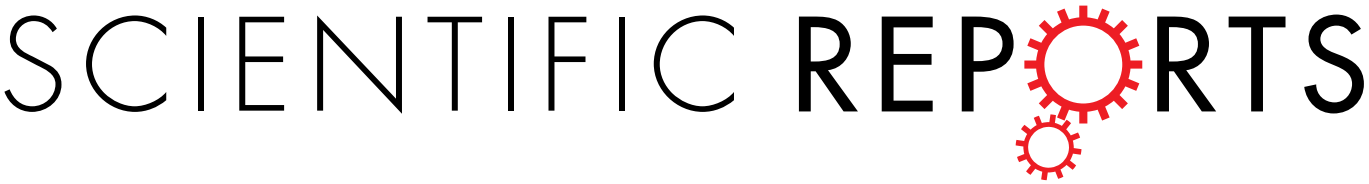


OPEN

A multiply-add engine with monolithically integrated 3D memristor crossbar/CMOS hybrid circuit

Received: 30 September 2016
Accepted: 09 January 2017
Published: 14 February 2017

B. Chakrabarti[1], M. A. Lastras-Montaño[1], G. Adam[1], M. Prezioso[1], B. Hoskins[2], K.-T. Cheng[1,3] & D. B. Strukov[1]

Silicon (Si) based complementary metal-oxide semiconductor (CMOS) technology has been the driving force of the information-technology revolution. However, scaling of CMOS technology as per Moore's law has reached a serious bottleneck. Among the emerging technologies memristive devices can be promising for both memory as well as computing applications. Hybrid CMOS/memristor circuits with CMOL (CMOS + "Molecular") architecture have been proposed to combine the extremely high density of the memristive devices with the robustness of CMOS technology, leading to terabit-scale memory and extremely efficient computing paradigm. In this work, we demonstrate a hybrid 3D CMOL circuit with 2 layers of memristive crossbars monolithically integrated on a pre-fabricated CMOS substrate. The integrated crossbars can be fully operated through the underlying CMOS circuitry. The memristive devices in both layers exhibit analog switching behavior with controlled tunability and stable multi-level operation. We perform dot-product operations with the 2D and 3D memristive crossbars to demonstrate the applicability of such 3D CMOL hybrid circuits as a multiply-add engine. To the best of our knowledge this is the first demonstration of a functional 3D CMOL hybrid circuit.

Resistance switching or memristive devices are metal-insulator-metal structures that can switch between at least two different resistance states upon application of an electrical impulse (voltage or current). Although the phenomenon of resistance switching has been known since the 1960 s, research interest has significantly grown in the last decade after resistance switching devices were identified not only as one of the leading candidates for next generation memory[1–7] but also for analog computation[8–10], neuromorphic circuits[11–15], reconfigurable logic[16,17] and other applications[18,19]. However, practical applications of memristive devices, which are passive circuit elements, may often need integration with active CMOS components. Recently a hybrid architecture termed as "CMOL" (CMOS + molecular devices) has been proposed that can combine the novel functionalities of memristive devices with CMOS technology to potentially lead to ultra-high density memory, reconfigurable logic and neuromorphic applications[20–22]. Performance of such hybrid circuits can be further improved manifold by integration of 3D memristive crossbars on CMOS technology. Reports of memristor-CMOS integration where memristors were fabricated between the two metal layers of a CMOS process exist in the literature[5,23]. H. Li *et al.* has recently shown 3D vertical resistive memory devices integrated with FinFET selectors[24]. The integrated devices were used to demonstrate in-memory computation capability. However, demonstrations of CMOL hybrid circuits has been extremely rare[16,25]. Xia *et al.* demonstrated reconfigurable logic functionalities with memristive components integrated on a CMOS substrate. However, only 2D planar devices were integrated. Moreover, discussions on the quality of interface between the CMOS and memristors as well as the memristor device characteristics were also very limited. We have been recently able to show 3D memristive crossbars with analog computation capability although the memristive crossbars were not integrated on a CMOS substrate[26]. In this work, we demonstrate vertical monolithic integration of 3D memristive crossbars on a foundry-processed 5 mm × 5 mm CMOS chip. High integration yield with low contact resistance between the CMOS layer and memristive devices was

[1]Electrical and Computer Engineering Department, University of California, Santa Barbara, CA, 93106, USA. [2]Materials Department, University of California, Santa Barbara, CA, 93106, USA. [3]School of Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Correspondence and requests for materials should be addressed to B.C. (email: bchakrabarti@ece.ucsb.edu)

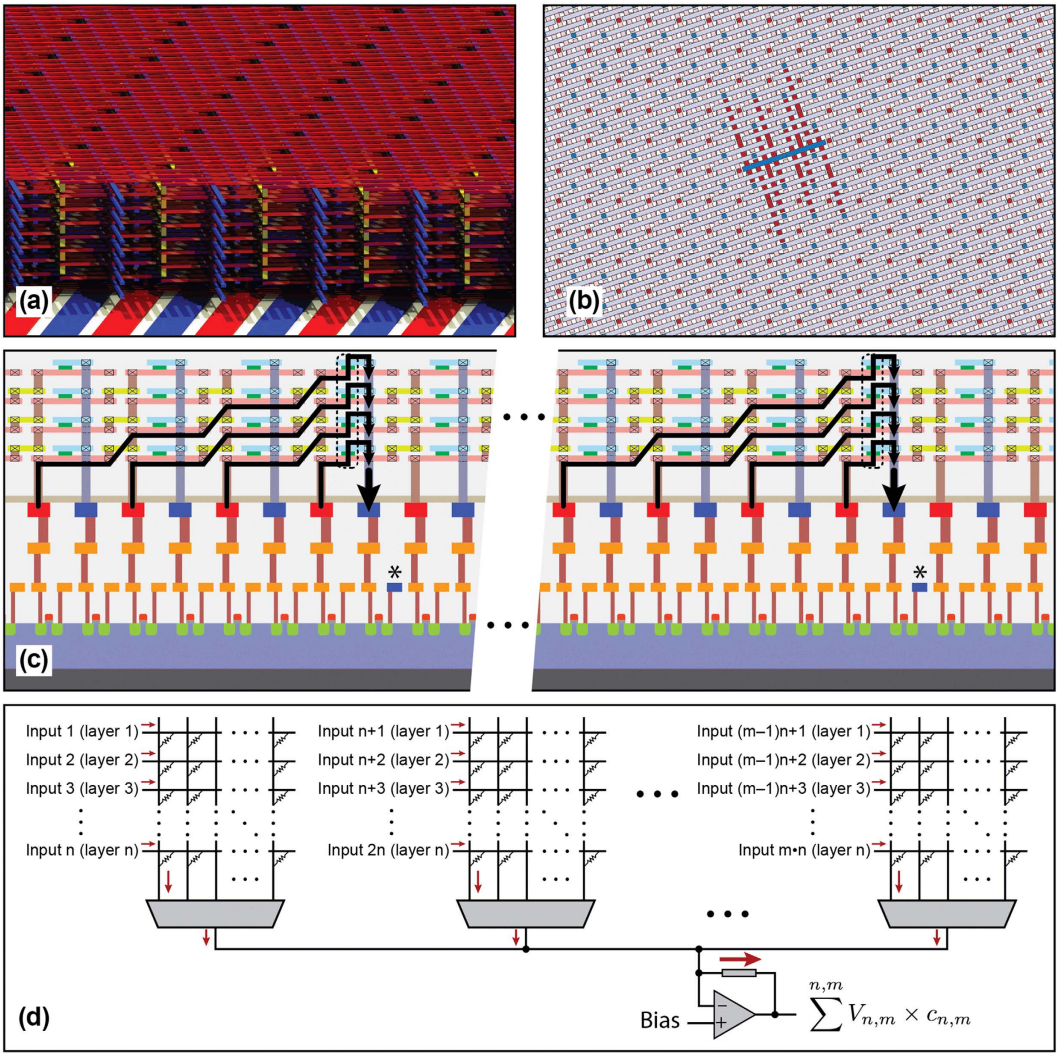


Figure 1. Conceptual representation of a 3D CMOL multiply-add engine. (**a**) Conceptual view of a 3D multi-layer crossbar integrated on CMOS substrate. The red and blue thinner wires represent the bottom and top electrodes of the crossbar respectively, the thicker red and blue wires illustrate the CMOS interconnections. (**b**) Top-view of one layer of a CMOL crossbar: the red and blue dots represent the contact vias for the bottom (BE) and top electrodes (TE) respectively. The highlighted electrodes (one TE and eight BE) demonstrate the CMOL connectivity. (**c**) Cross-sectional view of (**b**) showing multiply-add operation at the 'blue' pin of the eight input signals fed through the eight 'red' pins. (**d**) Multiply-add operation in a (m × m × n) 3D crossbar where n is the number of layers.

achieved. We demonstrate the dot-product operation, a critical computation for many applications involving linear transforms, using the 3D memristive crossbars with high-precision tunable analog devices. We believe that this work is a significant step towards practical realizations of highly efficient 3D hybrid CMOL circuits.

Ultra-high bandwidth multiply-add engine with 3D CMOL hybrid circuits

Figure 1 illustrates a high bandwidth multiply-add engine using a 3D CMOL hybrid circuit. A memristive crossbar array can naturally perform dot-product operation. If an array of voltage signals is applied to the rows of a crossbar, the current measured at a column will be a weighted summation of the inputs with each input signal being multiplied by the conductance or 'weight' of the corresponding cross-point memristive device. Multiply-add operation is performed concurrently in all the layers as well within each layer of crossbar and the resulting currents are summed at the output using the CMOS circuitry. Such massive parallelism and the effectiveness of CMOL area-distributed interface enable very high bandwidth computations, which can be further improved in 3D CMOL circuits[27]. Figure 1a is an illustration of a 3D CMOL crossbar with 8 layers, while Fig. 1b shows one layer of a CMOL crossbar. The red and blue via openings provide connectivity to the area-distributed interface of the CMOS substrate for the BE and TE arrays of the memristive crossbar. Each TE (blue line) in this CMOL architecture is connected to eight BE (red line). Figure 1c is a partial cross-sectional view of the 3D CMOL in Fig. 1a showing only 4 layers. In each layer 8 BE rows are connected to a TE column through 8 memristive

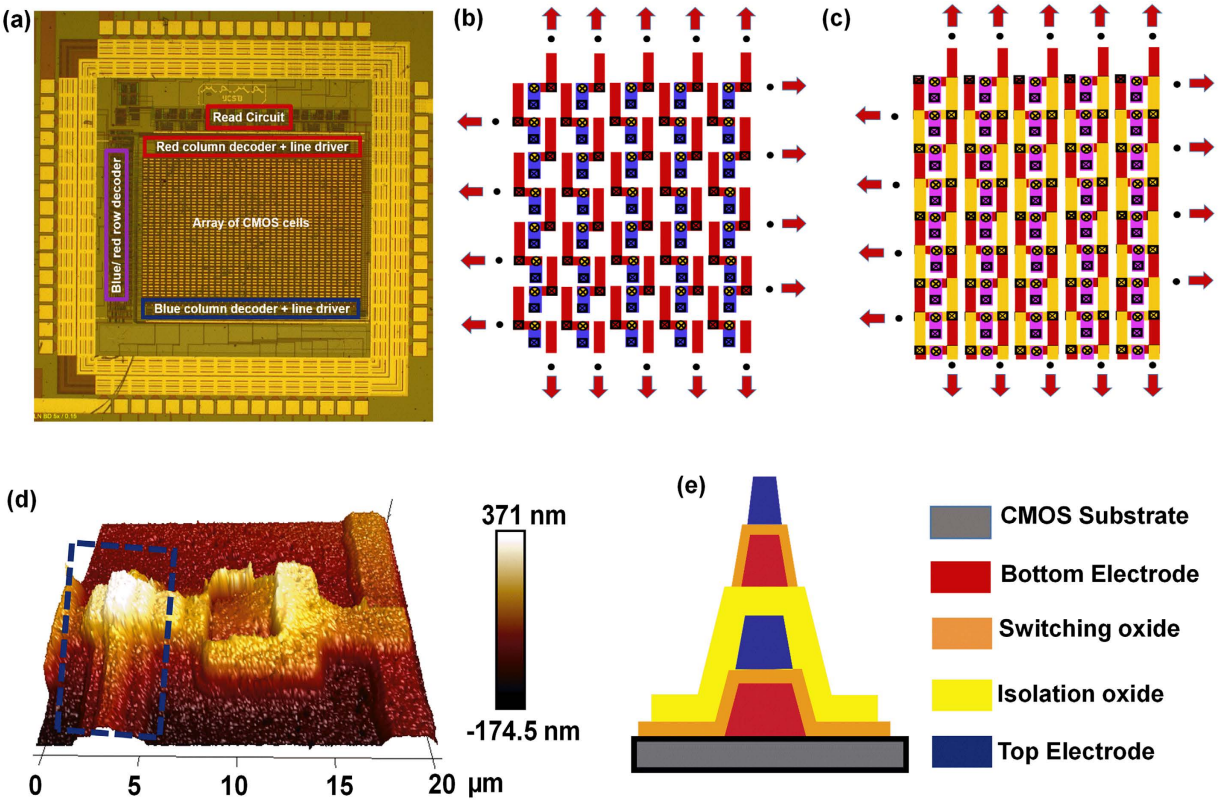


Figure 2. Fabrication of a 3D CMOL crossbar. (**a**) Optical image of the foundry-processed CMOS chip showing the on-chip decoder, 'Read' and 'Write' circuitry as well as the 24 × 36 array of CMOS cells. Each 'CMOS cell' houses a pair of pins ('Red' and 'Blue'), (**b**) structure of the first layer of the 3D crossbar, (**c**) the final structure of the 3D crossbar with 2 layers (**d**) AFM image of a section of the 3D crossbar: the highlighted section shows two stacked device layers, (**e**) Cross-section of the region highlighted in (**d**).

devices while the TE columns are connected to one another through 8 vertically stacked devices. The overall multiply-add operation therefore has 2-dimensional and 3-dimensional components. Within one layer each TE line receives the weighted input signals from eight BE lines, each weight corresponding to the conductance state of the individual memristors. The output signals generated at each TE line in each layer are then summed at the TE column over all the layers. Figure 1d shows the equivalent circuit representation of the dot product operation in the 3D crossbar.

