ADVANCED MATERIALS

Supporting Information

for *Adv. Mater.*, DOI: 10.1002/adma.202003984

Filament-Free Bulk Resistive Memory Enables Deterministic Analogue Switching

Yiyang Li, Elliot J. Fuller, Joshua D. Sugar, Sangmin Yoo, David S. Ashby, Christopher H. Bennett, Robert D. Horton, Michael S. Bartsch, Matthew J. Marinella, Wei D. Lu, and A. Alec Talin**

WILEY-VCH

Supporting Information

Filament-Free Bulk Resistive Memory Enables Deterministic Analogue Switching

Yiyang Li[*1], Elliot J. Fuller[1], Joshua D. Sugar[1], Sangmin Yoo[2], David S. Ashby[1], Christopher H. Bennett[3], Robert D. Horton[1], Michael S. Bartsch[1], Matthew J. Marinella[3], Wei D. Lu[2], A. Alec Talin[*1]
[1]Sandia National Laboratories, Livermore, CA, USA
[2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA
[3]Sandia National Laboratories, Albuquerque, NM, USA

*Corresponding authors: yiyangli@umich.edu; aatalin@sandia.gov

Fabrication of bulk-RRAM devices on single-crystal YSZ:

Single-crystal YSZ substrates, 1-cm square and 100, 300, and 1000-μm thick, were purchased from MTI (www.mtixtl.com) and used without further processing. These substrates have a nominal concentration of 8 mol% Y_2O_3 in ZrO_2. Reactive DC sputtering using a Kurt J. Lesker CMS18 system was used to deposit TiO_2 from a 76-mm-diameter Ti metal target (99.99%) using a sputter power of 300 W in a 4 mtorr gas environment consisting of 33% O_2 and 67% Ar. Sputtering was conducted on both sides on the substrate: the TiO_2 film on the switching layer side is 60-nm thick; the TiO_2 film on the base layer side is 120-nm thick. The sputter rate was about 0.5 nm min^{-1}. The film was then annealed at 600ºC in air for 1 hr to crystallize the anatase phase. A 5-nm Ti adhesion layer and 50-nm Pt contacts were deposited on the 60-nm-thick switching layer by electron beam evaporation at a rate of ~1 Å s^{-1}. A stainless-steel shadow mask was used to pattern two contacts separated by 250 μm on the switching layer to measure the conductance. The area of the switching layer between the contacts is 8000-μm long and 250-μm wide.

Next, we reduced the base and switching layer from TiO_2 to TiO_{2-X} in a reducing environment containing 2 bar of pure H_2 at 400 ºC for 1 hr using a Setaram PCTPro gas sorption tool. The base layer TiO_2 side as well as the part of the switching layer not blocked by the contacts were reduced, which the switching layer under the Pt contacts was not reduced because Pt blocks oxygen from leaving TiO_2. After reduction, the switching layer conductance was measured to be ~100 μS using the previously-fabricated contacts. An unpatterned Ti/Pt contact was again formed by electron beam evaporation on the 120-nm-thick TiO_{2-X} base layer to complete the device.

Fabrication of bulk-RRAM devices with thin-film YSZ on Si/SiO_2 substrate:

We first fabricated the Pt contacts with a 2-μm gap using optical photolithography (Supporting Figure S2). Photoresist was spun on a 150-mm silicon wafer containing a 100-nm-thick dry thermal oxide layer. The photoresist was patterned using a Karl Suss contact aligner and developed with an MF-319 developer. Afterwards, 5 nm of Ti and 50 nm of Pt were evaporated, and the remaining metal was removed by lift-off in PG Remover to define the contacts. Next, 60 nm of TiO_2 was evaporated by reactive sputtering as described previously and annealed at 600 ºC for 1 hour to crystallize into anatase for the switching layer. 400 nm of YSZ was grown by RF sputtering on a 76-mm YSZ target (8 mol% Y_2O_3 in ZrO_2, Plasmaterials LLC) using a sputter power of 240 W and a growth rate of ~0.25 nm min^{-1}. The sample was

then annealed at 700 ^{o}C for 2 hours. X-ray diffraction and Raman spectroscopy of the YSZ and TiO_2 films are shown in Supporting Figure S3, showing pure phase for both materials.

180 nm of oxygen-deficient TiO_{2-X} layer was grown above the YSZ for the base layer. This oxygen-deficient film was grown by changing the sputter gas composition from 67% Ar and 33% O_2 to 87% Ar and 13% O_2; the deposition rate was ~10 nm min^{-1}. This oxygen-deficient film is electronically conductive, whereas the oxidized anatase film for the switching layer is not. Finally, a 30-nm TiO_X layer was sputtered above TiO_{2-X} by changing the sputter gas to 91% Ar and 9% O_2; the sputter rate was ~5 nm min^{-1}. These two final layers were not annealed to preserve the oxygen-deficient TiO_{2-X} layer. The parameters were chosen based on past experience using similar materials, and were not optimized for this purpose.

We note that these two methods yielded different conductivities because they differed in the amount of oxygen vacancies that were incorporated. The direct sputtering of nonstoichiometric TiO_{2-X} was done for the thin-film devices because the YSZ films delaminate in the presence of two bars of hydrogen. Reducing the annealing temperature to be compatible with a back-end-of-line process is an important avenue of future work.

Materials Characterization:

Scanning transmission electron microscopy: Cross-samples of the entire bulk-RRAM stack were prepared in a Dualbeam FIB/SEM (Helios 660 Nanolab, ThermoFisher). The conventional FIB liftout approach was used with 30 kV Ga ions and a final step at 5 kV to remove ion-induced damage. STEM imaging of these cross-sections was performed on a probe-corrected 300 kV Titan Themis Z (ThermoFisher) equipped with a 4-quandrant SuperX EDS detector system.

Conductive AFM: Conductive AFM measurements were performed on devices fabricated on 500-μm-thick single-crystal YSZ. The base layer and contacts were fabricated identically to the other devices. The switching layer differed because the c-AFM tip needs to contact a TiO_{2-X} region on Pt. For the switching layer, two sets of Pt contacts were first patterned and grown on YSZ; each set of contacts contains 10-μm-wide Pt strips separated by a 10-μm-wide Pt-free region. The distance between the two sets of contacts is 100 μm. The 60-nm-wide TiO_2 switching layer were sputtered next; half the film resides above the Pt contacts, while the other half grown on the Pt-free region resides above YSZ. We electrochemically set the device conductance using electrochemistry to a high-conductance state of 5 μS and a low-conductance state of 1 μS. The conductance measures the conductance of the 5-mm-long, 100-μm-wide TiO_{2-X} region between the two sets of contacts.

