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, A. Alec Talin*

Dr. Y. Li, Dr. E. J. Fuller, Dr. J. D. Sugar, Dr. D. S. Ashby, R. D. Horton, Dr. M. S. Bartsch, Dr. A. A. Talin, Sandia National Laboratories, Livermore, CA, 94550 USA

S. Yoo, Prof. W. D. Lu, Department of Electrical Engineering and Computer Science, Ann Arbor, MI, 48109, USA

Dr. C. H. Bennett, Dr. M. J. Marinella, Sandia National Laboratories, Albuquerque, NM, 87185, USA

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

#Present address, Department of Materials Science and Engineering, University of Michigan

Keywords: resistive switching, deterministic, point defects, neuromorphic computing

Abstract: **Digital computing is nearing is physical limits as computing needs and energy consumption rapidly increase. Analogue memory-based neuromorphic computing can be orders of magnitude more energy efficient at data-intensive tasks like deep neural networks, but has been limited by the inaccurate and unpredictable switching of analogue resistive memory. Filamentary resistive random access memory (RRAM) suffers from stochastic switching due to the random**

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/adma.202003984.

This article is protected by copyright. All rights reserved.

Author Manuscript

kinetic motion of discrete defects in the nanometer-sized filament. In this work, we overcome this stochasticity by incorporating a solid electrolyte interlayer, in this case yttria-stabilized zirconia (YSZ), towards eliminating filaments. Filament-free, bulk-RRAM cells instead store analogue states using the bulk point defect concentration, yielding predictable switching because the statistical ensemble behavior of oxygen vacancy defects is deterministic even when individual defects are stochastic. Both experiments and modeling show bulk-RRAM devices using TiO_{2-X} switching layers and YSZ electrolytes yield deterministic and linear analogue switching for efficient inference and training. Bulk-RRAM solves many outstanding issues with memristor unpredictability that have inhibited commercialization, and therefore could enable unprecedented new applications for energy-efficient neuromorphic computing. Beyond RRAM, our work shows how harnessing bulk point defects in ionic materials can be used to engineer deterministic nanoelectronic materials and devices.

Although CMOS-based digital memory and processors have achieved enormous advances in computing, they may not be optimal to meet future computing requirements. Data-intensive operations including machine learning and artificial neural networks are particularly costly in energy due to the need to move information between the memory and the processor. On the other hand, analogue neuromorphic computing processes information directly on the memory elements[1–3] to bypass this bottleneck, making such systems hundreds of times more energy efficient[4]. Neuromorphic computing architectures for fully-connected[5–7] and convolutional[8,9] neural networks have been developed. Despite significant research into memory technologies such as conductive-bridge random access memory[10–12], ferroelectric memory[13], phase-change memory[14–16], among others, the search for a CMOS compatible analogue non-volatile memory element, or artificial synapse, with accurate and efficient switching has been elusive.

Filamentary-RRAM has demonstrated tremendous potential as analogue memory due to its scalability, non-volatility, fast switching, and CMOS compatibility[17–26], but suffers from severe challenges in achieving predictable analogue behavior. Filamentary-RRAM stores analogue

This article is protected by copyright. All rights reserved.

Author Manuscript

information in a conductive filament formed by oxygen vacancy ($V_O^{\cdot\cdot}$) defects inside a metal oxide switching layer. The localized, nanometer-sized filament arises from the instability of the electroforming process that results from positive thermal feedback[20,27,28]. Applying a write voltage combines two effects to change the resistance state[27,28]: first, a large electronic current creates internal joule heating to several hundred degrees Celsius to locally activate $V_O^{\cdot\cdot}$ mobility[29]; second, a much smaller electrochemical current directs the motion of $V_O^{\cdot\cdot}$ to or away from the filament, changing its size and/or composition.

A well-known challenge of filamentary devices is that the analogue resistance state, set by the position and motion of a discrete number of $V_O^{\cdot\cdot}$ defects in this nanosized filament[20,21,30], is stochastic due to the probabilistic nature of microscopic atomic behavior (e.g. thermally-activated defect "hopping"). This stochasticity is responsible for the intrinsically unpredictable and irreproducible analogue switching in filamentary devices[31–33]. The challenges surrounding stochasticity become acute for analogue devices engineered for higher-resistance operation due to fewer atoms in the critical conduction path[30,32]. Since energy-efficient neuromorphic computing demands both more analogue states and higher resistances[4,34], state-of-the-art analogue filamentary devices often switch in the correct direction only ~60% of the time[5,8,10,25,35], slightly better than random (50%). As a result, it can take nearly 500 switching events to tune the memory cell to one of 32 analogue states[8], expending significant time and energy to achieve modest 5-bit resolution. Additional challenges include nonlinear and asymmetric changes in analogue state that reduces the training accuracy of artificial neural networks[1,34,36].