In the next sections we will discuss practical implementation of a 3D CMOL multiply-add engine. For demonstration purposes our 3D CMOL hybrid circuit consists of 2 layers of memristive crossbars. However, this process can be extended to fabricate as many crossbar layers as needed.

Results

3D Crossbar integration on CMOS substrate. Figure 2a shows an optical image of a typical CMOS chip designed for fabrication of the hybrid circuits. The chip, a 5 mm × 5 mm die was fabricated in a commercial foundry with 0.5 μm technology. The area-distributed interface required for the CMOL architecture consists of an array of 'CMOS cells' in the top metal level (Al) with each cell containing a pair of pads (Red and Blue). The chip also houses the decoder, 'Read' and 'Write' circuits for addressing as well as writing and sensing operations on the integrated memristive crossbars. More details of the chip level architecture and the 'Read' and 'Write' circuitry on the CMOS chip is discussed in the next section. A scratch-protect oxynitride layer covers the CMOS pads. The topography on the top surface (see Supplementary Figure 1a,b) originates from the thickness of the Al pads in the top metal layer. The severe topography over the area-distributed interface (>800 nm) necessitates planarization before memristive crossbars can be fabricated on top. Among the different planarization techniques available only chemical mechanical planarization (CMP) provides both local and global planarity. However, CMP is suitable for wafer level applications and reports of CMP on die level are extremely rare. In one demonstration approximately 5 × 5 mm^2 chips were planarized using multiple 10 × 10 mm^2 dummy Si pieces around the chip[28]. However this method is unlikely to produce reproducible results as the relative height distribution between the dummies and the actual sample will vary between experiments. To achieve reproducible planarization we use a 4 inch Si wafer with a cavity slightly larger than the die. The thickness of this wafer (holder) was carefully adjusted so that the holder is approximately 3–4 μm thinner than the CMOS chip (see methods and Supplementary Figure 2 for details). The scratch-protect oxynitride on the CMOS chip was removed and 2.5 μm silicon dioxide (SiO_2) was deposited before polishing (see methods and Supplementary Figure 3). Both the holder and the chip were

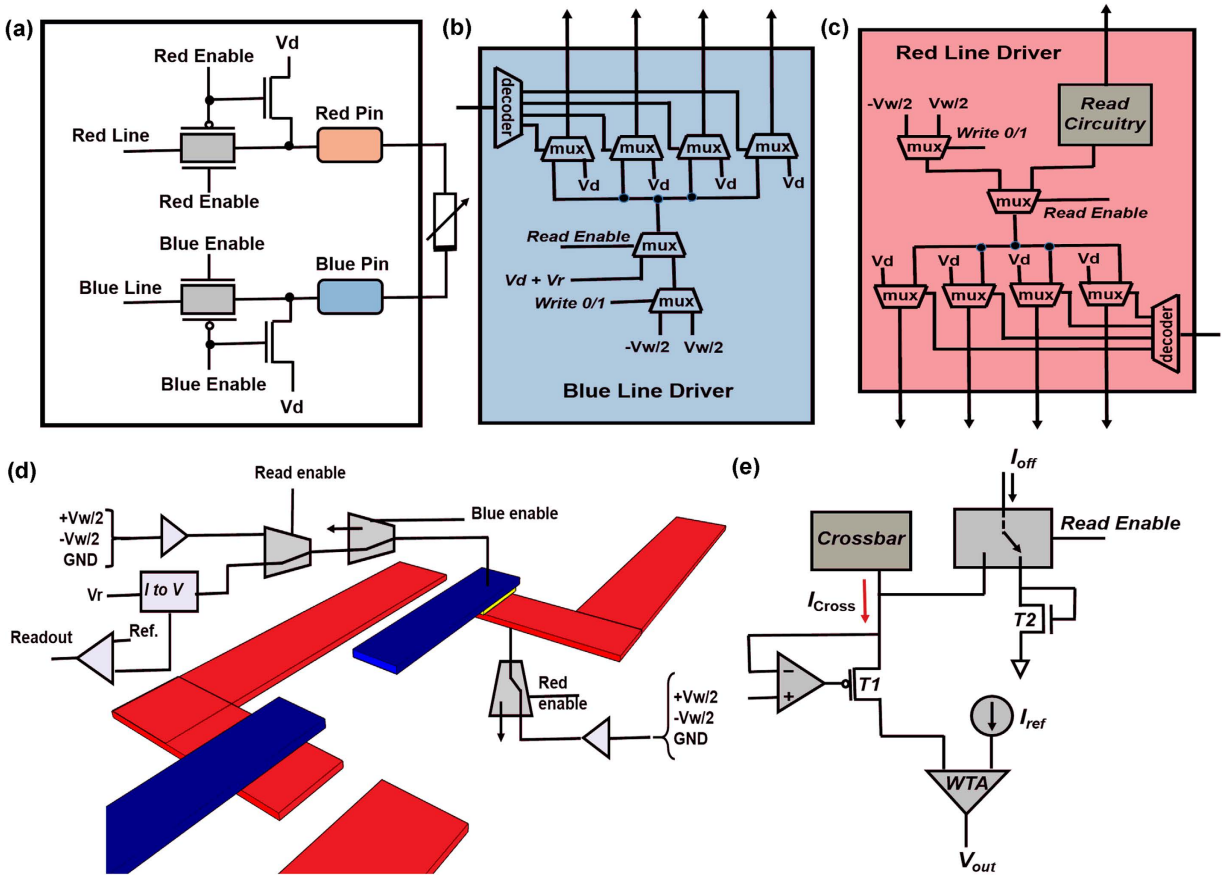


Figure 3. CMOS circuitry. (**a**) Circuit of a CMOS cell. (**b**,**c**) Blue and Red line driver for supplying the appropriate biases to the Blue and Red pins respectively. (**d**) Cartoon of the write/read operation for the memristive crossbars. (**e**) Schematic of the sensing circuit.

then attached on to a second Si carrier wafer by crystal bond (Supplementary Figures 2 and 3) and the whole assembly was polished by CMP. Planarization was performed in two steps. The first step employs a fast polishing that reduces the topography from >800 nm to ~30–40 nm (Supplementary Figure 4b,c). In the second step the topography is further reduced by a slow polish down to <10 nm (See Supplementary Figure 4d,e). The surface roughness after planarization process can be a key factor for reliable device performance. In our experiments, the post-CMP surface roughness (<4 nm, see Supplementary Figure 6) is negligible compared to the thickness of the switching oxide (~33 nm) and creation of false devices due to roughness can be ruled out. The memristive crossbars were fabricated on top after the planarization (see methods and Supplementary Figure 7 for details). Note that in this experiment we performed the planarization process once to remove the initial topography originating from the CMOS top-metal layer since we fabricated only 2 layers for demonstration. However, for integration of more than 2 layers where each layer would add further topography additional planarizations after fabrication of each crossbar layer would be required.

Figure 2b,c show the structure of the integrated 3D crossbar. Figure 2b shows a section of the first layer of the crossbar with the arrows indicating continuation of the structure in the indicated directions. Here the red and blue lines indicate the BE and TE of layer 1. Each electrode (BE or TE) is connected to an underlying CMOS cell through a BE or a TE via. For the BEs, the via is through the planarized SiO_2 layer while for the TE the via opening is through the switching oxide+ the SiO_2 layer (see methods for fabrication details). The BEs of the 2nd layer (yellow lines) are connected to the BEs of the 1st layer by a set of vias through the isolation oxide (SiO_2) between the 1st and the 2nd layer. The TEs of the 2nd layer are fabricated on top of the TEs of the 1st layer. As such, this structure allows integration of more than 2 layers without any additional lithography masks. For example a 3rd crossbar layer will have the BEs in the same positions as in the BEs of the 1st layer (Fig. 2c) while the TEs of the 3rd layer can be fabricated on top of the TEs of the 1st and 2nd layers. For test purposes we also fabricated single devices and 2-dimensional crossbars (See Supplementary Figure 8). The devices (both for 2D and 3D crossbars) employ a bilayer Al_2O_3/TiO_{2-x} dielectric as the switching material. The non-stoichiometry of the TiO_{2-x} layer is controlled by the sputtering conditions. The devices have Ta/Pt BEs and Ti/Pt TEs.

CMOS circuit details. As mentioned before, addressing the memristive crossbars as well as the memory operations are all performed through built-in circuitry in the CMOS chip itself. Figure 3a shows the circuit details of a CMOS cell which contains two transmission gates that connect the red and blue pins to the red and blue lines respectively when the gates are enabled. The red and blue lines provide the appropriate read/write voltages to the

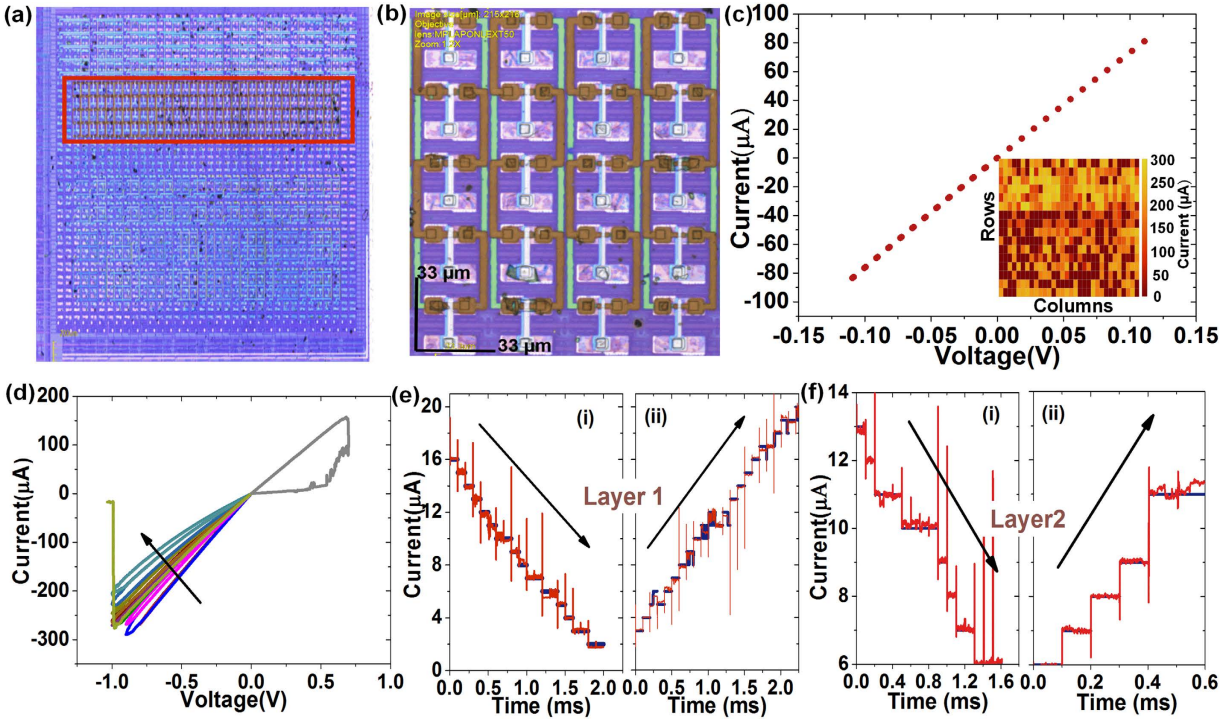


Figure 4. Electrical characteristics. (**a**) An optical image of the CMOS chip with vertically integrated memristive crossbars: the region highlighted in the red box has the 3D CMOL crossbar structures, (**b**) A high-resolution optical image of a section of the integrated 3D crossbars. (**c**) A typical *I-V* plot measured on a test structure to test the contact resistance between the memristive layer and the underlying CMOS. Inset shows a map of the current measured at all the 24 × 36 array of 'red pins' showing contact between the integrated devices and the CMOS cells. Current is measured at 0.2 V for each measurement. (**d**) An I-V plot showing typical DC switching characteristics of the integrated memristive devices: the device is initially turned on and is gradually reset demonstrating analog switching behavior, (**e**) pulsed switching characteristic of a device in layer 1 showing gradual reset and set operation: the device is tuned to each desired state using a tuning algorithm (discussed in the supporting info.). The arrow indicates the direction of programming in each case, (**f**) pulsed switching characteristic of a device in layer 2 showing gradual reset and set operation: the device is tuned to each desired state using the tuning algorithm.