Conductive AFM measurements were carried out using standard contact mode AFM (Bruker Dimension Icon). The AFM cantilevers had a spring constant of 2.8 N m^{-1} and had doped, single-crystal diamond tips (Adama Probe SS). The nominal radius of curvature of the AFM tip was estimated to be <5 nm based on the manufacturer's specifications. The applied voltage was -2V.

X-ray diffraction: Polycrystalline thin-film X-ray diffraction was conducted using a Rigaku Smartlab X-ray diffractometer using a Cu K-α source. Samples were measured using a 2θ scan with an incident angle θ=1 degree. The samples consist of thin films of TiO_2 or YSZ grown to silicon and annealed at 600^{o}C for 1 hr for TiO_2 and 700^{o}C for 4 hr for YSZ.

Raman spectroscopy: Raman spectroscopy was conducted using a Renishaw invia confocal Raman microscope. A 532 nm green laser was used. The same samples as the Raman was used.

X-ray photoelectron spectroscopy: The XPS data was acquired on a Thermo-Fisher Scientific K-alpha+ using a monochromatic Al (1486eV) X-ray source. The ultimate vacuum for this tool is below 5×10^{-10} Torr. The data was processed using Thermo's Avantage software using relative sensitive factors, backgrounds, and curve fits appropriate for the samples. The pristine sample consists of a TiO2 thin-film sputtered on Si and annealed at 600^{o}C for 1 hr. The reduced film was reduced in a H_2 environment at 400^{o}C, and its electronic conductance was ~2 MΩ, approximately equal to that of the high-conductance state in Fig. 1f (450 nS).

Electrochemical cycling and retention:

Bulk-RAM was characterized electrochemically on a hot plate in an Ar glovebox. The temperature of the hot plate was calibrated using a thermocouple adhered to a reference YSZ substrate using silver paint. One probe was connected to each of three contacts; a Bio-logic SP-300 bipotentiostat was used to study the sample, whereby potentiostat channel 1 was used to apply voltage and current between contact 1 and 3, and potentiostat channel 2 was used to measure the electronic conductance of the switching layer by applying a constant 100-mV bias between contacts 2 and 3 and recording the current. We cycled the device using cyclic voltammetry in Supporting Figure S8a at 169 ^{o}C.

Retention measurements:

Long-term retention measurements in Figure 4 were made using a SP-300 bipotentiostat. The single-crystal samples were heated to 160 ^{o}C using a Nextron heated probe station and charged to 20 μS by applying a 1V on contact 1, and measuring the current between contact 2 (100 mV) and contact 3 (common). Afterwards, the device was quickly cooled (<30 s) to the desired measurement temperature. Once it was cooled, we shorted the base and switching layers by applying 0 V on contact 1 and measured the conductance of the switching layer using contact 2 and 3 over time. The thin-film devices were measured identically, except that potentiation was conducted at 90^{o}C instead of 160 ^{o}C, and the switching layer was charged to 500 nS.

Electrochemical impedance spectroscopy:

The ionic conductivity of the YSZ electrolyte were measured using electrochemical impedance spectroscopy (Bio-logic SP-300) on a hot plate in the glovebox. We applied alternating currents on opposite sides of single-crystal YSZ with Pt contacts from a frequency of 10^5 Hz to frequencies as low as 10^{-4} Hz at 50^{o}C with an amplitude of 100 mV. Representative impedance results are shown in Supporting Figure S10.

Data acquisition (SET/RESET) measurements:

The experiments in Figure 1f,g and Figure 3a-c, as well as all other experiments where pulses were less than 1 second, were conducted using a National Instruments Data Acquisition Device (DAQ-6358) controlled using the LabView program. An analogue switch (MAX327CPE) served as the extrinsic selector during weight updates. This selector was ON by applying 2V concurrent with the write pulses, and OFF at all other times to retain state during weight updates. This procedure was described in ref. [1]. Analogue outputs from the DAQ were connected to contacts 1 and 2 as well as V_{IN} of the analogue switch, while contact 3 was connected to a virtual ground created by the inverting input of an operational amplifier. An operational amplifier (AD820) was used to create an operational trans-conductance amplifier circuit using a 1 MΩ (for the single-crystal YSZ devices) or 100 MΩ (for the thin-film YSZ devices) resistor between the inverting input to the output, and the operational amplifier output was connected to an analogue input of the DAQ to measure the read current through the switching layer.

The write speed for the single-crystal devices was quantified by determining the write pulse length needed to obtain 100 weight updates between the highest and lowest operational conductance state of the device. As shown in Supporting Figure S9, the weight update is proportional to the pulse length.

Filamentary-RRAM fabrication and measurement:

The filamentary-RRAM data was obtained from the raw data set in a previously-published work on wafer-scale Ta/TaO_x devices[2]. Each device was sputtered on a CMOS chip in a back end of the line process, and consists of 20-nm TiN bottom contact, 10-nm TaOx switching layer, 15-nm Ta base layer, and 20-nm TiN top contact. W and Al lines were grown below and above the memristor to connect the device to the CMOS layers. A train of 1000 pulses were used for both SET and RESET operations; however, because the conductances quickly saturate, only the data for the first 50 pulses are shown and analyzed in Figure 1b,c.

MNIST crossbar simulations:

The MNIST crossbar simulations (Support Figure S6) were performed using the CrossSim simulation platform which is a physics realistic approach to using emerging non-volatile memory for neural networks simulations. CrossSim models every cross bar of non-volatile memory devices as a neural core, and transforms ideal numerical updates as computed through the back-prop algorithm into device-realistic updates that would reflect within the devices using a probabilistic map known as a look-up-table . More details on the Numeric algorithm and the way in which look-up-tables are used to emulate realistic device-to-device non-linearity, asymmetry, and stochasticity are given at: https://cross-sim.sandia.gov/_assets/documents/crosssim_manual.pdf . For the MNIST task, the neural network chosen was a multi-layer perceptron with 2 neural cores (crossbars) and one hidden layer. This network has the following dimensionality: 784x300x10. A fixed learning rate of 10^{-4} was used for both bulk-RRAM and contrasting TaOx MNIST simulations, using respective look-up-tables for these two devices constructed based on experimental measurements.

Electrical write energy estimate

The electrical write energy was calculated by multiplying the write voltage (1.5V) by the amount of charge (Q) needed for a weight update. Q was computed by taking the integrated charge needed to change the switching layer conductance by the desired amount, in this case from 5 μS to 15 μS, from the cyclic voltammetry in Supporting Figure S8a, and dividing it by 100 analogue states. This results in 43 nC per weight update, or 65 nJ, for an $(8\text{-mm})^2$ device. The projected energy was then computed by assuming the write energy is proportional to the device size, yields 1 fJ for a $(1\ \mu m)^2$ device and 10 aJ for a $(100\text{-nm})^2$ device.