To address these issues, we introduce the filament-free bulk-RRAM using TiO_2 and YSZ to achieve deterministic switching. An electron-blocking, ion-conducting solid electrolyte interlayer eliminates the positive thermal feedback that creates the dominant conductive filament. Instead,

This article is protected by copyright. All rights reserved.

Author Manuscript

due to uniform temperatures provided by external heating sources, $V_O^{\cdot\cdot}$ defects are homogeneously distributed as a solid solution in the bulk. By widening the active information storage volume from the dominant nanosized filament to the bulk volume of the switching layer, bulk-RRAM employs the statistical ensemble concentration of $V_O^{\cdot\cdot}$ defects in the bulk to store the analogue resistance states. This leads to predictable switching because the ensemble behavior of all defects is deterministic even if individual defects are stochastic. The analogue switching accuracy at high resistances improves from ~60% in filamentary memristors to between 96% and 99% in these first-generation bulk-RRAM devices. Eliminating internal joule heating also enables linear changes in the analogue state for accurate training. While external heating in bulk-RRAM may be less space efficient than internal joule heating, it provides the uniform temperatures to eliminate the filament, and provides a path towards an analogue non-volatile resistive memory solution.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

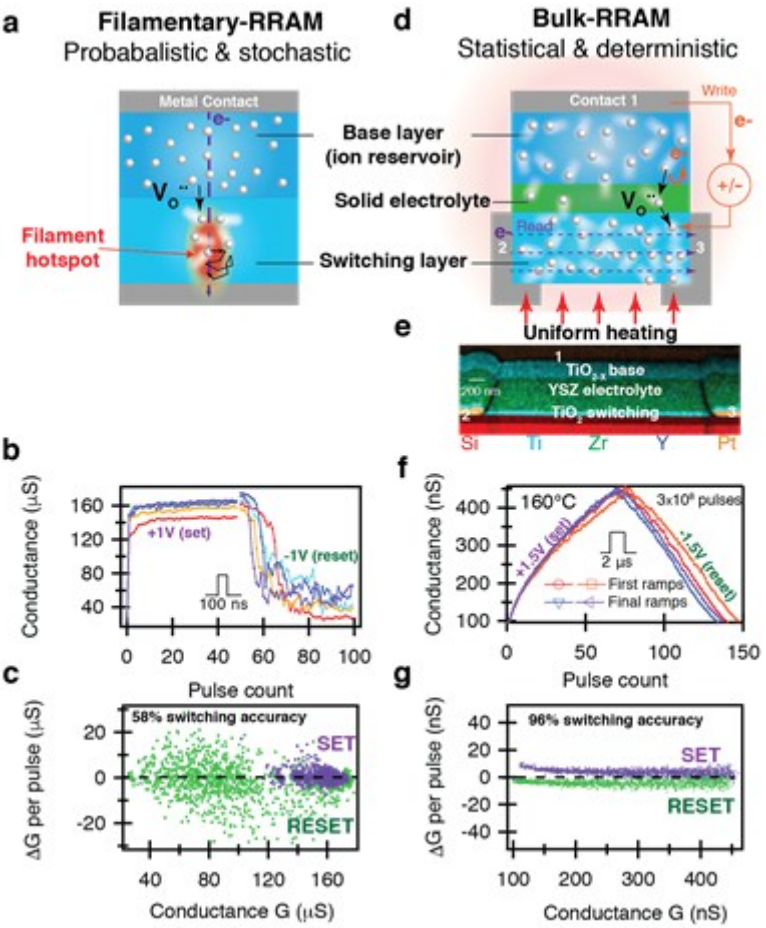


Figure 1: Bulk-RRAM compared to filamentary-RRAM. (a) Filamentary-RRAM uses localized internal joule heating to mobilize $V_O^{\cdot\cdot}$ defects within the nanometer-sized conductive filament. Due to the discrete number of defects within the dominant conducting filament, this memory device switches stochastically and probabilistically, schematically illustrated using a random walk. (b) Analogue switching behavior in a TaO_x filamentary-RRAM device; this data was also previously used for neural network simulations in ref. [35]; the first fifty SET and RESET pulses in each of five ramps are plotted. (c) The switching distribution plot from 50 ramps. The switching accuracy, defined as the number of SET or RESET pulses that changes the state in the correct direction, is ~58%, slightly better than random (50%). SET from the low-conductance state results in ΔG~100 μS, beyond the scale of the graph. (d) Bulk-RRAM utilize the average concentration of $V_O^{\cdot\cdot}$ defects in the switching layer to store analogue information. While each defect follows probabilistic behavior, the statistical behavior of all defects is deterministic, leading to accurate and predictable switching. (e) Cross-sectional energy dispersive spectroscopy (EDS) map taken by scanning transmission electron microscopy, of the bulk-RRAM cell used in this work. Contact 1 is a metallic, highly oxygen-deficient TiO_x, and the substrate is Si with 100 nm of thermal oxide. (f) Bulk-RRAM shows much more linear and deterministic behavior despite lower electronic conductance and more analogue states. The first and final two ramps over 3×10^8 pulses are shown. The ramping switches between SET to RESET when the conductance limits of 100 nS and 450 nS are reached. (g) The switching distribution for ~30 ramps show 96% switching accuracy.