pins. Both the red and blue pins are connected to a default voltage V_d when the transmission gates are disabled. The red and blue drivers shown in Fig. 3b,c supply the required voltages to the red and blue lines. Depending on the memory operation (read/write) the blue and red drivers can supply a read voltage of V_r and write voltage of $\pm V_w/2$ in both the drivers. Row/column decoders are employed to select the appropriate 'Blue' and 'Red' pins for applying 'read'/'write' bias to the desired crosspoint (s). The row and column decoders on the chip surround the array of CMOS cells (see Fig. 2a). The entire CMOS cell array is divided into many units called multi-cells. A unique crosspoint within one multi-cell can be chosen using a 12 bit address provided as user input through the interface we have developed. At a time eight such crosspoints corresponding to eight multi-cell columns can be selected concurrently, leading to eight bit operation (Supplementary Figure 9). More details on the addressing scheme for the memristive crossbars can also be found elsewhere[29]. Figure 3d shows a cartoon of the 'Read'/'Write' operations on the integrated crossbars. For simplicity only the first layer of the crossbar structure is shown (see also Fig. 2b). During the 'Write' operation a bias of either $+V_w/2$ (set) or $-V_w/2$ (reset) is applied through the blue line driver to the TE (blue wire), while the BE is connected to bias of $-V_w/2$ (set) or $(+V_w/2)$. This leads to application of a resultant bias of $+V_w$ (set) or $-V_w$ (reset) across the device. During a 'Read' operation a bias of $+V_r$ is applied to the TE (with respect to the BE). A schematic of the 'Read' circuitry is shown in Fig. 3e. When the Read Enable signal is asserted, the current from the crossbar I_{Cross} is sensed through a winner-take-all (WTA) circuit by comparing against a reference current I_{ref}. When the Read Enable signal is not asserted, a very small current I_{off} (50 pA) is passed through the WTA circuitry to avoid a delay in the loop for the Read operation[30].

Switching characteristics in 2D and 3D devices. Figure 4a shows an optical image of the processed CMOS chip. The highlighted region contains the integrated 3D crossbars. Integrated 2D crossbars and single devices can be seen in other regions on the area-distributed interface. A zoomed-in view showing a section of the integrated 3D crossbars can be seen in Fig. 4b. The underlying CMOS pads can be clearly seen as well. Figure 4c shows *I-V* characteristic of a test structure to evaluate contact resistance. The test structure has the BE and TE shorted with no switching dielectric in between. The measured resistance for the specific test structure is ~1.5 kΩ. Figure 4c inset shows a map of the measured current at all the contacts between the CMOS pads and the

bottom electrodes ('red' pins) of the 3D crossbars. More than 90% of the contacts (145 out of 160) have current of 100 μA or more at a V_{read} of 0.3 V indicating that most of the contacts between the memristive devices and the underlying CMOS substrate have contact resistance less than 3 kΩ. This is an important step towards realization of large-area CMOL circuits. Figure 4d shows typical DC switching characteristics of an integrated memristive device. The devices are initially in a low resistance state. Analog reset operation is observed as the reset-stop voltage is increased incrementally. A more abrupt reset operation can be achieved with sufficiently high reset bias (~−1 V). Similarly the devices can be turned on significantly with a positive bias ~0.8 V. No current compliance was used to control the set operation. The device can be turned more on to a less resistive state by applying a higher set bias. However, an internal compliance (~250 μA) employed by the CMOS circuitry protects the devices from permanent damage. This internal compliance is tunable through the CMOS circuitry. The switching bias is always applied at the TE while the BE is connected to electrical ground. It is to be noted that abrupt change in device conductance can occur, especially under stress at voltages similar to the maximum voltage applied for analog switching. For example the abrupt reset in Fig. 4d occurs after several consecutive analog reset steps. However, any abrupt change in the conductance is adjusted and the device is tuned to the desired state using a write-and-verify tuning algorithm. The algorithm employs pulse trains of increasing amplitude to incrementally increase/decrease the resistance of the devices with controllable precision. Figure 4e, f exhibit examples of high-precision multi-level tuning operation for the 1st and 2nd layer devices in a 3D crossbar, respectively (See Supplementary Figure 10 and Supplementary note 1 for a detailed description of the tuning algorithm). The tuning operation was optimized to minimize the number of iterations required to tune to each state. The number of pulses required to tune the device to each level was used as a measure of the tuning speed. Optimization was achieved by adjusting the user-defined inputs of the tuning procedure (Supplementary Figures 10 and 11). An optimized tuning algorithm with 10% precision was employed to tune the devices shown in Fig. 4e and f. The red lines in each figure indicate the actual current during the tuning operation while the blue lines indicate the desired states. Devices in both layers can be tuned to at least 8 levels with 10% precision (see also Supplementary Figure 12). Under the optimized conditions all the levels can be tuned within 150 applied pulses (Supplementary Figure 12). In a typical experiment we also tuned the devices to 8 distinct levels with each level being programmed 1000 times. All the levels exhibit clearly distinguishable distribution over 1000 switching operations showing stable endurance under the tuning procedure (Supplementary Figure 12). Stable room temperature retention was observed as well up to 10^4 seconds (Supplementary Figure 12). Note that the pulses applied in the tuning procedure have amplitudes significantly lower than the amplitudes required to turn the devices fully on/off (Supplementary Figure 13). As resistive switching devices typically have strongly non-linear switching dynamics with respect to applied voltage[31], it is expected that the stress caused by the tuning procedure is significantly less compared to the switching of the devices fully on/off.

Dot product operations with 3D crossbar devices. Figure 5 describes the multiply-add operation utilizing the 3D memristive crossbars. We would like to emphasize that the entire multiply-add operation is carried out through the in-built CMOS circuitry in accordance to the write and sensing schemes discussed previously. An entire 3D crossbar with 5 devices in each layer was employed for the operation. However, for the simplicity of demonstration we only programmed one device in each layer during the operation. Figure 5a represents the set-up used for the dot-product operation. Input voltage signal V1 is applied to the TE (Blue pin) of the layer 1 device (conductance C1) while signal V2 is applied to the TE of the device in layer 2 (conductance C2). The summation of the current outputs of the devices is carried out at the two connected BE (red) terminals. The output is a weighted summation of the input voltage signals. Figure 5b shows an example of the multiply-add operation with the bottom layer device being programmed to different states. The input signals for both bottom and top layer devices have an amplitude of 300 mV. However, the frequency of the signal for channel 2 is 10x more than that of channel 1. The evolution of the output current waveform shown in Fig. 5b reflects the multiply-add operation as the conductance of the 1st layer device (channel 1) varies from C1 to C2 to C3 in the decreasing order. With the reduction of weight for device 1 the amplitude of the low frequency component (envelope) of the output waveform reduces while the amplitude of the high-frequency component (ripples on the envelope) remains unchanged. Figure 5c provides more examples of the dot-product operations in another set of 3D memristive crossbars. In the experiment shown in Fig. 5c we first change the weight of the layer 1 device while keeping the state of the layer 2 device unchanged. Figure 5c(ii) shows a gradual increase of channel 1 (layer 1) weight from 0.39 mS to 0.48 mS in 5 steps, using the tuning algorithm. The evolution of the output waveform is shown in Fig. 5c(i), which confirms the correct operation. The amplitude (peak-peak) of the output current for the layer 1 component changes from 72 μA to 95 μA as a result (see Supplementary Figure 14b). In the next step, the state of the device in layer 1 is kept unchanged while the device in layer 2 is gradually turned on (Fig. 5c(v),(vi)). The corresponding output waveforms are displayed in Fig. 5c(iv). As expected, as the weight for channel 2 increases the amplitude of the high-frequency component in the output waveform increases. As shown in Supplementary Figure 14f, the layer 2 component (peak-peak) changes from 10–50 μA. The margins for the change in output current components for both the layer 1 and layer 2 devices are ~5 μA or more (Supplementary Figure 14) and it is to be noted that this margin can be easily adjusted by the tuning procedure. These results indicate that the devices in each layer in the 3D CMOL crossbars can be controlled independently and used for matrix multiplication operation. However, practical implementation of high bandwidth multiply-add operation using 3D CMOS/memristor crossbars will also require overcoming the challenges due to finite line resistance, sneak-path and other sources of noise. Increase in the number of layers in a 3D crossbar is equivalent to increasing the size of a 2D crossbar array, thereby also increasing the sneak-currents[32]. Several remedies proposed to mitigate the effect of sneak currents include use of mapping algorithms (to map target weight matrix onto actual crossbar conductance values)[33] or train the hardware through supervised or unsupervised learning schemes[14,34]. A detailed analysis of the effect of sneak-paths in 3D hybrid memristor/CMOS and accurate benchmarking is beyond the scope of this work.

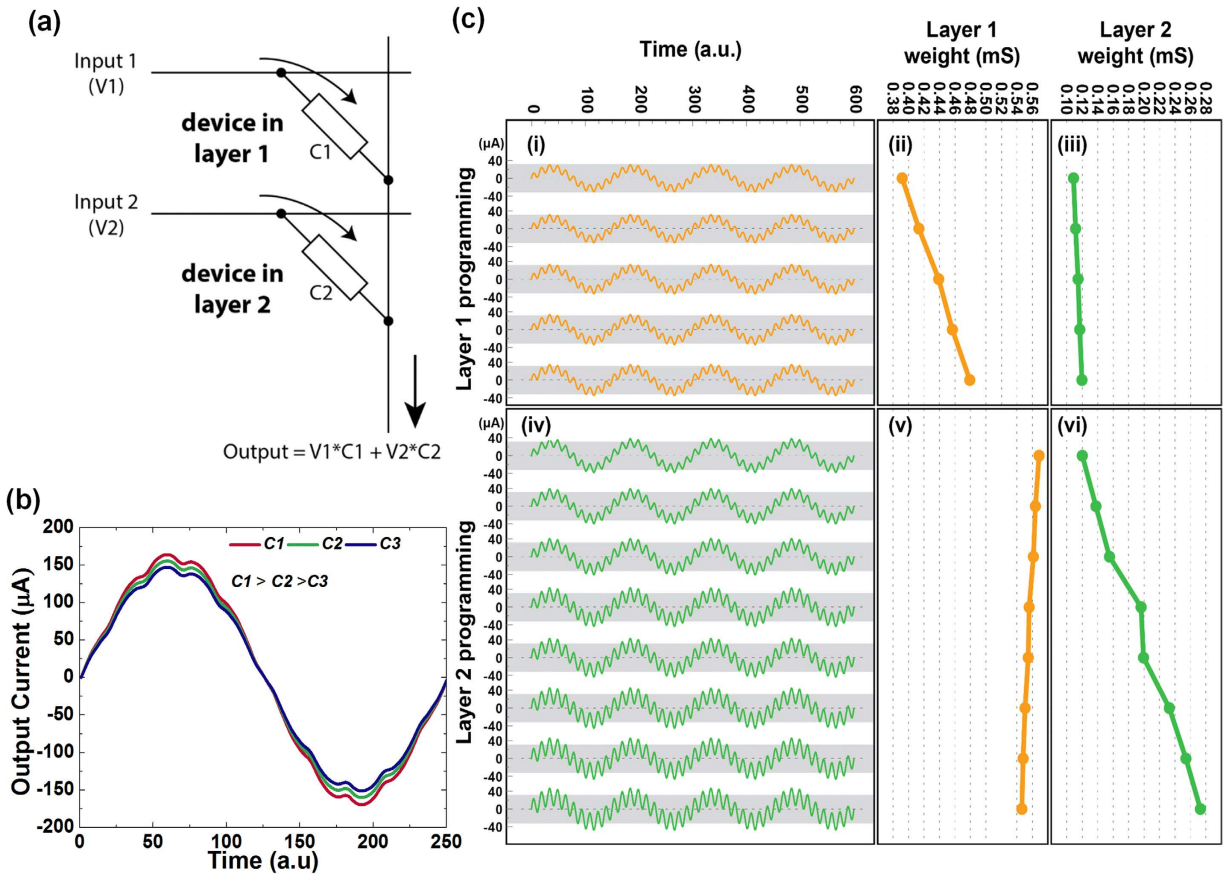


Figure 5. Dot-product operation in the integrated 3D memristive crossbars. (**a**) Schematic of the set-up for the dot-product operation utilizing two devices in two different layers of the 3D crossbar, (**b**) An example of the dot product operation with two sinusoidal inputs applied to two devices in a 3D crossbar and the device in layer 1 being programmed to decreasing conductance values, (**c**) (i) evolution of the output waveform with the weight of the device in layer 1 changing (ii) and the weight for the device on the 2nd layer being kept constant (iii); (iv) evolution of the output waveform when weight of the device in layer 1 is kept constant (v) while layer 2 device changes it's state (weight) (vi).