Thermal heating power estimate

The thermal heating power was estimated assuming that bulk-RRAM can be designed to operate at the same temperature as the Si junction. We anticipate operating around 150°C, which is around the maximum junction temperature for Silicon[3] and only slightly less than the bulk-RRAM temperature for micro-second switching (160°C, Figure 1f). A 1 cm^2 Si chip at 150°C is expected to consume ~10W of thermal power; this is based on the ~5W consumed for a similarly-sized Qualcomm 855 system-on-a-chip[4] that operates at a junction temperature ~90°C and the assumption that the operating power is proportional to the junction temperature minus the ambient temperature.

If each bulk-RRAM device occupies $(100\ nm)^2$ of chip space, and each switching event requires 10^{-5} seconds, then the 10 W of power equates to ~10^{-14} J per switching event. We note that this assumes that bulk-RRAM operates in a highly efficient "training accelerator" mode whereby the chip is continuously training. For inference operations that conducts matrix vector

multiplications with pretrained weights, the device would only need heating for small amounts of time (e.g., once a day to reload weights). As a result, even if the energy per switching even were higher, there still exists significant energy advantages from conducting inference using crossbars of resistive memory in a highly-efficient manner.

Physical modelling for RRAM

The physical model for filamentary-RRAM was conducted using a COMSOL model is detailed in past work[5]. The switching layer is 4 nm thick; the base layer is 30 nm thick. Both layers are 40 nm wide. The filament was formed (Figure 2f) by applying a forming voltage of 1.9V for 1.9 seconds, and then a reset voltage of -1 V for 1 seconds. Several areas of high oxygen vacancy concentration were placed to seed the filament. The SET/RESET simulations in Supporting Figure S7a were conducted in the following manner. 20 set pulses were applied with a compliance current of 250 μA; each set pulse consists of a triangular wave that starting from 0V initially, 1.9V at 5 μs, and 0V at 10 μs. Next, 20 triangular reset pulses were also applied, except with an amplitude of -1.2V without current compliance. The conductance was measured at the end of a switching cycle, when the device has cooled, by dividing the simulated current from the applied voltage of -0.1V. We note that the voltage here is applied on the contact closer to the base layer, while the contact of the switching layer is grounded. This is to be consistent with the bulk-RRAM, where one of the switching layer contacts must be grounded to "read" the conductance.

The physical model for bulk-RRAM modified the filamentary-RRAM model. To minimize changes to the model and precisely determine the origins for the difference in behavior, we use the same geometry and material property of Ta_2O_5 and TaO_x for the switching and base layers. A 4-nm-thick electrolyte was added between the base and switching layer; the ionic conductivity of the electrolyte is 1/100 of the value in Figure 3d because real electrolytes are much thicker. The ion flux in the base and switching layer are solved using the drift-diffusion equation combined with the continuity equation, just as in the filamentary-RRAM model:

$$\frac{\partial n}{\partial t} = \nabla \cdot (D\nabla n - v\, n_d) \tag{1}$$

where n is the spatial concentration of oxygen vacancies, D the temperature-dependent diffusivity, and v the drift velocity. The thermoelectric Soret diffusion term was neglected in this model due to isothermal conditions. The first term on the right-hand side is Fick's first law of diffusion, while the second term is the drift term. D is given by

$$D = \frac{1}{2} a^2 f \exp\left(-\frac{E_A}{k_B T}\right)$$

where a is the lattice jump distance (0.1 nm), f the attempt frequency (10^{12} s^{-1}), and E_A the diffusion activation energy (0.85 eV), k_B the Boltzmann constant, and T the temperature. The drift velocity v is given by

$$v_n = a\, f \exp\left(-\frac{E_A}{k_B T}\right) \cdot \sinh\left(\frac{qa\, \nabla\Psi}{k_B T}\right)$$

where q is the charge of an oxygen vacancy, which is twice the elementary charge. Ψ is the electrical voltage, and its gradient $\nabla\Psi$ is the electric field.

Equation (1) is combined with Ohm's Law to define the voltage:

$$\nabla \cdot (\sigma \nabla \Psi) = 0$$

where the conductivity σ equals the sum of the electronic conductivity (σ_e) and the ionic conductivity (σ_{ion}) of oxygen vacancies. $\sigma \nabla \Psi$ gives the current density, which can be provided by electrons or ions. The applied voltage is given as the boundary condition at the top and bottom edges of the device.

In practice, within the base and switching layers, the electronic and ionic conductivity act in parallel, so the conductivity effectively equals the much larger σ_e. In the electrolyte, where the electronic conductivity is nonexistent, the σ_i dominates. Additionally, σ_i in the electrolyte is more than eight orders of magnitude smaller than σ_e in the base and switching layers, so the voltage drop effectively happens only in the electrolyte. The much lower electrode σ_i, which is connected in series with the electronically conductive base and switching layers, explain why the total current in bulk-RRAM is much lower than that of filamentary-RRAM, and cannot induce significant joule heating.

Finally, because the current is exclusively carried by oxygen vacancies in the electrolyte, we set the flux of oxygen vacancies to equal the current divided by the charge of oxygen vacancies:

$$j_{ion,electrolyte} = \frac{\sigma_i \nabla \Psi}{q}$$

this j_{ion} denotes the flux of ions in the electrolyte and also serves as the boundary condition for equation (1) for the base and switching layer. To minimize the changes from the RRAM model, we assume no interfacial resistance between the electrodes and the electrolyte. Because the voltage drop occurs exclusively in the electrolyte within this model, we define the electric field $\nabla \Psi$ as the applied voltage divided by the electrolyte thickness. A full bulk-RRAM model would need to account for both this interfacial resistance as well as a dependence of the electrochemical potential on the ion concentration and may lead to some deviations from perfectly linear and symmetric behavior. Experimental results show minimal amounts of nonlinearity (Figure 1f, 3a).

A thermal boundary condition set to 473K is also used to represent external heating. We retained the joule heating equation in order to confirm that the current densities here are too low to induce internal heating (Figure 2e).

Initially, 200 pulses of 10 milliseconds each with +1V was applied to increase the conductance during an effective "forming" process (Figure 2e) with the same initial conditions as in the filamentary-RRAM case. To switch, we apply 100 pulses of -1V (reset), then 100 pulses at +1V (set), and 100 pulses at -1V (reset). The conductance was measured by averaging the electronic conductance of the switching layer normalized by the conductance of the final pulse. The simulated conductance of the final 200 pulses were plotted in Supporting Figure S7c.

Proposed temperature correction for synaptic weight

It is ideal for the synaptic weight to be constant during reading and writing even though the electronic conductances of bulk-RRAM differ with temperature. We show here that this is possible by conducting a column-by-column correction to the weight that depends on the temperature.