This article is protected by copyright. All rights reserved.

We first compare the switching stochasticity of filamentary- and bulk-RRAM. Filamentary devices contain a direct electronic path between the top and bottom contacts formed by electron-donating $V_O^{\cdot\cdot}$ defects in the metal oxide layers (**Figure 1**a). An initial voltage applied between the top and bottom contacts concentrates the current into a thermal hotspot to "electroform" a nanometer-sized conducting filament in the switching layer[20]. The number and position of $V_O^{\cdot\cdot}$ defects in the filament controls its conductance, which dominates the overall device conductance state. Subsequent write pulses lead to local joule heating that activates $V_O^{\cdot\cdot}$ migration within the filamentary hotspot to change the defect concentration and distribution. To illustrate the effect of stochastic switching and introduce a method to quantify the switching variability, we show a typical analogue ramping profile for a TaO_x filamentary memristor (Figure 1b, see supporting information for details). This data was also used to conduct neural network simulations in ref. [35]. Switching is unpredictable and stochastic, and the distribution of switching events varies considerably from cycle to cycle and device to device[35] . For higher resistance devices, only ~60% of SET or RESET pulses switch in the desired direction (Figure 1c), a result comparable to those of other published works[5,8,10,25] (Supporting Figure S1).

Previous studies have shown that stochastic switching arises because the discrete defects in a single nanometer-sized filament dominate the device conductance. A low probability of defect "hopping" over an ~1 eV activation barrier (~10^{-9} at 300°C) combined with the "random walk" of defects from kinetic theory dictate that discrete defects behave probabilistically and stochastically[31–33]. Poissonian switching statistics[37], shot noise[1], and random telegraph noise[30] adds to unpredictability and stochasticity. Filamentary-RRAM also presents nonlinear and asymmetric switching behavior that makes it difficult to accurately train neural networks[1,34,36].

Author Manuscript

This article is protected by copyright. All rights reserved.

To address these challenges, we design the bulk-RRAM cell (Figure 1d) that does not contain dominant conducting filaments and is instead sensitive to the average bulk defect concentration, a continuous variable that is generalizable to many transition metal oxides. To demonstrate the concept, we fabricated thin-film devices on silicon that adds a 400-nm thick solid electrolyte interlayer, yttria-stabilized zirconia (YSZ), between the mixed conducting base (120-nm thick) and switching (60-nm thick) layers, both consisting of TiO_{2-X} (see supporting information and Figures S2,3 for fabrication protocol and materials characterization). Figure 1e shows a cross-sectional energy dispersive spectroscopy (EDS) map taken using a scanning transmission electron microscope. High-resolution EDS maps do not show cation intermixing between the YSZ and TiO_2 layers (Supporting Figures S4,5). This structure resembles a solid oxide fuel cell[38] without the gas interface.

Because the electrolyte blocks the direct electronic pathway and resulting internal joule heating (Figure 1d), the heat to thermally activate $V_O^{\cdot\cdot}$ migration is supplied uniformly from an external hotplate or on-chip resistive heaters. No electroforming process is needed. Unlike filamentary devices, bulk-RRAM has separate read and write pathways: write pulses, ±1.5 V in magnitude, applied between contacts 1 and 3 shuttles $V_O^{\cdot\cdot}$ defects in and out of the switching layer. The measured resistance state between contacts 2 and 3 (100 mV) samples the average electronic resistivity of the switching layer, which depends on the defect concentration. To demonstrate analogue tunability, the switching layer was ramped from 100-450 nS. Figure 1f plots the first and last two ramps over 3×10^8 write operations, each consisting ±1.5V write pulses applied between contacts 1 and 3 for 2 μs.

Bulk-RRAM shows linear, symmetric, and predictable switching profiles. In stark contrast to that of filamentary devices (Figure 1b,c), the switching behavior is essentially deterministic with minimal cycle-to-cycle variations. Over 96% of the switching pulses change the conductance state in

This article is protected by copyright. All rights reserved.