Development of a selector technology can also be a possible solution. We will investigate these avenues in our future work.

In summary, we demonstrate the first 3D CMOL hybrid circuit with 3D memristive crossbars monolithically integrated on a CMOS substrate. High integration yield in terms of good electrical contact between the memristive components and the CMOS substrate was achieved by planarization of the CMOS chips. The integrated 3D crossbars can be fully controlled by the underlying CMOS circuitry. The memristive devices display forming-free switching with low voltage operation. They are analog tunable using a write-and-verify algorithm. The multi-level control of the states for the memristive devices allows them to be used in multiply-add operations where their conductance values can be used as controllable weights. Demonstration of multiply-add operation utilizing memristive devices both in the 1st and 2nd layer of the 3D crossbars opens up promise for ultra-high bandwidth multiply-add engines with high density memristor/CMOS 3D hybrid circuits.

Methods

Preparation of planarization holder. The CMOS chips used in this experiment have a dimension of 5 mm × 5 mm and a thickness of 256 μm. Thickness of a 4 inch Si wafer with initial thickness of ~260 μm (+0−4 μm) is reduced by 6–10 μm (depending on the initial thickness) using deep-Si Reactive Ion Etching (DRIE) to have a final thickness of ~254 μm. The wafer is then polished by CMP in SF1 slurry (alkaline colloidal silica) for 4 mins to remove the roughness generated by the DRIE process. A 3 μm SiO_2 film is then deposited by Plasma Enhanced Chemical Vapor Deposition (PECVD) on the wafer. A 5.5 mm window is patterned on the oxide by photolithography with negative resist (AZnLOF2020) and using a 5.5 mm Si piece as the mask. The oxide in the window region is etched back with CHF3 plasma. The wafer is then subjected to DRIE to completely etch Si within the window to make a 5.5 mm × 5.5 mm hole in the Si wafer.

Chemical mechanical planarization of the chip. The as-received chip has a 1.3 μm scratch-protect oxynitride layer with unknown composition (undisclosed from the Foundry). Due to the unknown composition/quality of the oxynitride it is difficult to precisely control processing of this layer. Therefore we completely

remove this layer and use a planarization dielectric of known quality/composition. After completing removing the oxynitride layer by dry etching in CHF_3 plasma, the organic residues were removed by cleaning in AZ300T for 15 minutes. Next the CMOS pads in the active region are covered with Ti/Au (10/100 nm) to prevent oxidation of Al. A 2.5 μm SiO_2 layer is then deposited by Inductively Coupled Plasma based PECVD (ICP-PECVD) at low temperature (50 °C). The planarization holder is then crystal bonded on a second Si substrate and the CMOS chip is placed in the middle of the holder. The entire ensemble was then polished in CMP with SF1 slurry for 4 mins. After CMP the final topography of the chip is verified by atomic force microscopy (AFM). The oxide thickness on top of the CMOS pads post-CMP is measured by a reflectance measurement unit. The post-CMP oxide thickness is ~1.5 μm across the chip. Next the planarization oxide (SiO_2) is etched back to the desired thickness (180 nm).

3D memristive crossbar fabrication. $4 \times 4\,\mu m^2$ via holes for contact between the BEs of the crossbars and the CMOS pads are first created by photolithography and CHF_3 plasma etching through the 180 nm planarization oxide. Next, Ta/Pt (5/60 nm) BEs for layer 1 devices (width 2 μm) were patterned by photolithography and E-beam evaporation. The Al_2O_3/TiO_x (3/30 nm) switching stack is deposited by reactive sputtering in Ar/O_2 plasma. Stoichiometry of the TiO_x layer was controlled by controlling the O_2 flow. Next, via holes ($4 \times 4\,\mu m^2$) for contact between the TEs and CMOS pads are created by photolithography and dry etching in CHF_3 plasma. TEs of Ti/Pt (15/60 nm) are defined by optical lithography and E-beam evaporation. After fabrication of the first crossbar layer an isolation oxide of 200 nm is deposited by ICP- PECVD. The 2nd layer of crossbars is fabricated by performing the same fabrication steps used for layer 1, namely patterning of BE via holes, deposition of Ta/Pt BEs, deposition of the switching oxide stack, patterning of TE via holes and defining the Ti/Pt TEs. In a final lithography step, via holes are opened on the wire-bonding pads by photolithography and CHF_3 plasma etching. The chip is then annealed at 300 °C for 15 mins in forming gas ($N_2 + H_2$). The processed chip was wire bonded and packaged in a commercial facility before electrical measurements were performed.

References

1. Chen, Y. S. *et al.* Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity. *IEDM Tech. Dig. 5.5.* 1–4 http://dx.doi.org/10.1109/IEDM.2009.5424411 (2009).
2. Hsu, C. W. *et al.* 3D vertical TaOx/TiO2 RRAM with over 103 self-rectifying ratio and sub-μA operating current. *IEDM Tech. Dig. 10.4.* 1–4 http://dx.doi.org/10.1109/IEDM.2013.6724601 (2013).
3. Govoreanu, B. *et al.* 10 × 10 nm2 Hf/HfOx crossbar resistive RAM with excellent performance, reliability and low-energy operation. *IEDM Tech. Dig. 31.6.* 1–4 http://dx.doi.org/10.1109/IEDM.2011.6131652 (2011).
4. Lee, S. R. *et al.* Multi-level switching of triple-layered TaOx RRAM with excellent reliability for storage class memory. *Dig. Tech. pap. - VLSI Technol.* (*VLSIT*), *2012 Symp.* **52,** 71–72 (2012).
5. Sheu, S. S. *et al.* A 4 Mb embedded SLC resistive-RAM macro with 7.2 ns read-write random- access time and 160 ns MLC-access capability. *Dig. Tech. pap.- Int. Solid-State Circuits Conf.* (*ISSCC*) *11.2.* 200–202 http://dx.doi.org/10.1109/ISSCC.2011.5746281 (2011).
6. Sheu, S. S. *et al.* A 5 ns fast write multi-level non-volatile 1 K bits RRAM memory with advance write scheme. *Dig. Tech. pap.- VLSI Circuits, 2009 Symp.* 82–83 (2009).
7. Kim, Y. B. *et al.* Bi-layered RRAM with unlimited endurance and extremely uniform switching. *Dig. Tech. pap.- VLSI Technol.* (*VLSIT*), *2011 Symp*. 52–53 (2011).
8. Laiho, M. & Lehtonen, E. Arithmetic operation within memristor based analog memory. *Proc. Int. Workshop CNNA* 1–4 http://dx.doi.org/10.1109/CNNA.2010.5430319 (2010).
9. Merrikh-Bayat, F. & Shouraki, S. B. Memristor-based circuits for performing basic arithmetic operations. *Procedia Comp. Sci.* **3,** 128–132 (2011).
10. Shin, S., Kim, K. & Kang, S. M. Memristor Applications for Programmable Analog ICs. *IEEE Trans. Nanotech.* **10,** 266–274 (2011).
11. Gaba, S., Sheridan, P., Zhou, J., Choi, S. & Lu, W. Stochastic memristive devices for computing and neuromorphic applications. *Nanoscale* **5,** 5872–5878 (2013).
12. Jo, S. H. *et al.* Nanoscale Memristor Device as Synapse in Neuromorphic Systems. *Nano lett.* **10,** 1297–1301 (2010).
13. Kim, K.-H. *et al.* A Functional Hybrid Memristor Crossbar-Array/CMOS System for Data Storage and Neuromorphic Applications. *Nano lett.* **12,** 389–395 (2012).
14. Prezioso, M. *et al.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521,** 61–64 (2015).
15. Yu, S., Wu, Y., Jeyasingh, R., Kuzum, D. & Wong, H. S. P. An Electronic Synapse Device Based on Metal Oxide Resistive Switching Memory for Neuromorphic Computation. *IEEE Trans. Electron Dev.* **58,** 2729–2737 (2011).
16. Xia, Q. *et al.* Memristor−CMOS Hybrid Integrated Circuits for Reconfigurable Logic. *Nano lett.* **9,** 3640–3645 (2009).
17. Yang, J. J., Borghetti, J., Murphy, D., Stewart, D. R. & Williams, R. S. A Family of Electronically Reconfigurable Nanodevices. *Adv. Mat.* **21,** 3754–3758 (2009).
18. Gao, Y., Ranasinghe, D. C., Al-Sarawi, S. F., Kavehei, O. & Abbott, D. Memristive crypto primitive for building highly secure physical unclonable functions. *Sci. Rep.* **5,** 12785 (2015).
19. Gao, L., Alibart, F. & Strukov, D. B. Programmable CMOS/Memristor Threshold Logic. *IEEE Trans. Nanotech.* **12,** 115–119 (2013).
20. Likharev, K. K. & Strukov, D. B. CMOL: Devices, Circuits, and Architectures. *Lect. Notes Phys.* **680**, 447–477 (Springer, 2005).
21. Likharev, K. K. Neuromorphic CMOL circuits. *Proc. IEEE-NANO.* 339–342 http://dx.doi.org/10. 1109/NANO. 2003.1231787 (2003).
22. Strukov, D. B. & Likharev, K. K. Prospects for terabit-scale nanoelectronic memories. *Nanotech.* **16,** 137–148 (2005).
23. Liu, T. Y. *et al.* A 130.7 mm 2-layer 32-Gb ReRAM memory device in 24-nm technology. *IEEE J. Solid-State Circuits* **49,** 140–153 (2014).
24. Li, H. *et al.* Four-layer 3D vertical RRAM integrated with FinFET as a versatile computing unit for brain-inspired cognitive information processing. *Dig. Tech. pap.- VLSI Technol.* (*VLSIT*), *2016 Symp*. 1–2 (2016).
25. Lin, P., Pi, S. & Xia, Q. 3D integration of planar crossbar memristive devices with CMOS substrate. *Nanotech*. **25,** 405202 (2014).
26. Adam, G. C. *et al.* Highly-uniform multi-layer ReRAM crossbar circuits. *Proc. ESSDERC* 436–439 (2016).
27. Strukov, D. B. & Williams, R. S. Four-dimensional address topology for circuits with stacked multilayer crossbar arrays. *Proc. Nat. Academy of Sci.* **106,** 20155–20158 (2009).
28. Lee, H. D., Miller, M. H. & Bifano, T. G. Planarization of a CMOS die for an integrated metal MEMS. *SPIE Proc.* **4979,** 137–144 (2003).
29. Lastras-Montaño, M. A., Ghofrani, A. & Cheng, K.-T. Architecting energy efficient crossbar-based memristive random-access memories. *Int. Symp. Nanoscale Archit.* (*NANOARCH*) 1–6 http://dx.doi.org/10.1109/NANOARCH.2015.7180575 (2015).

30. Payvand, M. *et al.* A configurable CMOS memory platform for 3D-integrated memristors. *Int. Symp. Circuits and Syst.* (*ISCAS*) 1378–1381 http://dx.doi.org/10.1109/ISCAS.2015.7168899 (2015).
31. Yang, J. J., Strukov, D. B. & Stewart, D. R. Memristive devices for computing. *Nat. Nanotech.* **8,** 13–24 (2013).
32. Xia, L. *et al.* Technological Exploration of RRAM Crossbar Array for Matrix-Vector Multiplication. *J. Comp. Sc. Tech.* **31,** 3–19 (2016).
33. Hu, M. *et al.* Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication. *Proc. Design Automation Conf.* (*DAC*) 1–6 http://dx.doi.org/10.1145/2897937.2898010 (2016).
34. Liu, B. *et al.* Vortex: variation-aware training for memristor x-bar. *Proc. Design Automation Conf.* (*DAC*) 1–6 http://dx.doi.org/10.1145/2744769.2744930 (2015).

Acknowledgements

This work was supported by the Air Force Office of Scientific Research (AFOSR) under the MURI grant FA9550-12-1-0038 and DARPA under Contract No. HR0011-13-C-0051UPSIDE via BAE Systems.