In order to obtain negative weights[6], we would pair a memory element with conductance G_i with a reference "fixed bias" material G_{fixed}. We start with the room temperature (RT) configuration, where the vector sum Z^{RT} of each column can be written as:

$$Z^{RT} = k^{RT} \sum_i V_i [G_i^{RT} - G_{fixed}^{RT}]$$

where *k* is a scaling factor implemented in software.

We assume that G_{fixed} has the same temperature dependence as G_i, which can be done if G_{fixed} is also a bulk-RRAM cell, biased at its equilibrium value. Since the electronic conductance is linearly dependent on temperature (Supporting Figure S8b), we can write the following expression:

$$G^{hot} = \alpha\, G^{RT} + \beta$$

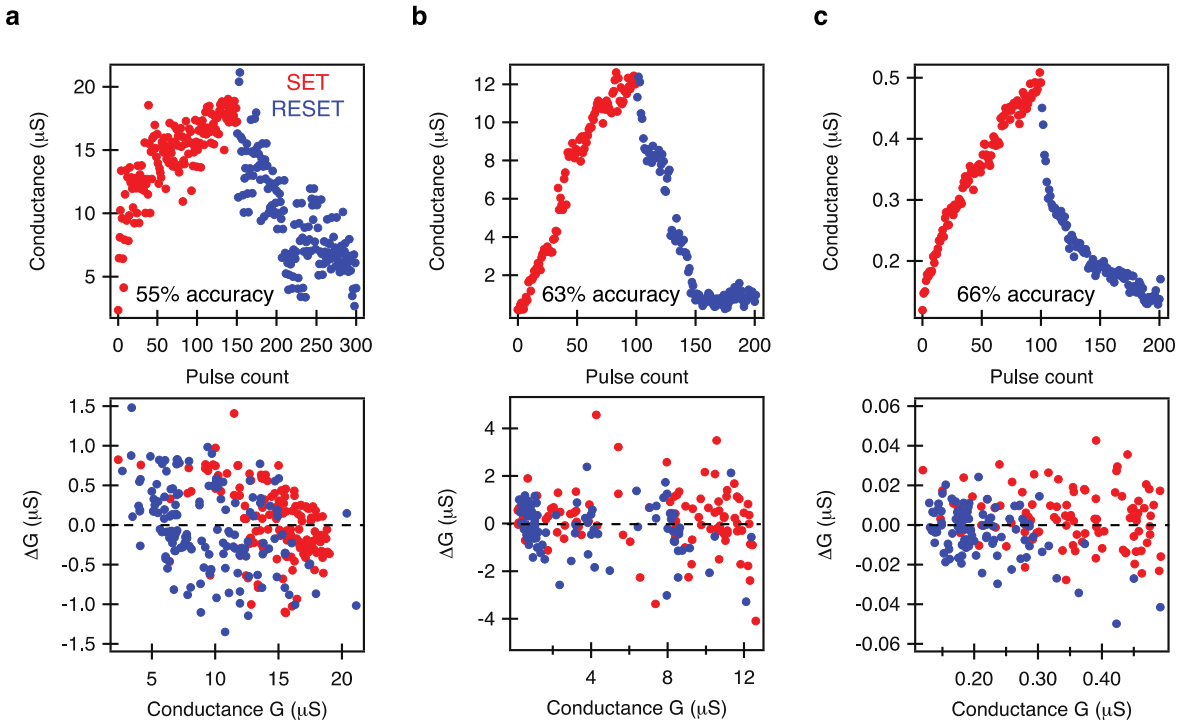
where α and β are arbitrary values to designate the linear dependence of G^{hot} and G^{RT} for all conductance values. Substituting these into Z, we write:

$$Z^{hot} = k^{hot} \sum_i V_i [G_i^{hot} - G_{fixed}^{hot}] = k^{hot} \sum_i V_i [\alpha\, G_i^{RT} + \beta - \alpha\, G_{fixed}^{RT} - \beta]$$

