et al. "Four-layer 3D vertical RRAM integrated with FinFET as a versatile computing unit for brain-inspired cognitive information processing", IEEE Symposium on VLSI Technology, 2016.

[7] K. H. Kim et. al. "A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications", Nano Letters, vol. 12 (1), pp. 389-395, 2011.

[8] Q. Xia et al. Memristor−CMOS Hybrid Integrated Circuits for Reconfigurable Logic, Nano lett. 9, 3640-3645, 2009.

[9] G.C. Adam, D.B. Hoskins, M. Prezioso, D.B. Strukov, D.B., 2016. Optimized stateful material implication logic for three-dimensional data manipulation. Nano Research, vol.9, no.12, pp.3914-3923, 2016.

[10] B.D. Hoskins, D.B. Strukov, "Maximizing stoichiometry control in reactive sputter deposition of TiO_2". Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films, 35(2), p.020606, 2017.

[11] M. Prezioso, F. Merrikh Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors", Nature, vol. 521, pp. 61-64, 2015.

[12] G.C. Adam, et al. 3-D memristor crossbars for analog and neuromorphic computing applications. IEEE Transactions on Electron Devices vol. 64, pp. 312-318, 2017.

[13] B. Chakrabarti et al. "A multiply-add engine with monolithically integrated 3D memristor crossbar/CMOS hybrid circuit", Scientific Reports, 7, 42429, 2017.

[14] M. Payvand, A. Madhavan, M.A. Lastras-Montaño, A. Ghofrani, J. Rofeh, K.T. Cheng, D. Strukov, L. Theogarajan, "A configurable CMOS memory platform for 3D-integrated memristors". Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), May 2015, pp. 1378-1381.

[15] M.A. Lastras-Montano, A. Ghofrani, K.T Cheng, "Architecting energy efficient crossbar-based memristive random-access memories" Proceedings of the IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH), 2015.

[16] P. Sun et al., "Thermal crosstalk in 3-dimensional RRAM crossbar array", Nature Scientific Reports, vol. 5, art. 13504, 2015.

[17] F. Alibart, L. Gao, B.D. Hoskins, and D.B. Strukov, "High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm", Nanotechnology , vol. 23, 075201, 2012.

[18] M.J. Lee et al. "A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5− x/TaO2− x bilayer structures" Nature materials, vol. 10, no. 8, pp.625-630, 2011.

[19] Borghetti, J.; Snider, G. S.; Kuekes, P. J.; Yang, J. J.; Stewart, D. R.; Williams, R. Set al. "Memristive" switches enable "stateful" logic operations via material implication. Nature 2010, 464, 873–876.

[20] H. Nili et al. "Highly-Secure Physically Unclonable Cryptographic Primitives Using Nonlinear Conductance and Analog State Tuning in Memristive Crossbar Arrays", arXiv preprint arXiv:1611.07946, 2016.

[21] L. Gao, P.-Y. Chen, R. Liu, and S. Yu, "Physical unclonable function exploiting sneak paths in resistive cross-point array", IEEE Transactions on Electron Devices , vol 63, pp. 3109-3115, 2016.

[22] Xia, L. et al. Technological Exploration of RRAM Crossbar Array for Matrix-Vector Multiplication. J. Comp. Sc. Tech. 31, 3-19 (2016).