Author Contributions

B.C. wrote the manuscript and fabricated the hybrid CMOS/3D memristor chip. M.A.L.-M. designed the CMOS chip and created the user-interface for electrical measurements. Both B.C and M.A.L.-M. conducted the electrical characterizations and analyzed the data. G.A. contributed in the chemical mechanical planarization of the CMOS chip as well as the tuning operation of the memristors. M.P. contributed to develop strategies for electrical characterization of the memristor crossbars. B.H. was involved in the non-stoichiometric TiO_{2-x} thin film depositions. K.T.C. and D.B.S. have supervised the overall project. All authors have seen and approved of the manuscript before submission.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Chakrabarti, B. *et al.* A multiply-add engine with monolithically integrated 3D memristor crossbar/CMOS hybrid circuit. *Sci. Rep.* **7**, 42429; doi: 10.1038/srep42429 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A multiply-add engine with monolithically integrated 3D memristor crossbar/CMOS hybrid circuit

B Chakrabarti[1*], M A Lastras-Montaño[1], G Adam[1], M Prezioso[1], B Hoskins[2], M Payvand[1], A Madhavan[1], A Ghofrani[1], L Theogarajan[1], K-T Cheng[1,3] and D B Strukov[1]

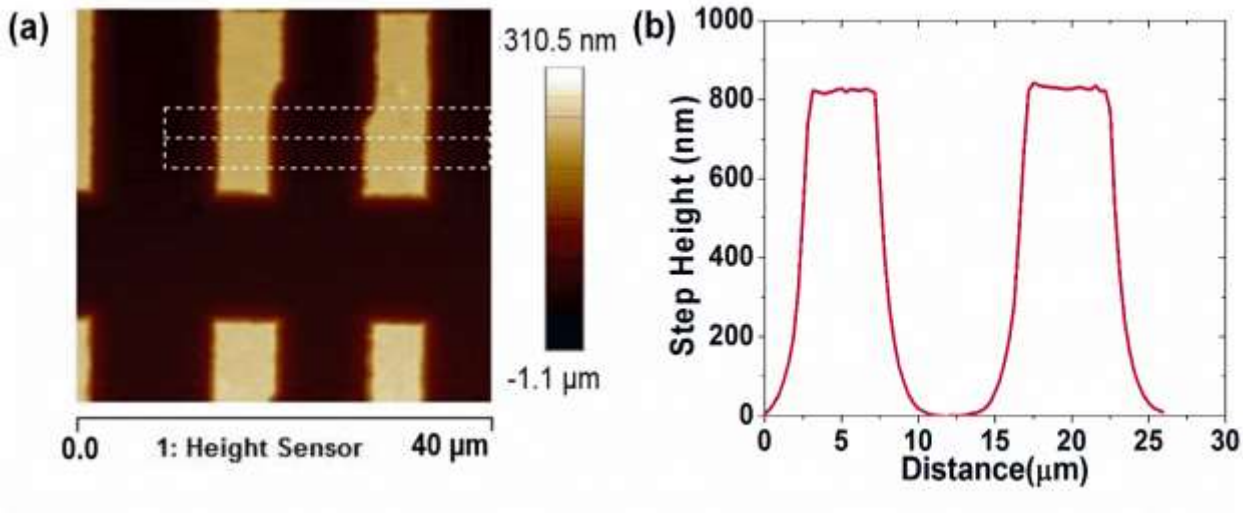
[1] Electrical and Computer Engineering Department, University of California, Santa Barbara, CA, 93106.

[2] Materials Department, University of California, Santa Barbara, CA, 93107.

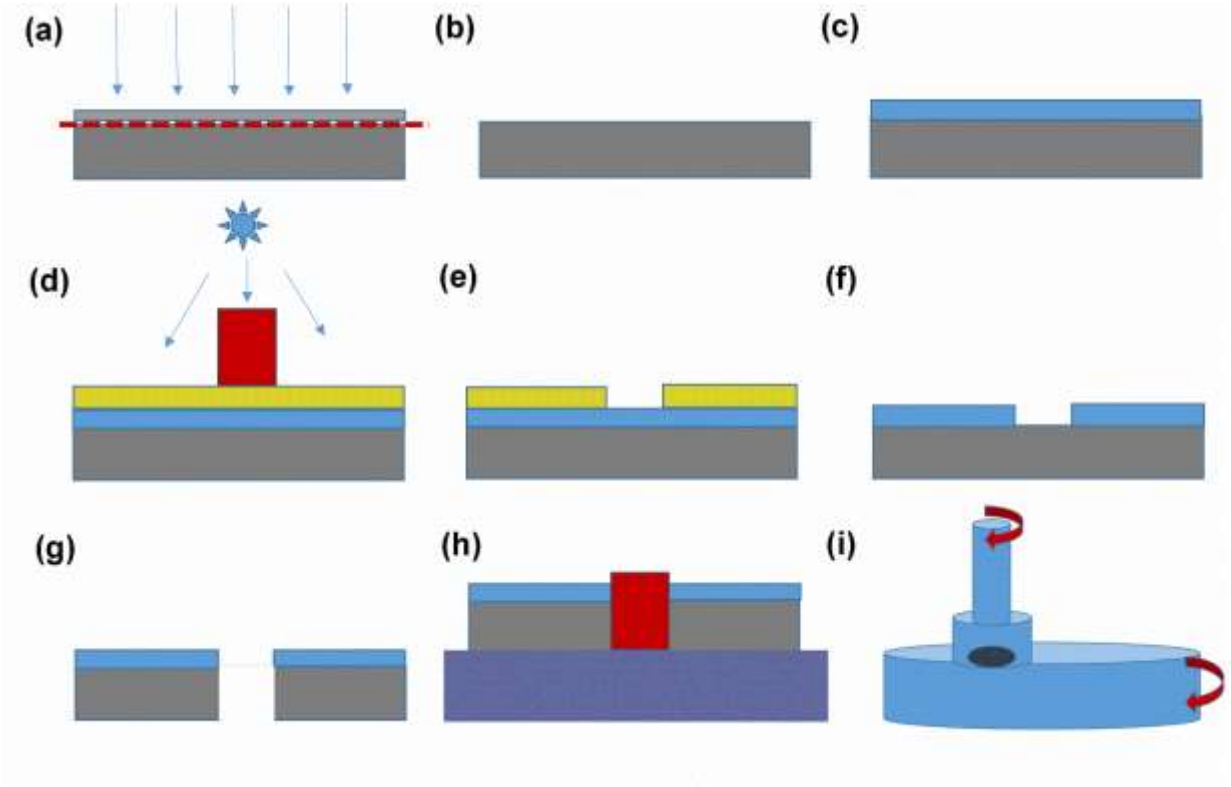
[3]School of Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

E-mail: bchakrabarti@ece.ucsb.edu

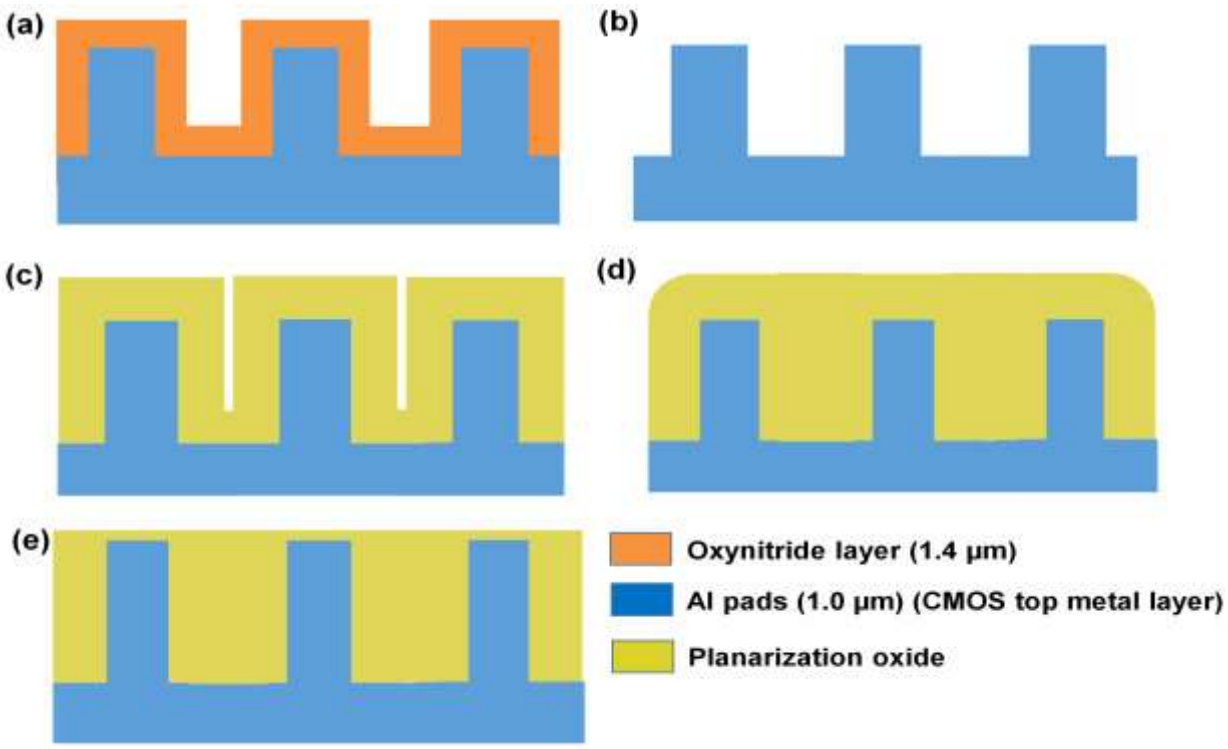
Supplementary Information



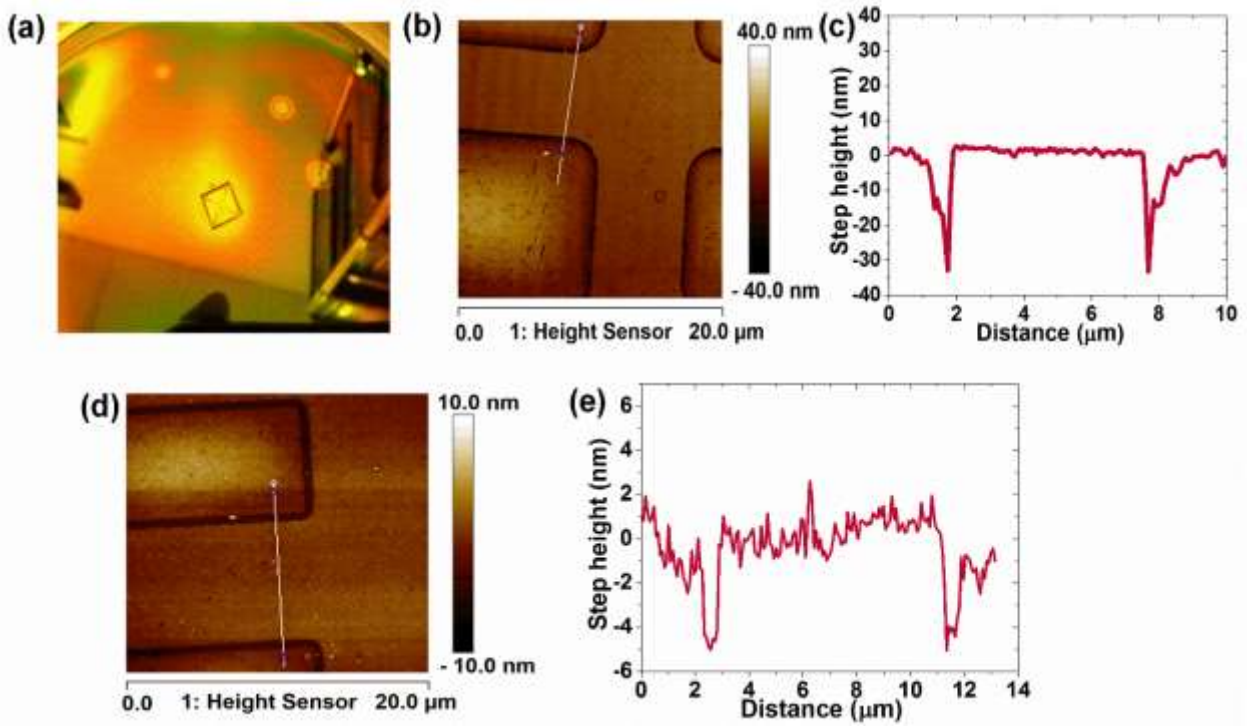
Supplementary figure 1. Initial topography (before planarization) of the as-received CMOS chip. (a) AFM image of a section of the CMOS chip before planarization showing the CMOS pads covered under SiO_2, **(b)** the topography of the region shown in (a).