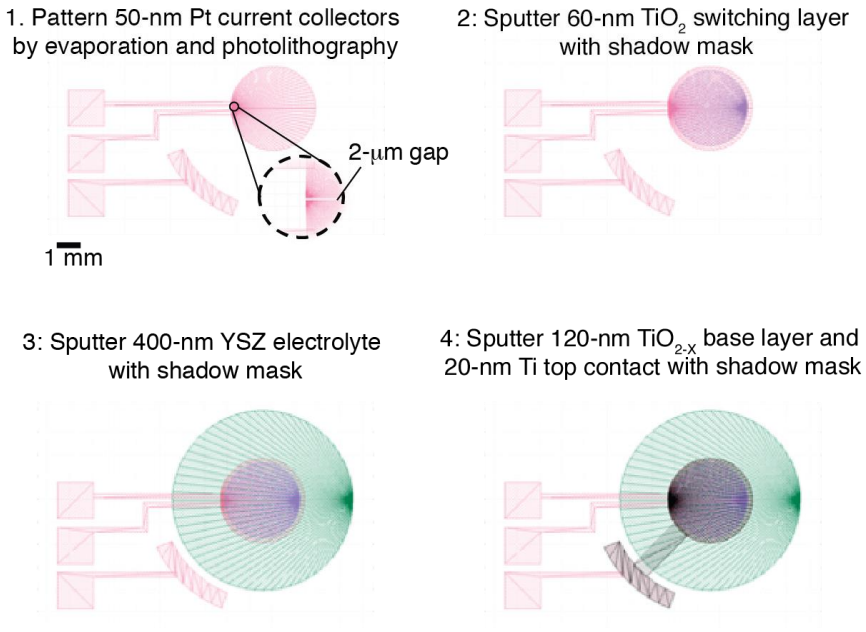
this equates to

$$Z^{hot} = \alpha\, k^{hot} \sum_i V_i [G_i^{RT} - G_{fixed}^{RT}]$$

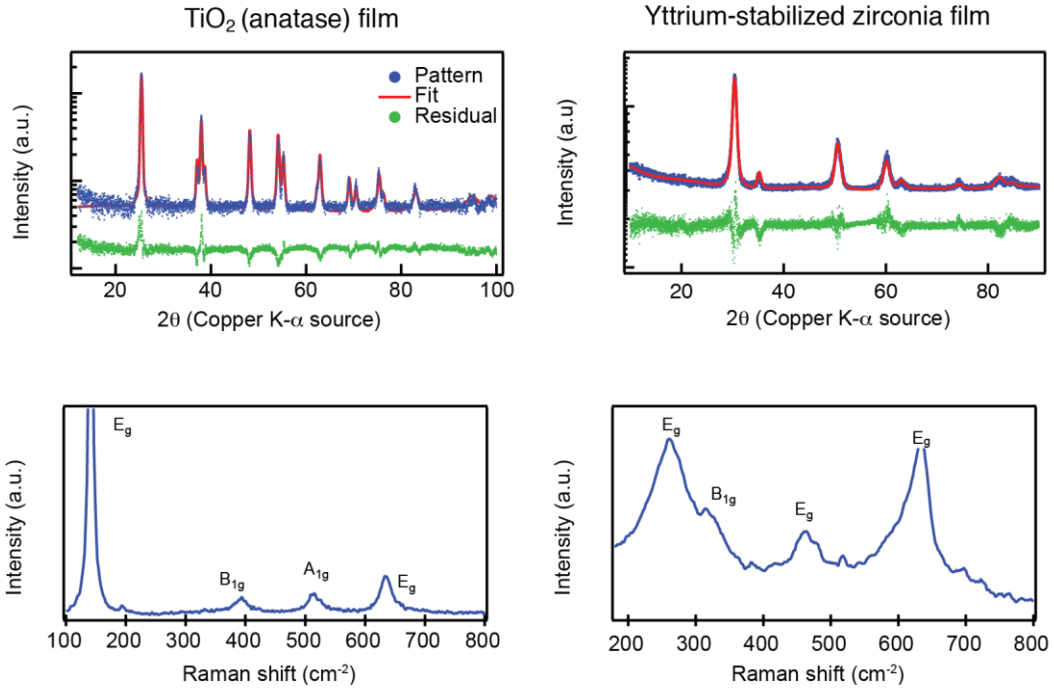
In order for $Z^{hot} = Z^{RT}$, we simply need to control the software scaling factors of k^{RT} and k^{hot} such that $k^{RT} = \alpha k^{hot}$.



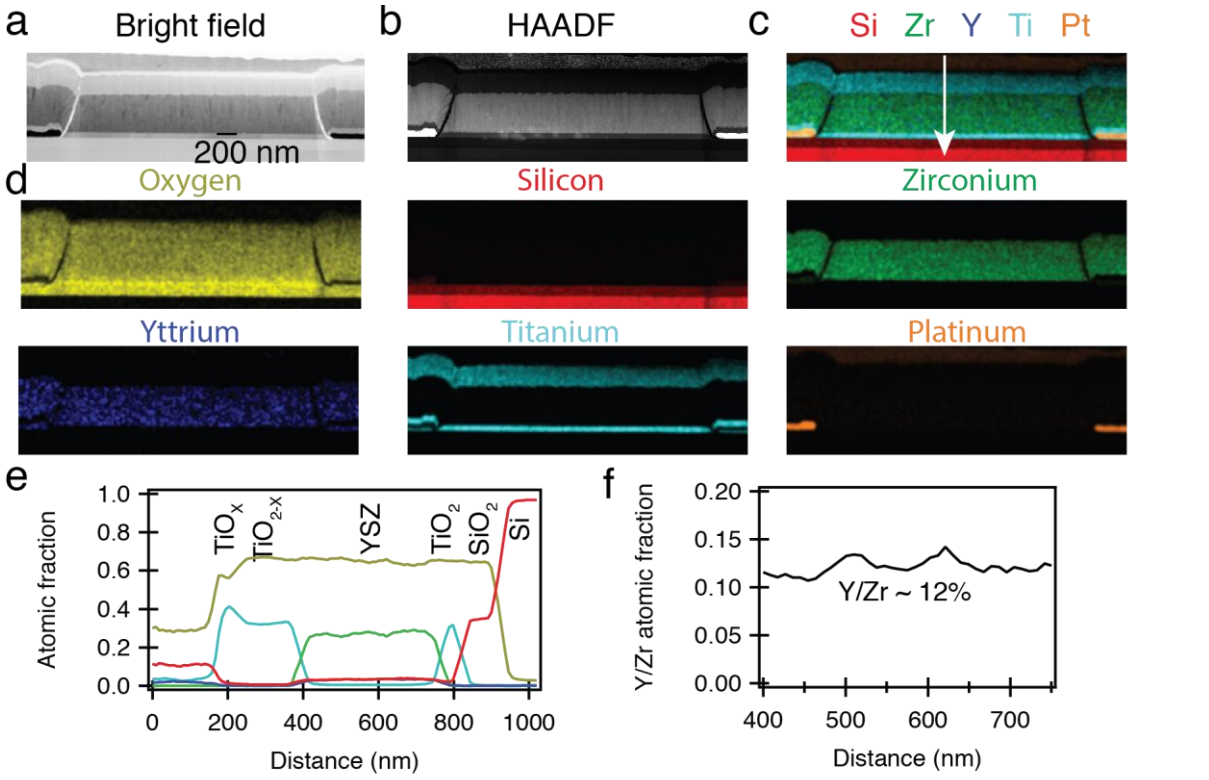
Supporting Figure S1: Analogue switching profiles for several filamentary memristors, both cation and anion types, published in the literature show ~60% of pulses switches in the correct direction versus 50% for purely random switching and 96-99% for bulk-RRAM in Figure 1g, 3b. (a) Yao et al.'s bilayer HfO_x/TaO_x RRAM[7] has a switching accuracy ~55%. (b) Choi et. al's SiGe epitaxial conducting-bridge random access memory[8] has ~63% switching accuracy. (c) Zhu et al.'s Li_xMoS_2 layered memristor[9] has ~66% switching accuracy. The switching accuracy is defined as the fraction of switching pulses that change the conductance state in the correct direction (higher or lower), consistent with Figure 1c, 1g, and 3b.



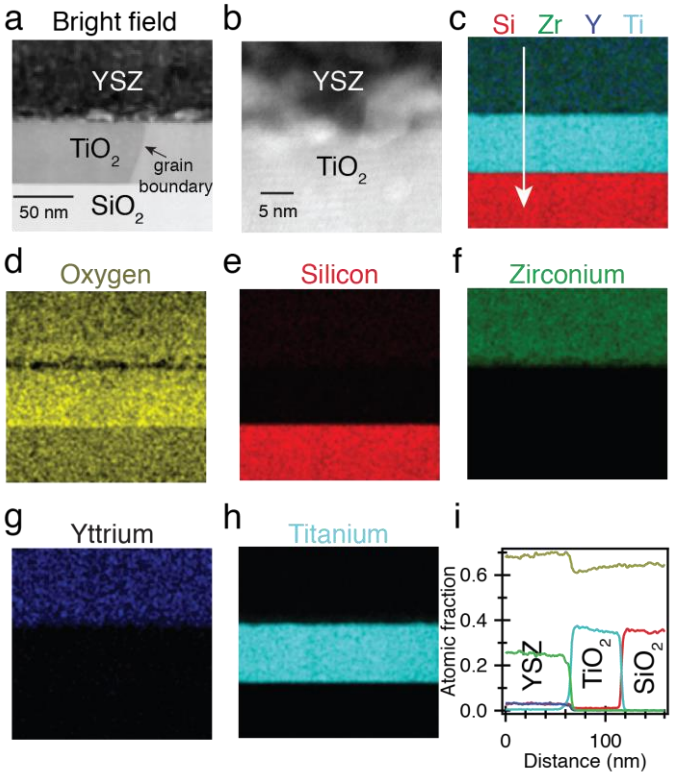
Supporting Figure S2: Fabrication process flow for the thin-film bulk-RRAM device characterized in Figure 1d-g. Each color indicates a different layer, where red is Ti/Pt, purple is TiO_2, yellow is YSZ, and black is TiO_{2-X}/Ti base layer and contact.