Author Manuscript

the desired direction (Figure 1g), despite >2 MΩ resistance, more analogue states, and lack of materials optimization compared to the filamentary devices. This predictable switching arises from the absence of localized internal heating: under uniform temperatures, mobile defects are homogeneously distributed as a solid solution due to configurational entropy[38], without the positive feedback conditions for heterogeneous filament growth[28]. Because the number of $V_O^{\cdot\cdot}$ defect sites exceeds 10^6 even in a small $(30\text{-nm})^3$ volume[39], defect migration firmly obeys collective deterministic statistical behavior like Fick's Laws of Diffusion. Moreover, switching is linear and symmetric, essential attributes required to achieve high network training accuracy[1,34,36]: Crossbar simulations (Cross-Sim[36]) show ~97% training accuracy for the MNIST training set (Supporting Figure S6). While this device is a type of electrochemical random-access memory (ECRAM) [40–49], the bulk-RRAM cells based on oxygen vacancies presented here provides significant advantages in terms of scalability, retention, and CMOS compatibility over previously-developed ECRAM, and will be discussed later.

Next, we probe the switching layer's local conductivity with both materials characterization and physical modelling to confirm the absence of filaments. Experimentally, we map the electrical conductivity with conductive atomic force microscopy (c-AFM) using a modified device geometry with exposed switching layers (see supporting information). **Figure 2**a,b shows the local tip current of a region of the switching layer in two conductance states, 5 μS and 1 μS. No current hotspots were detected, in stark contrast to hotspots frequently observed for filamentary-RRAM arising from conductive filaments[17,50]. The average tip current for all images in the high-conductance state is about 4.9 times higher than that of the low-conductance state, in agreement with macroscopically-measured values.

Author Manuscript

This article is protected by copyright. All rights reserved.

To quantify the inhomogeneity in local conductivity, we divide the c-AFM map into square regions of different sizes during analysis, and calculate the mean tip current $\bar{x}$ for each region (Figure 2c). The error bars represent two standard deviations (2s) of the distribution: our results show that 94/100 of the $(25\ nm)^2$ regions have tip currents between 70% and 130% of the mean current $\bar{x}$. Given that bulk-RRAM devices have 5× conductance range (Figure 1f), c-AFM results suggest that nanoscale devices with switching layers on the order of 25-50 nm would be sufficiently uniform for analogue memory crossbars. The $2s/\bar{x}$ ratio obtained from a number of c-AFM maps for both states are plotted in Figure 2d. The highly uniform, yet low conductance of the oxide at different states suggests that large crossbars with $> 10^6$ synapses are indeed possible to yield at least 10^{14} multiply-and-accumulate operations per joule, hundreds of times improvement over optimized digital computing[4].