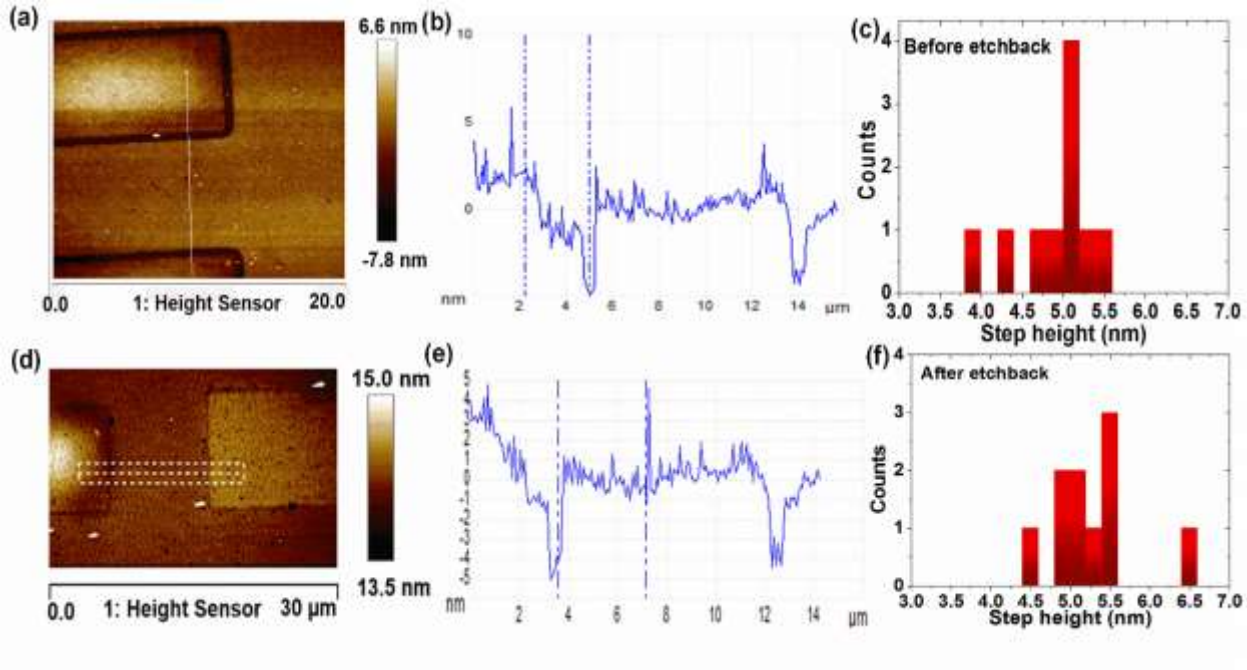
Supplementary figure 2. Preparation of the CMP holder followed by the planarization process. (a) initial deep-Si Reactive Ion Etching (DRIE) and planarization by CMP on the as received wafer (4 inch) to reduce the wafer thickness to desired value, **(b)** holder wafer after thickness reduction by CMP, **(c)** 3 µm oxide deposition by PECVD, **(d)** photolithography to define the etch window through the oxide using a dummy chip (6 mm), **(e)** profile of the resist after lithography and development, **(f)** dry etching and removal of oxide from the etch window, **(g)** etch through the Silicon wafer at the etch window by deep-Silicon Reactive Ion Etching (DRIE) process to make a cavity in the Silicon wafer, **(h)** a chip is placed inside the cavity of the holder and the whole assembly is bonded onto another 4 inch Silicon wafer by crystal bond: note that the height of the chip is slightly higher than the cavity (~ 3-4 µm) to ensure that during the CMP process the chip will be planarized, **(i)** the chip-holder-carrier assembly is planarized in the CMP tool.



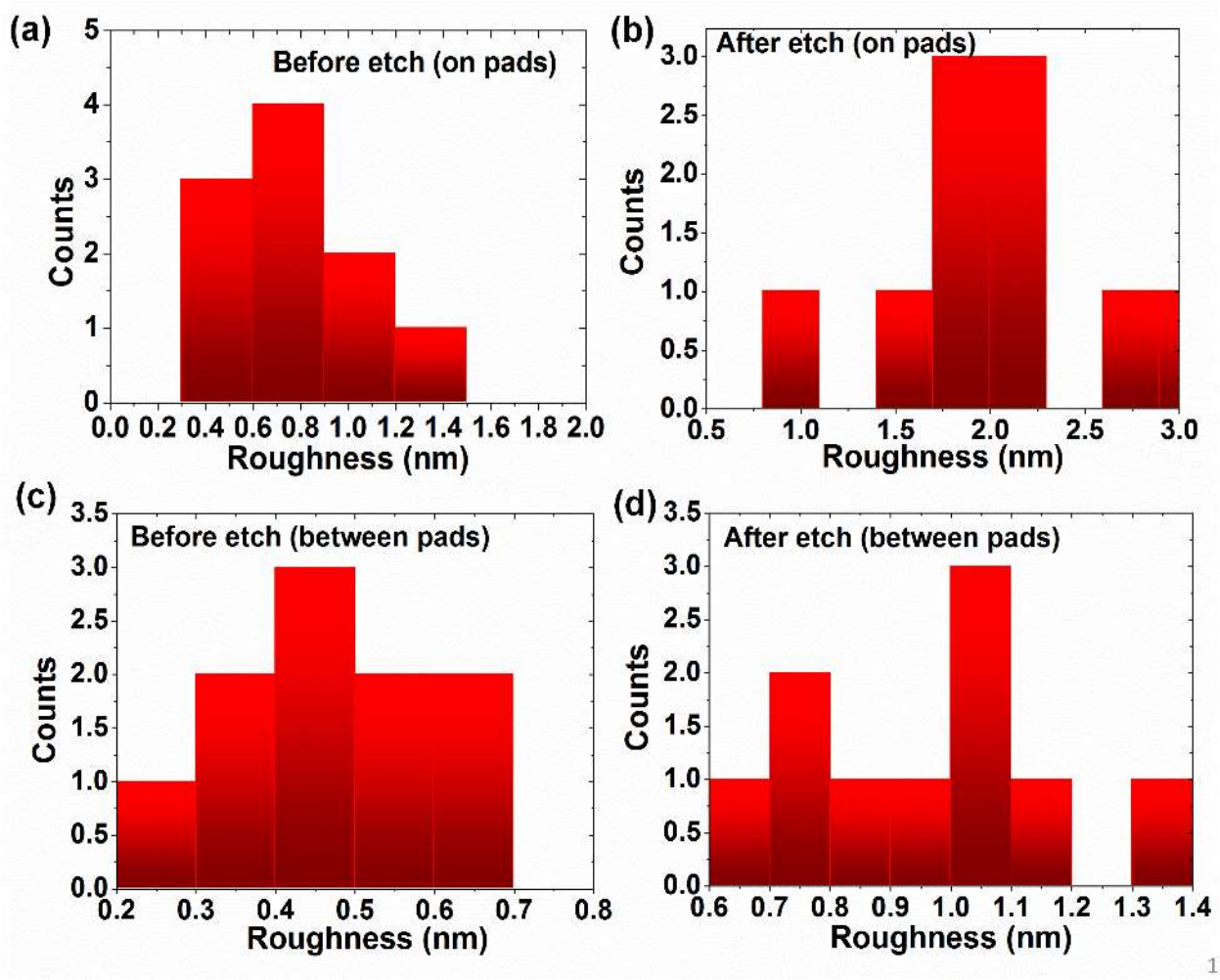
Supplementary figure 3. Planarization and etch back procedure. (a) The as-received chip with the scratch protect oxynitride covering the top metal (Al) of the CMOS layer, **(b)** removal of the oxynitride layer by plasma etching, **(c)** deposition of silicon dioxide by low temperature plasma enhanced chemical vapor deposition (PECVD): thickness of the oxide deposited is twice that of the thickness of the Al pads of the CMOS layer, **(d)** planarization of the oxide by chemical mechanical planarization (CMP) process, **(e)** etch-back of the planarization oxide after CMP by plasma etching to a desired thickness (~180 nm).



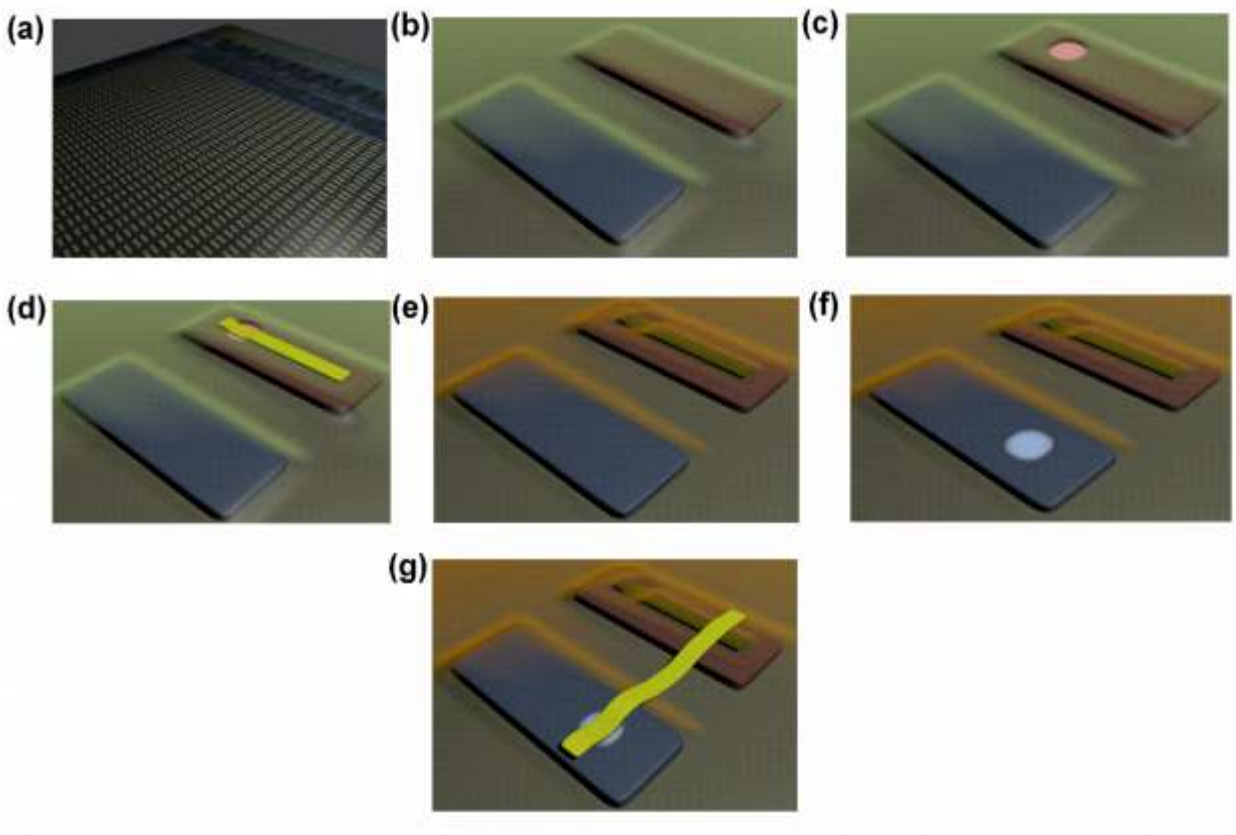
Supplementary figure 4. Surface topography after planarization. (a) Optical image of a CMOS chip inside the cavity of a Silicon carrier wafer with both the carrier and the chip crystal-bonded on to another Silicon holder wafer, **(b)** AFM image of a section of the CMOS chip after the first planarization step (fast polish) showing the CMOS pads covered under SiO_2. **(c)** The topography of the region shown in (b) indicates step-height reduction down to ~ 30-40 nm. **(d)** AFM image of a region after the final planarization (slow polish). **(e)** Topography of the region shows step height less than 10 nm.