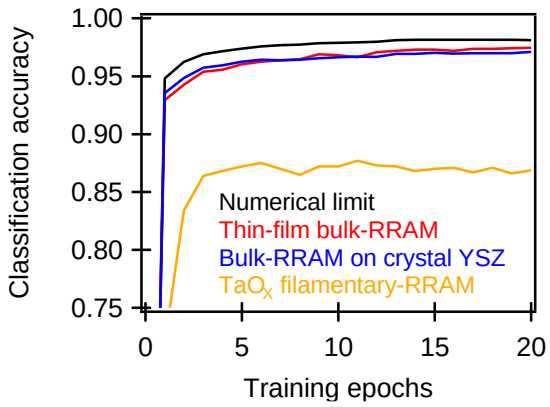
Supporting Figure S3: X-ray diffraction and Raman spectroscopy of the thin films used in this work. Rietveld refinement of TiO_2 to cif file 9015929 and yielded a tetragonal structure with lattice constants of a=3.780(1) Å and c = 9.485(2) Å. Rietveld refinement of YSZ to cif file 1528644 yielded a cubic structure with lattice constant of a=5.121(1). Raman spectroscopy shows the expected vibrations.



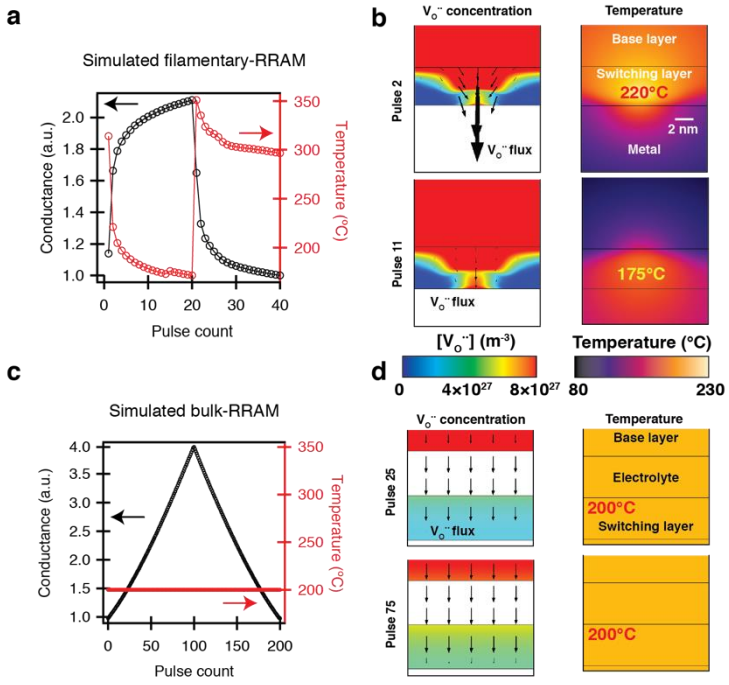
Supporting Figure S4: Chemical mapping using energy dispersive spectroscopy (EDS) inside a scanning transmission electron microscope of the bulk-RRAM cell. (a) Bright-field and (b) high-angle annular dark field (HAADF) image of the cell. (c) EDS map of several elements overlayed. (d) EDS map of individual elements. (e) Linescan of the atomic fractions of each element shows the presence of all layers with minimal mixing between the layers. The line is shown in (c). (f) The average Y/Zr atomic fraction as computed by energy dispersive spectroscopy is 12% ± 3%, slightly less than the 17% in the target; part of this may result from quantification error in the standard-less EDS mapping.



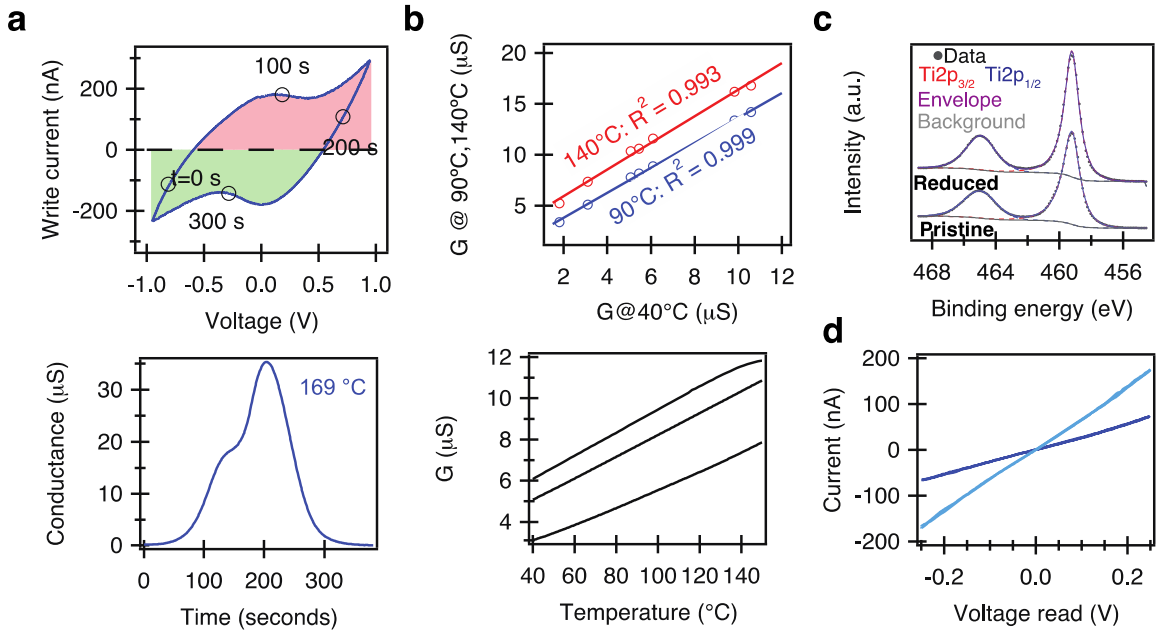
Supporting Figure S5: Scanning transmission electron microcopy of the interface between YSZ and TiO_2. (a-b) Bright field image of the interface. The YSZ is polycrystalline with very small grain sizes, while TiO_2 has larger grains with a visible boundary in (a) and lattice fringes in (b). (c-h) EDS mapping of all elements (c) and individual channels (d-h) of this interface. (g) Linescan (from c) of the atomic fractions from the EDS mapping. The individual EDS maps as well as the linescan show little evidence of cation mixing at the YSZ/TiO_2 interface.