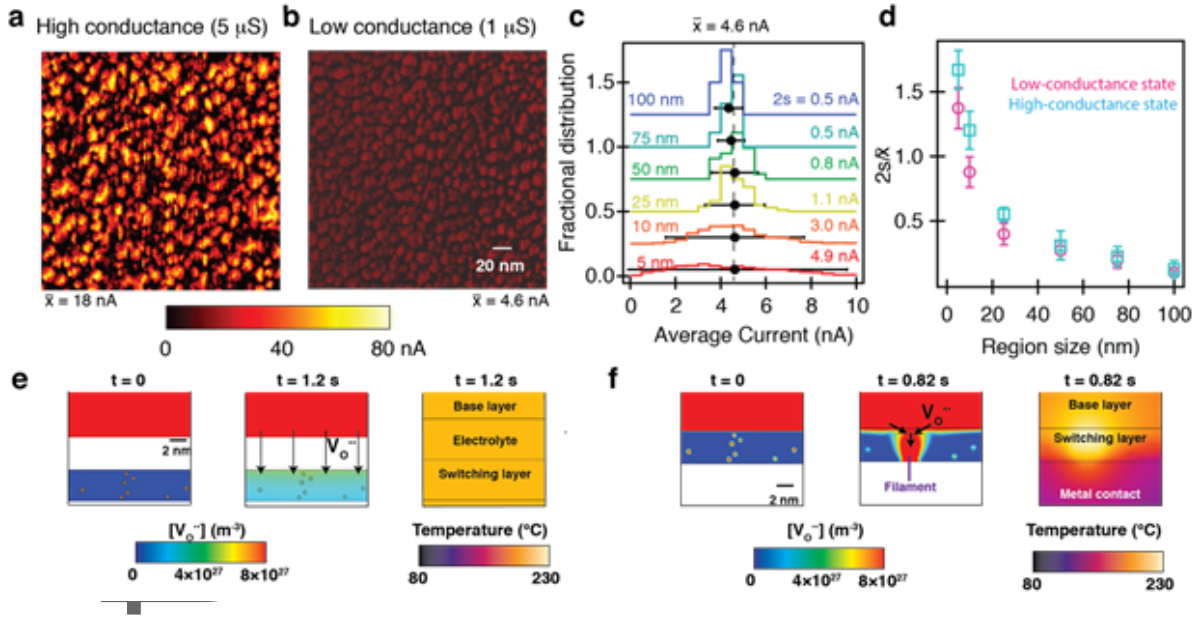


Figure 2: Nanoscale spatial distribution in electronic conductivity demonstrates non-filamentary nature of the device as well as potential scalability to smaller dimensions. (a-b) Conductive-AFM tip current distributions for the switching layer at high-conductance and low-conductance states. The tip voltage is -2V, and the absolute value for the current is plotted. No filaments were observed. (c) Histogram distribution of the tip current in (b) averaged across regions of different sizes, used to evaluate the uniformity of the conductance as a function of dimensions. The black dots represent the image-averaged current $\bar{x}$. The error bars *2s* (twice the standard

This article is protected by copyright. All rights reserved.

deviation) encompasses ~95% of the distribution of region-averaged currents. (d) Plot of the uniformity of the different regions using *2s/x̄* as a metric for the uniformity: for example, our results state that ~95% of the 50-nm regions in the low-conductance state have conductance within 25% of the mean value. The confidence intervals represent the standard deviation of (*2s/x̄*) across three c-AFM images for the low-conductance state and four images for the high-conductance state. (e) Physical modelling confirms the uniformity of the switching layer in bulk-RRAM as a result of diffusion and configurational entropy when the temperature is uniform, even as the initial distribution of $V_O^{\cdot\cdot}$ defects is nonuniform. The arrows indicate uniform oxygen vacancy flux. (f) In filamentary devices, the filaments form around these initial nonuniformities which act as nucleation seeds. Due to localized temperature increases and positive thermal feedback, defect migration is concentrated around a ~2-nm filament, where $V_O^{\cdot\cdot}$ flux is very high.

We adapt a quantitative and deterministic physical model from past work[27,28] to further study switching in bulk-RRAM (see Supporting Information for details). The temperature within bulk-RRAM with externally-controlled heating is constant (Figure 2e); as a result, $V_O^{\cdot\cdot}$ defects enter the switching layer uniformly from the base layer via the electrolyte. In contrast, internal joule heating in filamentary devices concentrates the electronic current into regions that are initially more conducting, turning them into hotspots (Figure 2f). These hotspots serve as positive feedback nucleation points for the filament. The filament is nanometer-sized even when the device is much larger, leading to the discrete stochastic behavior in Figure 1b,c.

We also use this deterministic model to simulate analogue switching. In filamentary devices, the change in conductance depends non-linearly on the present state, even when stochasticity is not simulated (Supporting Figure S7). This arises because the joule heating power (I^2R or V^2/R) depends on the resistance, yielding resistance- or state-dependent temperature which in turn affects the defect mobility. In contrast, the temperature of bulk-RRAM is controlled independently, so both the defect mobility and the number of defects shuttled per pulse is essentially constant, resulting in the highly linear switching shown experimentally (Figure 1f,g).

We next seek to elucidate and quantify the switching and retention properties using model bulk-RRAM cells fabricated on single-crystal YSZ substrates (100-1000-µm thick, see supporting

This article is protected by copyright. All rights reserved.

Author Manuscript

information for device fabrication details). Cyclic voltammetry between contacts 1 and 3 shows a strong hysteresis typical of electrochemical systems[51], and that the conductance of the switching layer (measured through contacts 2 and 3) increases as $V_O^{\cdot\cdot}$ is inserted (Supporting Figure S8a), suggesting that Ti^{4+} ions are reduced to Ti^{3+} and creating two mobile polarons (Ti'_{Ti}) for every $V_O^{\cdot\cdot}$ inserted. At negative currents, the process is reversed. The electronic conductivity of the switching layer increases with temperature (Supporting Figure S8b), characteristic of polaron conduction[52]. Because the devices are not cooled between weight update pulses, this behavior must be considered when converting analogue weights at different temperatures (see Supporting Information). Despite appreciable electronic conductance, the Ti3+ concentration at the surface is below the detection threshold of X-ray photoelectron spectroscopy (Supporting Figure S8c). Linear current-voltage curves confirm that our weight updates arise from bulk compositional modulation rather than interfacial effects at the Pt/TiO_{2-X} interface (Supporting Figure S8d).