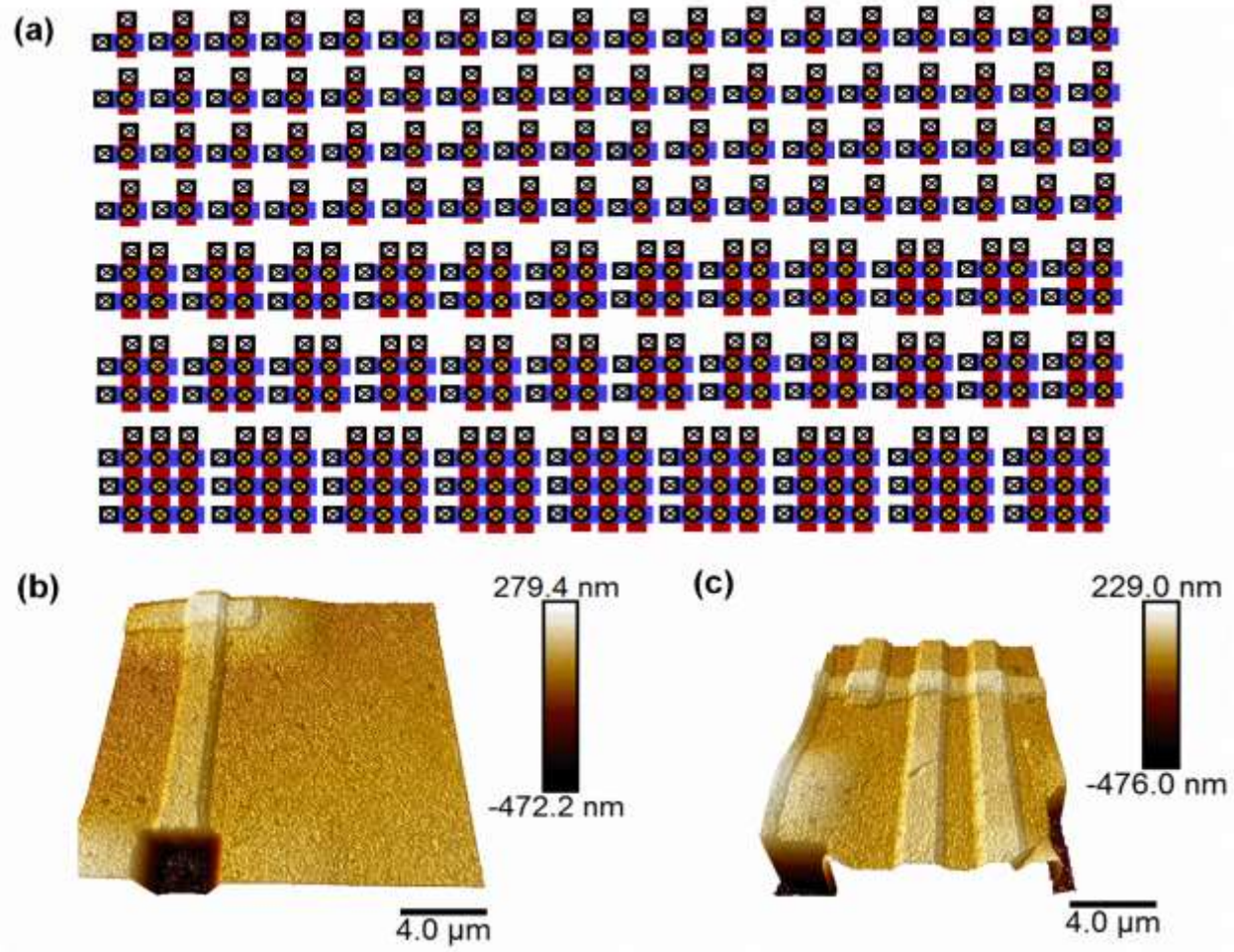
Supplementary figure 5. Step height measurements before and after the etch-back procedure. (a) AFM image of CMOS pads under planarization oxide before the etchback process, **(b)** surface topography of the region shown in (a), (c) histogram showing distribution of surface roughness in the region, **(d)-(f)** AFM image, surface topography and histogram of step height on the planarization oxide after the etchback process.



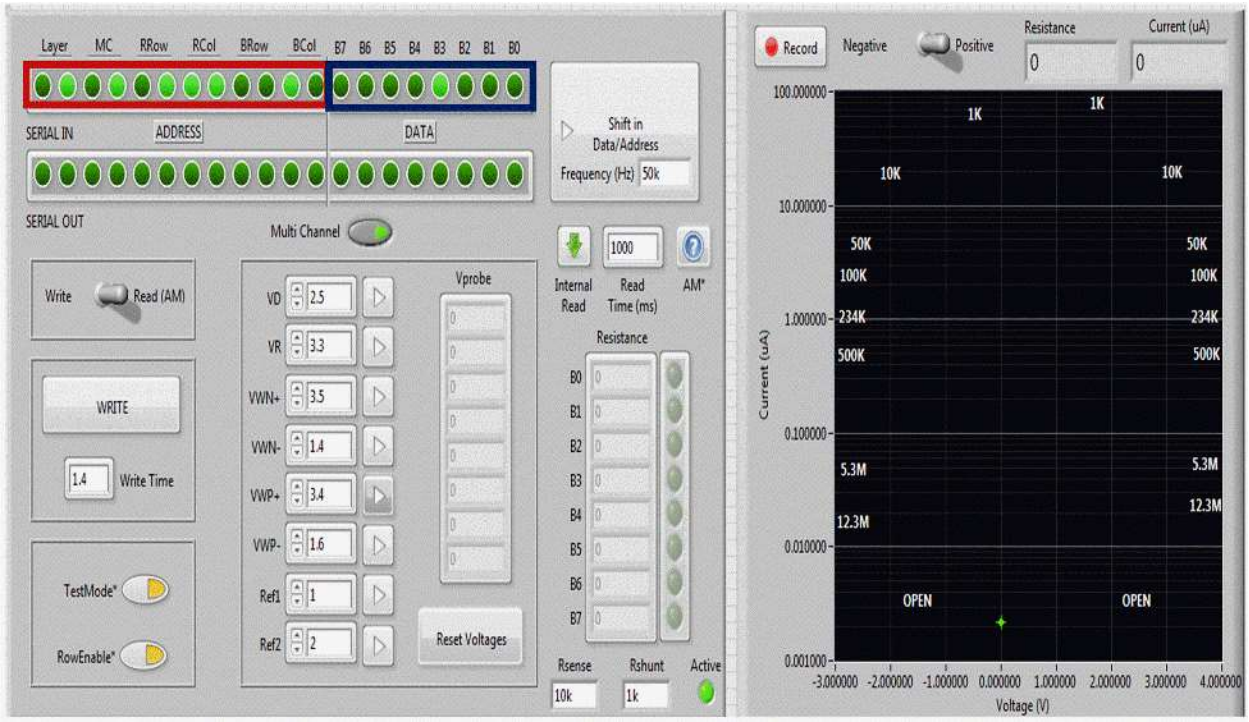
Supplementary figure 6. Comparison of the oxide roughness before and after planarization. **(a)** and **(b)** show the histograms of roughness of oxide measured on top the device pads ('red' and 'blue' pins) before and after the etch-back process respectively. **(c)** and **(d)** show the histograms of the oxide roughness measured between the device pads before and after the etch-back process respectively. In all the cases a slight increase is observed in the oxide roughness after the etch-back process.



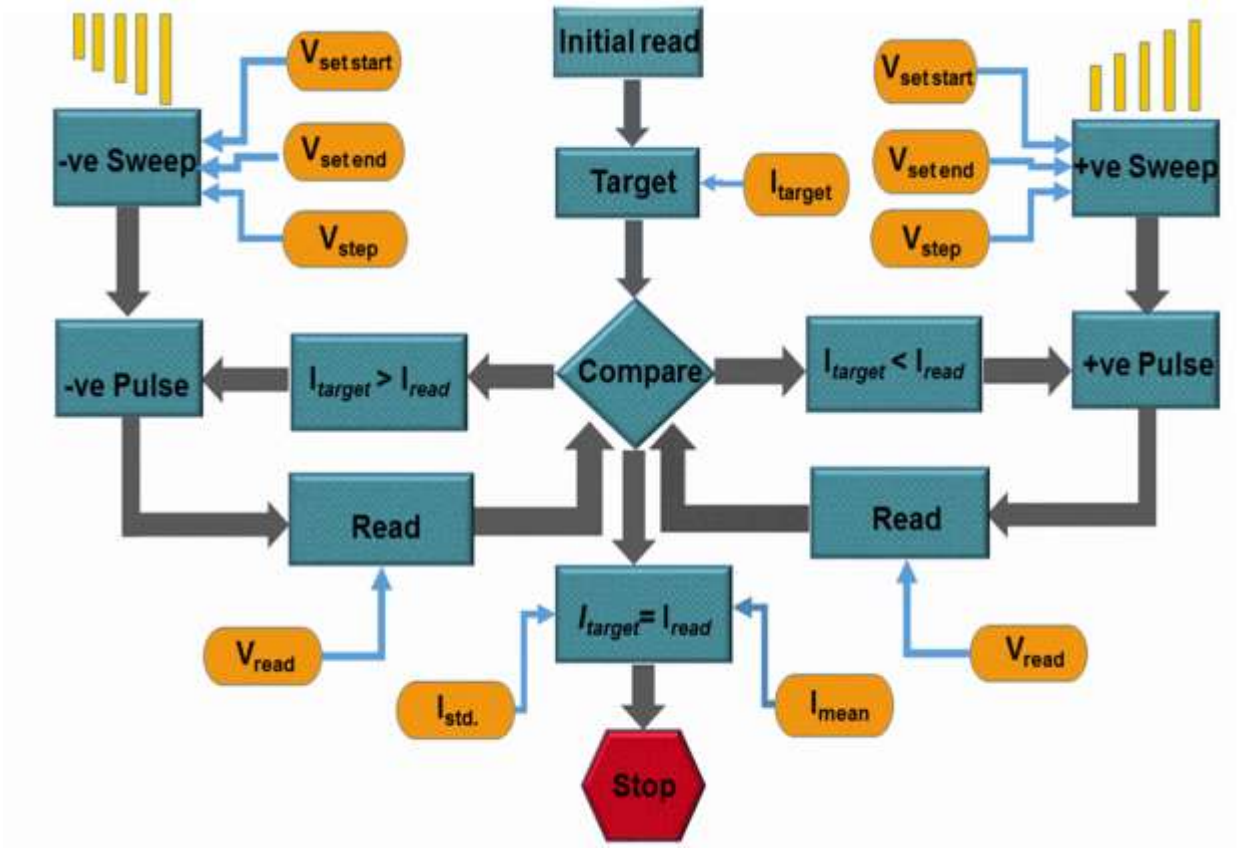
Supplementary figure 7. Device fabrication steps. (a) 3D image of the CMOS chip covered under the planarized and etched back oxide layer , **(b)** a pair of the 'Red' and 'Blue' pins corresponding to one 'CMOS cell' covered under the planarization oxide, **(c)** via opening on the 'red' pin by photolithography and dry etching for creating bottom electrode contact, **(d)** bottom electrode (Ta/Pt) deposition by photolithography and e-beam evaporation, **(e)** deposition of the switching dielectric (Al_2O_3/TiO_{2-x}) by reactive sputtering, **(f)** via opening on the 'blue' pin for creating top electrode contact, **(g)** deposition of top electrode (Ti/Pt) by photolithography and e-beam evaporation.



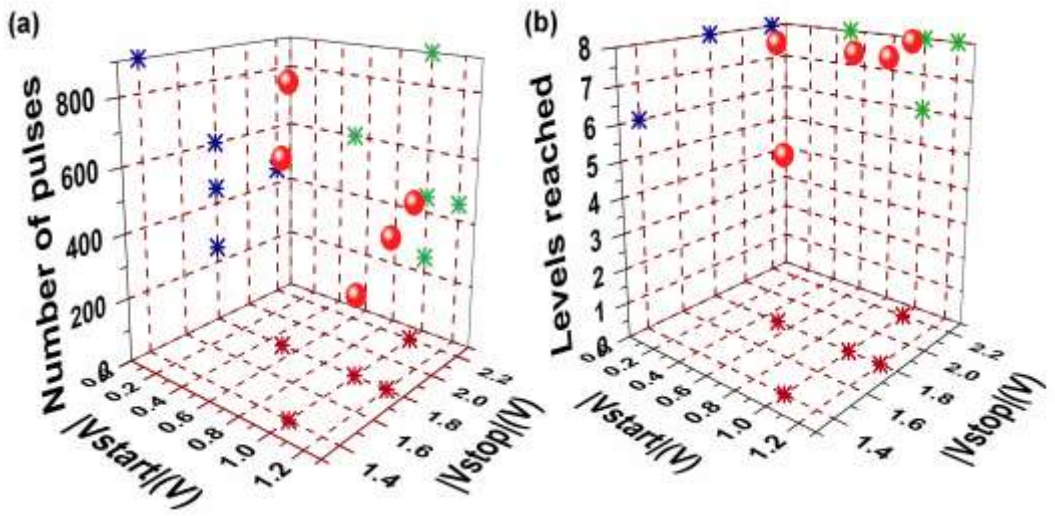
Supplementary figure 8. Details of the 2D single devices and crossbars. (a) Layout of the single devices as well as 2x2 and 3x3 crossbars (2 dimensional) integrated on the CMOS chip, **(b)** AFM image of a typical single cross-point device integrated, **(c)** section of a 2D crossbar integrated on the CMOS chip.