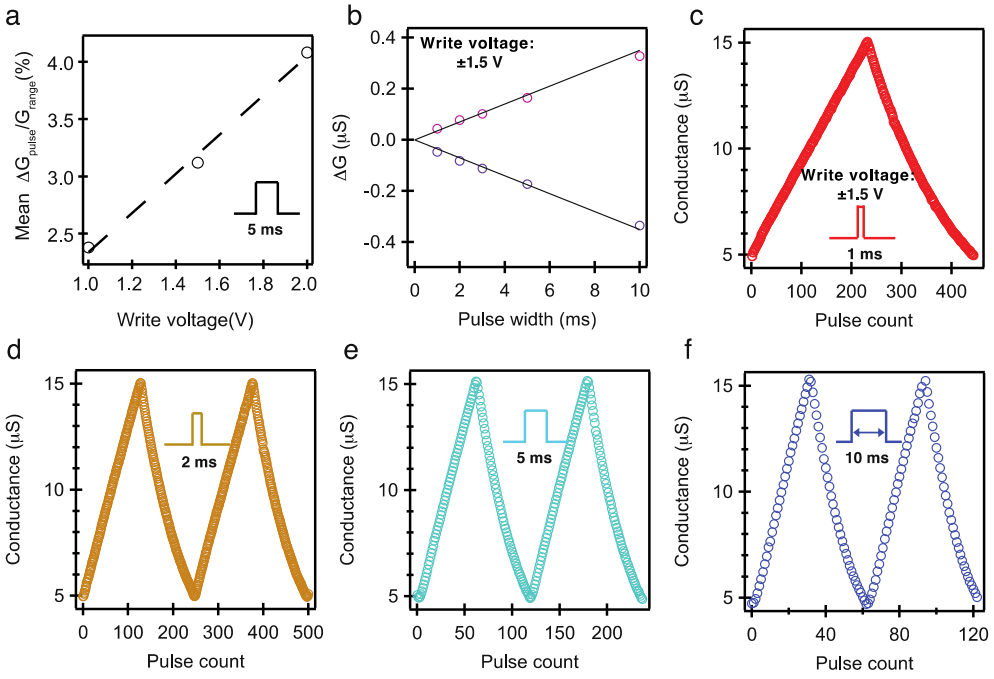
Supporting Figure S6: Crossbar simulations for the MNIST data set of bulk-RRAM exceeds 97%, close to the numerical limit of 98% and much higher than ~88% for filamentary RRAM using TaO_x. The distribution tables used for the simulations are shown in Figure 1c, 1g, and 3b.



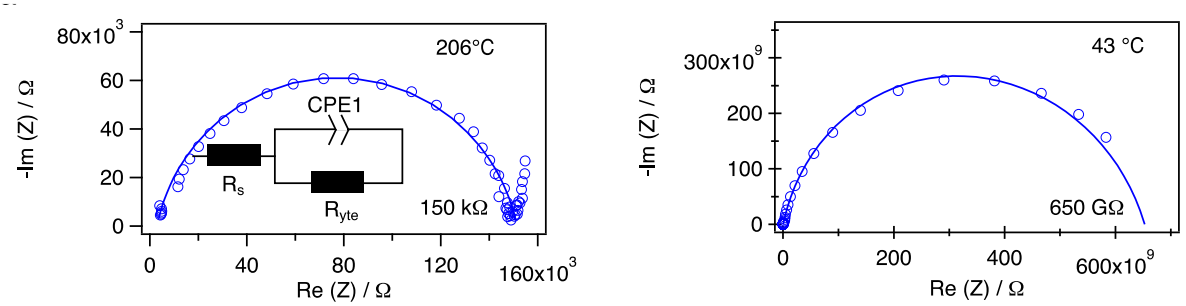
Supporting Figure S7: Physical modelling of switching for filamentary- and bulk-RRAM. (a) Simulated potentiation SET (pulse 1-20, 1.9V, 10 µS) and depression RESET (pulse 21-40, -1.2V, 10 µS) switching profiles for filamentary-RRAM show nonlinear and asymmetric switching due to differences in the devices temperature and resulting $V_O^{\cdot\cdot}$ mobility from internal joule heating. The joule heating power equals I^2G^{-1} for SET with a compliance current and V^2G for RESET without a compliance current. This deterministic model is able to simulate nonlinear switching but does not capture stochastic switching. (b) Because $V_O^{\cdot\cdot}$ mobility is strongly dependent on temperature, the $V_O^{\cdot\cdot}$ flux into the filament (black lines) can vary significantly between different pulses, 10^{23} $m^{-2}s^{-1}$ for pulse 2 but 10^{22} m^2s^{-1} for pulse 10, as their temperatures differ. (c) Bulk-RRAM achieves linear and predictable switching at constant temperatures. Pulses 1-100 is SET (1V, 10 ms), while 101-200 is RESET (-1V, 10 ms). (d) Because the temperature is constant, the $V_O^{\cdot\cdot}$ flux (~10^{19} $m^{-2}s^{-1}$) is uniform and independent of G, resulting in linear switching. (b) and (d) uses the same color scale.



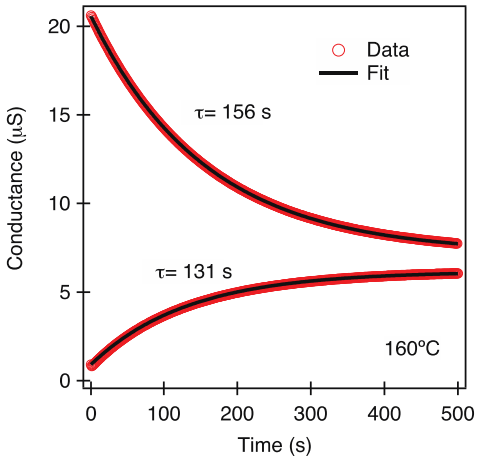
Supporting Figure S8: Electrochemical and electrical characterization of the model bulk-RRAM device built on a 100-μm-thick single-crystal YSZ substrate. (a) Cyclic voltammetry applied between the base and switching layer contacts at a sweep rate of 3 mV sec^{-1}. Positive currents between contact 1 and 3 indicate reduction of the switching layer (increased conductance), and negative currents indicates oxidation (decreased conductance). The relationship is hysteretic and typical of electrochemical systems whereby the defect concentration of the switching layer is not dependent on the voltage, but rather the total integrated current because every two electrons are associated with the migration of one oxygen vacancy[10]. The write applied voltage is lower than in the Figure 3 due to the much longer sweep times. The switching layer conductance measured by applying 100-mV "read" voltage between contacts 2 and 3 on the switching layer and measuring the current. (b) Temperature dependence of the conductance at different analogue information states. The conductance at an elevated temperature can be predicted from the conductance at a lower temperature using this line. (c) Despite the increased electronic conductance from oxygen vacancies, the concentration of Ti^{3+} at the surface is too low to be detected by X-ray photoelectron spectroscopy. We note that even 1% Ti3+ constitutes $>10^{20}$ cm^{-3} electron donors, significantly higher than what is needed to show a significant change in electronic conductivity. (d) Linear current-voltage profiles between contacts 2 and 3 of the switching layer are consistent with bulk compositional modulation as opposed to changes in the interfacial resistance at the Pt/TiO_{2-X} interface. Only 0.1V was applied between contacts 2 and 3 to measure the conductance in this work.