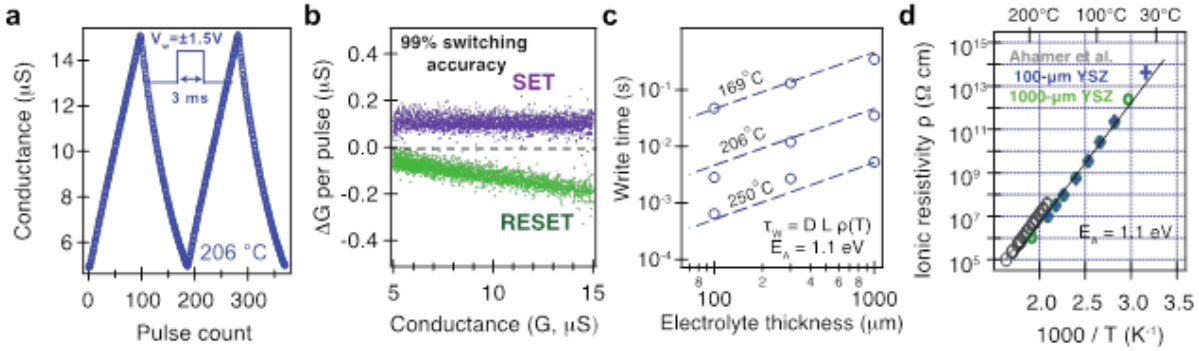


Figure 3: Switching behavior in model bulk-RRAM devices fabricated on single-crystal YSZ. (a) Analogue switching of a device fabricated on a 100-μm YSZ electrolyte substrate shows linear and symmetric switching. (b) The switching distribution shows an even higher switching accuracy than the thin-film YSZ devices. Over 99% of the SET and RESET pulses switch in the correct direction. (c) The switching time needed to achieve 100 analogue states within a 3× change in conductance for different electrolyte thickness and temperatures. All results are fitted using a common fit parameter D = 80 ± 20 nF cm^{-2} (95% confidence interval). (d) Ionic resistivity of the YSZ electrolyte as a function of the temperature show an activation energy ~1.1 eV, comparable to high-temperature results from Ahamer et al[53]. Representative Nyquist impedance plots are shown in Supporting Figure S10.

This article is protected by copyright. All rights reserved.

Figure 3a shows the analogue switching behavior of a device with a 100-µm-thick electrolyte. The TiO_{2-X} layers were more chemically reduced during fabrication to enable the bulk-RRAM cell to operate at different designed conductance levels. The switching accuracy is 99% (Figure 3b). The conductance change is linearly proportional to the write voltage and the pulse time (Supporting Figure S9a,b); this differs from the exponential (voltage) and logarithmic (time) relationship observed in metal-oxide-ECRAM[54], which were not heated, and suggests fundamentally different switching mechanisms. An even higher density of analogue conductance states is obtained by reducing the write pulse time (Supporting Figure S9c-f) without loss of accuracy.

In Figure 3c we plot the write time as a function of electrolyte thickness and temperature. The results are fitted to the simple model $\tau_W = DL\rho(T)$, where D = 80 ± 20 nF cm^{-2} (95% confidence interval) is the common fit parameter, $\rho(T)$ is the temperature-dependent ionic resistivity of the YSZ electrolyte plotted in Figure 3d, and L is the thickness of YSZ. Our results show that write time can be faster by decreasing L and by increasing T. The rate-limiting step is the resistance of the YSZ electrolyte, which is over 800 times thicker than the TiO_{2-X} layers. Reducing the electrolyte thickness from >100 µm to <1 µm can be used to decrease the series ionic resistance and the write time: as shown in Figure 1f, a 400-nm thick thin-film electrolyte result in devices ~2 µs write times. The excellent fit across all fabricated devices suggests minimal device-to-device variation.

Next, we consider the long-term information retention of the device. Like other types of RRAM, bulk-RRAM harnesses the reduced mobility of $V_O^{\cdot\cdot}$ at lower temperatures (Figure 3d) to "freeze" the information state. This contrasts strongly with other types of ECRAM cells, which also have three terminals but that need electronic switches to isolate the gate from the channel[40–48]. We plot the memory loss over time (**Figure 4**a) when shorting the base and switching layers after setting the device to a high-conductance state. The switching layer decay time strongly depends on

This article is protected by copyright. All rights reserved.

Author Manuscript

temperature: at room temperature, it decays less than 0.3% after one week. Devices in both high- and low-conductance states relax to equilibrium in a manner consistent with an RC circuit model with a similar time constant (Supporting Figure S11). The first-order exponential fit suggests that the relaxation time constant is not strongly dependent on the conductance state.

In Figure 4b we plot the retention time τ_R, defined as the time for the conductance to drop by 2%, as a function of temperature. The trendline fit is given by a simple RC circuit model $\tau_R = BL\rho(T)$, with B fitted to 1.4 ± 0.5 μF cm^{-2} (95% confidence interval), and relates to the chemical capacitance[55–57]. The thin-film devices from Figure 4b retain state longer than predicted, which suggests that $V_O^{\cdot\cdot}$ transport in the electrolyte is not rate limiting. Bulk-RRAM retain state for periods of days to weeks due to low $V_O^{\cdot\cdot}$ mobility at or near room temperature, sufficient for neuromorphic computing[1,3]. Materials with higher Arrhenius activation energy can be used to achieve longer retention (Figure 4c) by increasing the temperature dependence of the ionic resistivity. This is especially crucial to achieve sufficient retention time when other components in an integrated circuit need to operate above room temperature.