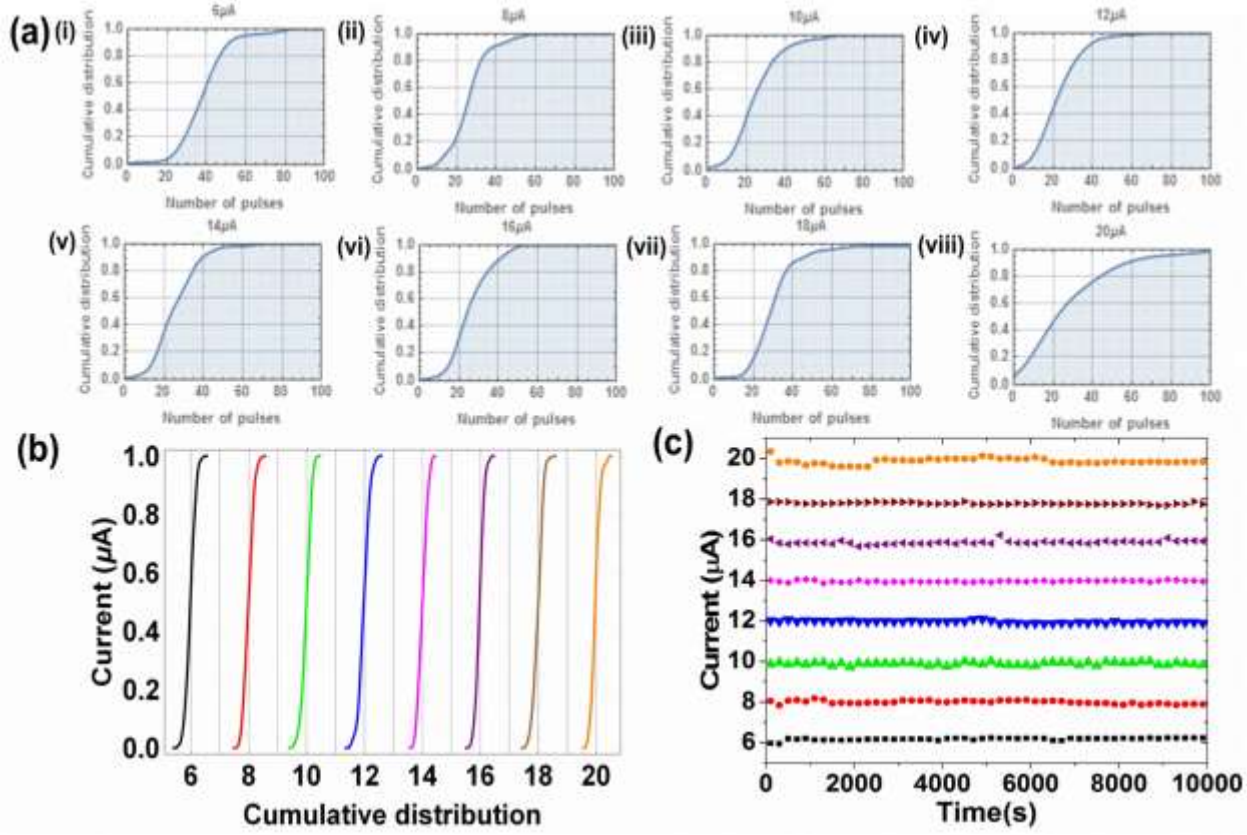
Supplementary figure 9. An image showing the interface to address each device integrated on a CMOS chip. The 12 bit address shown within the red box can select a cross-point within one multi-cell in the CMOS cell array. The 8 bits shown in the blue box can select 8 multi-cell columns concurrently.



Supplementary figure 10. A flowchart showing the tuning algorithm to tune the integrated memristive devices. Orange boxes denote the user-input variables. At each point of the tuning the read current Iread is compared with the target current Itarget. The tuning operation stops at any point if Iread = Itarget. The accuracy of the match is governed by Imean and Istd. The tuning stops if (Target current – mean read current) ≤ Imean and the standard deviation of the read current matches Istd.

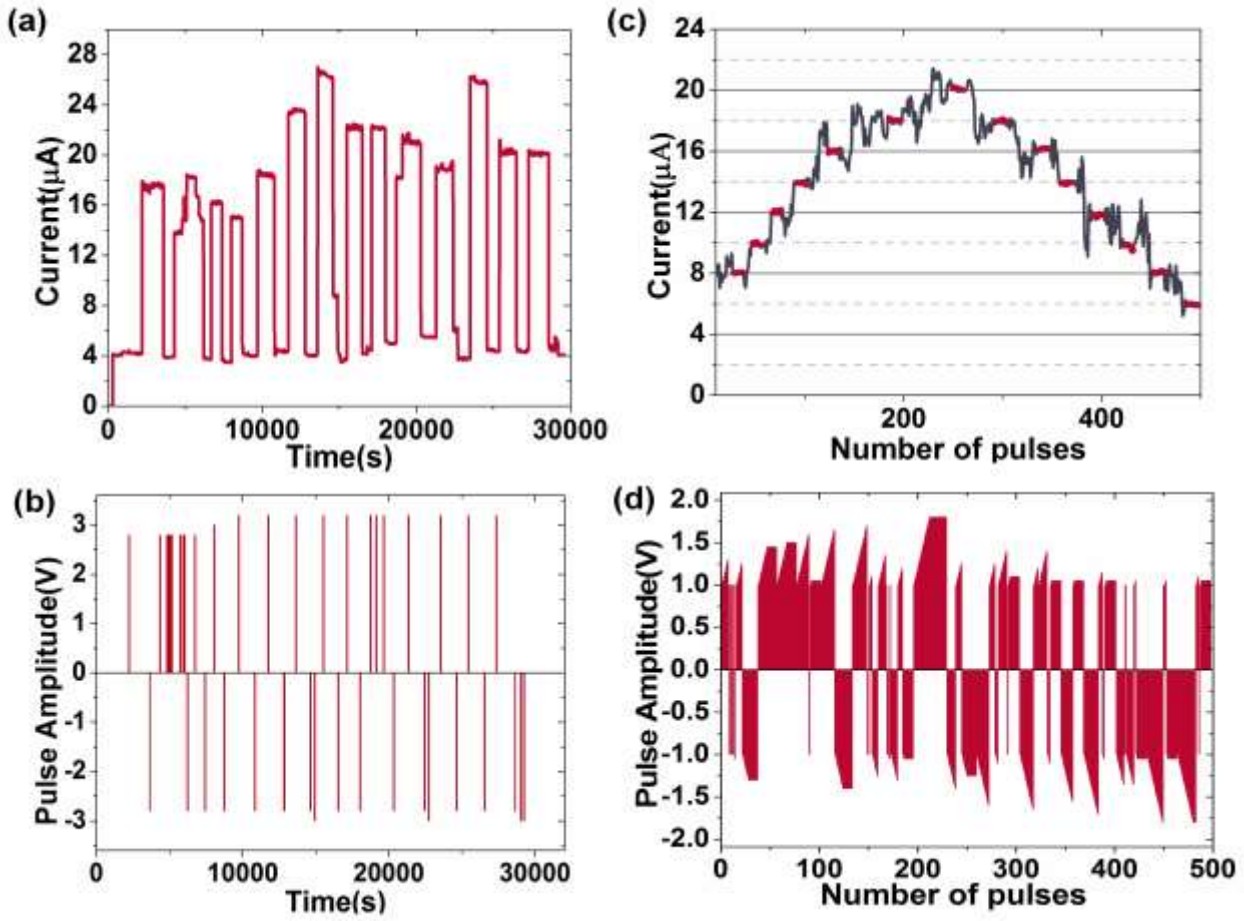


Supplementary figure 11. (a) optimization of pulse numbers as a function of Vstart and Vstop, **(b)** Number of levels reached for different Vstart and Vstop values.

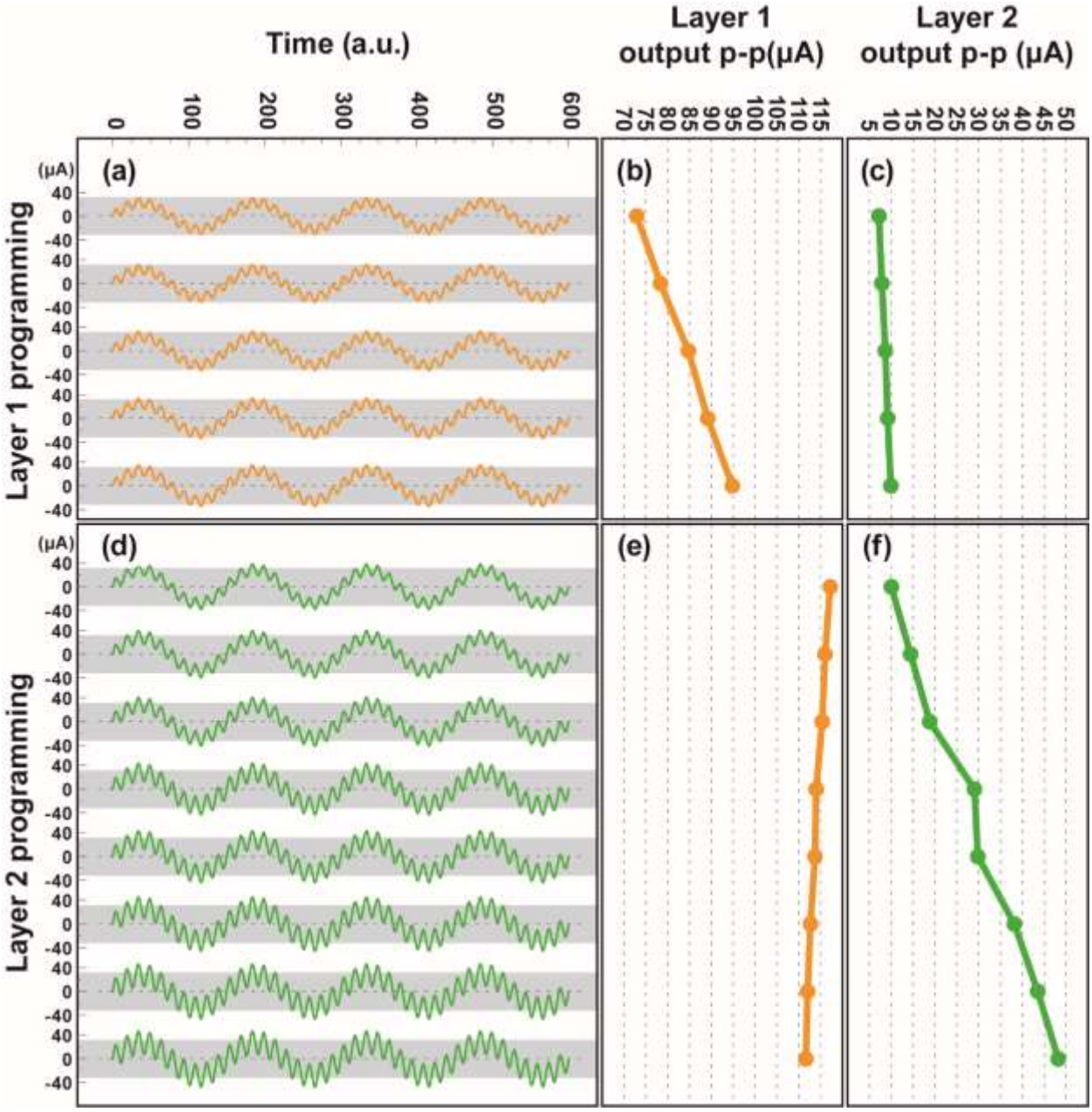


Supplementary figure 12. Performance of the optimized tuning algorithm for 1000 switching events randomly occurring between the 8 designated levels. (a) (i)-(a) (viii) depict the

cumulative distribution of the number of pulses required to tune to each of the eight levels, **(b)** cumulative distributions of the tuned current values for each level, **(c)** room temperature retention of each level measured over 10000 seconds.



Supplementary figure 13. (a) Pulsed switching of a device between fully on and off states. **(b)** The applied write pulses corresponding to the switching shown in (a). **(c)** Example of the tuning operation using the train of positive and negative pulses shown in **(d)**. Note that the amplitudes of the pulses applied for the tuning procedure are significantly less compared to the pulse amplitude required for turning the device fully on or off.



Supplementary figure 14. (a) Evolution of the output waveform as the conductance of the device in layer 1 is increased while the conductance of layer 2 device is kept unchanged. **(b)** Evolution of the component of the output current in layer 1 device as its conductance increases. **(c)** Component of the output current in layer 2 device remains unchanged. **(d)** Evolution of the output waveform as the conductance of the device in layer 2 is increased while the conductance of the layer 1 device is not changed. **(e)** Component of the output waveform in layer 1 corresponding to the experiment shown in (d). **(f)** Evolution of the component of output current in layer 2 as its conductance increases.

Supplementary Note 1. Tuning procedure

In each tuning step the read-current (Iread) measured at read-voltage Vread is compared against the target current (Itarget) corresponding to the desired state of the device. Depending on whether Iread is less or greater than the Itarget, a set of positive or negative pulses (1 μs) are applied in a ramp. Each write pulse is followed by a read operation (with a 1 ms read pulse) that evaluates the Iread. The tuning operation stops at any point if Iread matches Itarget. The precision of the tuning operation is governed by the user-defined inputs Imean and Istd. In case of an overshoot, the tuning operation can go in the opposite direction. All the user-defined inputs can be used to optimize the speed of the tuning operation.