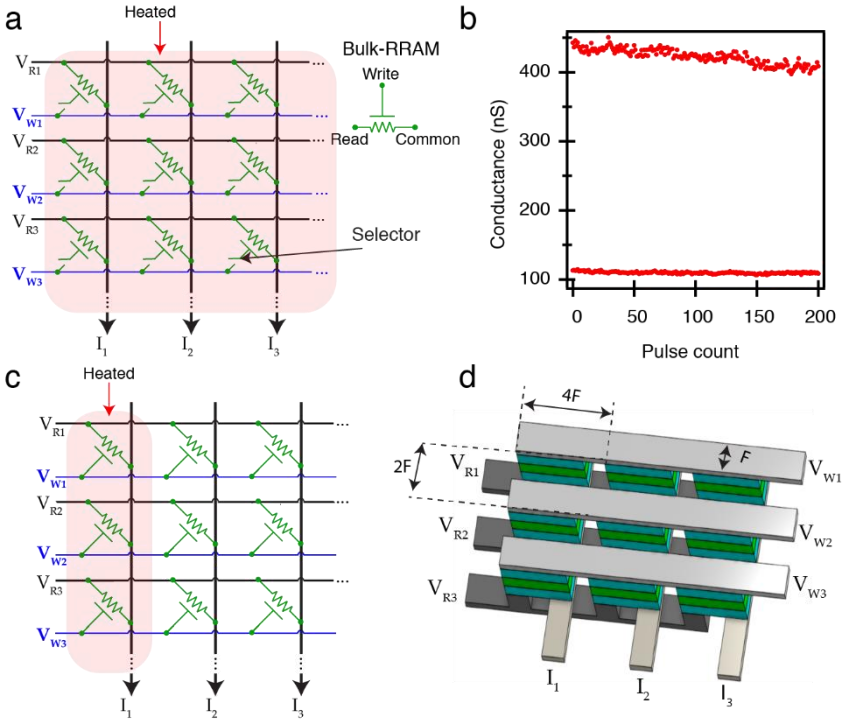
Supporting Figure S9: Highly deterministic and predictable switching profile of the bulk-RRAM device fabricated on 100-μm single-crystal YSZ substrate. All measurements were taken at 206ºC. (a-b) The change in conductance for one pulse is approximately proportional to the voltage and the pulse width. The conductance change can be predicted precisely from the voltage and pulse width. (c-f) Ramping profiles of the device at different pulse widths shows the ability to tune the number of analogue states as well as attain a specific state using the pulse width.



Supporting Figure S10: Nyquist impedance of the 100-μm-thick single-crystal YSZ used for determining the ionic resistivity in Figure 3d. The circuit model for each device is shown on the plot. The data is shown by unfilled circles and the fit is shown by the solid line. The R_{yte} is taken to be the ionic resistance of the electrolyte and used to compute the ionic resistivity.



Supporting Figure S11: Bulk-RRAM devices on single-crystal YSZ relaxing to equilibrium when the base and switching layers are shorted. The data are fit by a single exponential decay, consistent with an RC time constant model. During relaxation, the conductance values taken by high- and low- conductance curves span the entire conductance range. As a result, the time-shifted relaxation curves also provide an estimate for relaxation from intermediate conductance values.



Supporting Figure S12: Proposed array-level weight update schemes for the bulk-RRAM. (a) The parallel weight update scheme from ref. [11] would heat all devices simultaneously and use electronic selectors as switches to conduct parallel outer product updates. To maximize device density, these selectors will be configured in a 1T1R or 1S1R configuration similar to that of memristive crossbars[12]. Unlike the case for electrochemical synaptic transistors, these switches are used for selection, not for long-term state retention. (b) The selectors would provide sufficient short-term retention in order to complete training while the device operates at elevated temperatures. The experiment was conducted at 160℃ where each pulse is about 300 ms, significantly longer than the write time of 2 μs in Figure 1f. CMOS switches

(MAX327CPE) were used as the selector. (c) A simpler selector-free configuration for inference can be constructed; only one column is heated and updated at any one time. (d) Proposed $8F^2$ crossbar array schematic for bulk-RRAM, where F is the feature size.

Supporting References:

[1] Y. Li, E. J. Fuller, S. Asapu, S. Agarwal, T. Kurita, J. J. Yang, A. A. Talin, *ACS Appl. Mater. Interfaces* **2019**, *11*, 38982.

[2] C. H. Bennett, D. Garland, R. B. Jacobs-Gedrim, S. Agarwal, M. J. Marinella, in *IEEE Int. Reliab. Phys. Symp. Proc.*, IEEE, **2019**.

[3] *32-Bit Flash Microcontroller with MIPS32® MicroAptivTM UC Core with Low Power and Low Pin Count*, **2018**.

[4] K. Hinum, "Qualcomm Snapdragon 855," can be found under https://www.notebookcheck.net/Qualcomm-Snapdragon-855-SoC-Benchmarks-and-Specs.375436.0.html, **2019**.

[5] S. H. Lee, J. Moon, Y. Jeong, J. Lee, X. Li, H. Wu, W. D. Lu, *ACS Appl. Electron. Mater.* **2020**, *2*, 701.

[6] S. Agarwal, R. B. J. Gedrim, A. H. Hsia, D. R. Hughart, E. J. Fuller, A. A. Talin, C. D. James, S. J. Plimpton, M. J. Marinella, in *2017 Symp. VLSI Technol.*, JSAP, **2017**, pp. T174–T175.

[7] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, H. Qian, *Nature* **2020**, *577*, 641.

[8] S. Choi, S. H. Tan, Z. Li, Y. Kim, C. Choi, P. Chen, H. Yeon, S. Yu, J. Kim, *Nat. Mater.* **2018**, *17*, 335.

[9] X. Zhu, D. Li, X. Liang, W. D. Lu, *Nat. Mater.* **2019**, *18*, 141.

[10] E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, A. Salleo, A. A. Talin, *Science.* **2019**, *364*, 570.