Author Manuscript

This article is protected by copyright. All rights reserved.

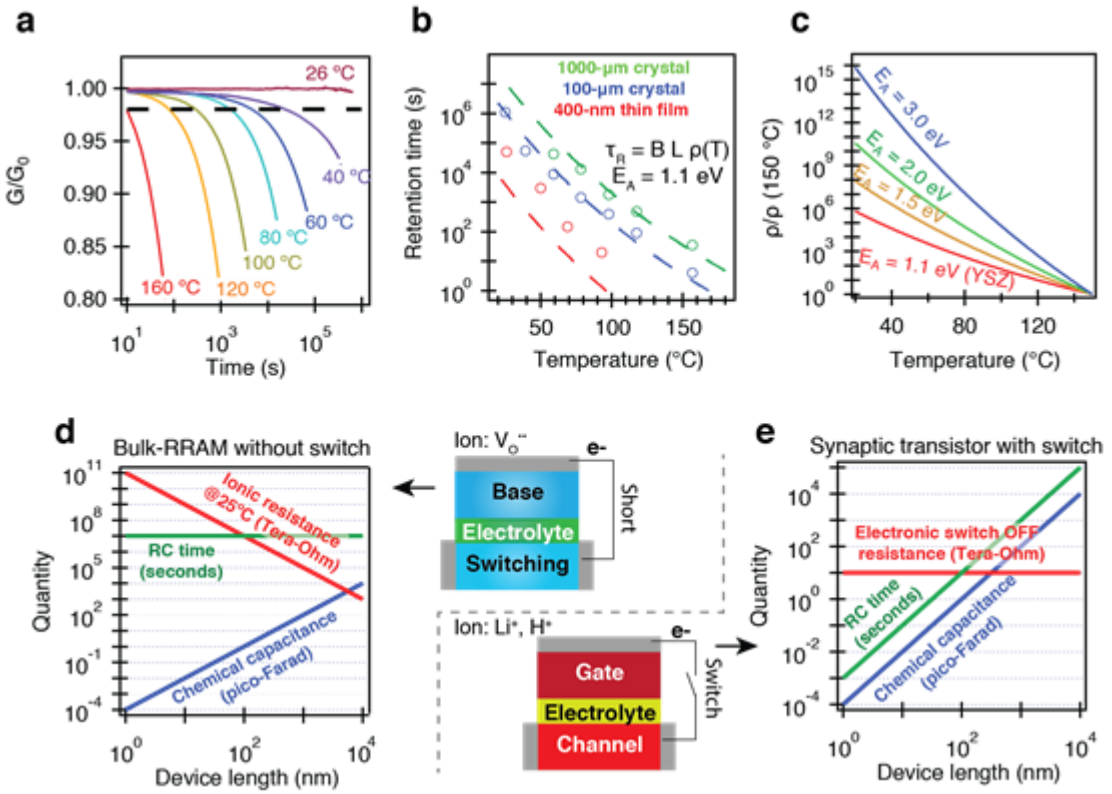


Figure 4: Retention of bulk-RRAM at various temperatures when the bulk and switching layers are shorted without an electronic switch. (a) Conductance decay over for a device constructed on 100-µm single-crystal YSZ. Retention time increases significantly at lower temperatures as the electrolyte resistance rises. (b) The retention time τ_R, defined as 2% change in conductance, as a function of temperature (T) and electrolyte thickness (L), with a single fit parameter B = 1.4 ± 0.5 µF cm^{-2} (95% confidence interval) fitted to all the single-crystal YSZ results (c) Computed ionic resistance normalized to the resistance at 150°C for different activation energies; the retention time is expected to be proportional to the ionic resistance. (d) RC time constants for bulk-RRAM do not depend on the device size because the ionic resistance increases at the same rate that the chemical capacitance decreases. Retention is achieved when the top and bottom contacts are shorted. We assume each layer is 100-nm thick, the specific chemical capacitance is 1000 F cm^{-3} (ref. [56,58]) and the ionic resistance is 10^{-14} Ωcm from Figure 3d. (e) The RC time decreases significantly with device size in ECRAM cells that use an electronic switch to isolate the top and bottom contacts. As chemical capacitance decreases with size, the switch's resistance does not decrease and assumed to be 10^{13} Ω here, resulting in drastic decrease in τ_R.

We now consider how bulk-RRAM would retain state when scaled to smaller lateral dimensions. We assume that the retention time is proportional to an RC time constant (τ_{RC}), where C is the chemical capacitance[55] of the switching layer while R is the total ionic resistance of the

This article is protected by copyright. All rights reserved.

device. A size-independent τ_{RC} is obtained because the chemical capacitance decreases proportionally with the area while the ion resistance increases at the same rate (Figure 4d). A τ_{RC} of 10^7 s, ~ 4 months, is computed when each layer in the stack is 100-nm thick.

Whereas bulk-RRAM cells retain state by immobilizing $V_O^{\cdot\cdot}$ and blocking ion migration, ECRAM cells that operate at constant temperature instead use electronic switches to block electron migration between the gate and channel[40–49]. This method does not scale well for smaller devices: R_{OFF} is constant and depends on the properties of the switch while C decreases proportionally with the area (Figure 4e). Assuming that R_{OFF} ~10^{13} Ω (ref. [59]), this scheme provides long τ_{RC} for relatively large areal devices (>10 μm)^2, but τ_{RC} decreases drastically for smaller devices, yielding only ~10 s for scaled (100 nm)^2 devices, many orders of magnitude lower than for a similarly-sized bulk-RRAM device (Figure 4d). A recent scaled proton-based ECRAM shows retention time of ~ 5 sec[49]. We note that the bulk-RRAM cells can also utilize an electronic switch for short-term information retention when the devices are heated at its elevated write temperature (Figure S12a,b). Unlike ECRAM based on Li+ and H+, bulk-RRAM also possess long-term information storage mechanisms by immobilizing oxygen vacancies using temperature.

The significant improvements in stochasticity of non-filamentary bulk-RRAM ultimately arise from using a solid electrolyte to eliminate the dominant conducting filament which arises from localized internal joule heating. Instead, the temperature of bulk-RRAM is uniform and controlled externally, so the defects are homogeneously distributed in the bulk as a solid solution. Without a dominant filament, the statistical ensemble behavior of all defects controls the resistance and analogue information state. This results in deterministic, predictable, and linear behavior for the bulk-RRAM devices, essential for analogue neural network applications, and compensates for the lower information density and higher complexity of using three terminals devices. While we

This article is protected by copyright. All rights reserved.

Author Manuscript

anticipate higher stochasticity for smaller devices with fewer defect sites, bulk-RRAM should be more deterministic than filamentary-RRAM because the switching layer's volume is always larger than the filament's volume, and stochasticity usually scales with the square root of the number of defects based on probability theory[33]. The switching layer's geometry can also be designed through lithography, whereas the size of the filament is sensitive to the material's mass, heat, and electron transport properties, and likely cannot exceed 10 nm. We hypothesize that bulk storage also makes these devices less sensitive to impurities and variations in oxide thickness, and not require extensive process control as in filamentary memristors[11]. Our ability to achieve deterministic switching is not specific to the properties of the materials used here, and is generalizable to many mixed-conducting transition metal oxides[38], opening up new opportunities for materials research.

Interfacial memristors[18,29] that switch by modulating the defects near the interface should be less stochastic than filamentary devices because of the higher number of defects at a two-dimensional interface. However, bulk-RRAM has even more defect sites by using the three-dimensional bulk. Furthermore, although interfacial[18,29] and three-dimensional memristors[60,61] without a solid electrolyte may not contain filaments, they rely upon internal joule heating[29], which leads to nonlinear and asymmetric switching behavior because the temperature and defect mobility would depend strongly on the present resistance state.

Another class of analogue memory with significant recent research is ECRAM [40–48] based on bulk transport of Li^+ or H^+ defects. Our statistical ensemble analysis also explains why these ECRAM switch more linearly and deterministically than filamentary devices. Bulk-RRAM is a type of ECRAM that uses oxygen vacancies, and has significantly better retention for smaller devices by immobilizing defects using temperature, as opposed to using switches to electronically isolate the gate and channel in past ECRAM cells (Figure 4d,e). Moreover, there are wide numbers of CMOS-compatible

This article is protected by copyright. All rights reserved.

Author Manuscript

materials that conduct oxygen vacancies at elevated temperatures, including many transition metal oxides. ECRAM that switch at room temperature must utilize more mobile defect ions like Li^+ in metal oxides[41,44,46], H^+ in electroactive polymers[42,45], or O^{2-} in ionic liquids[40,48], which often necessitates CMOS-incompatible materials.

Finally, we consider the energy of reading and switching in an array. Read energies are low when bulk-RRAM is engineered to have high electronic resistances; our devices already achieved several megaohm (Figure 1f) and can be further improved to tens to hundreds of megaohm using "square" geometries. High read resistances not only minimize the read current but also enables larger (>10^6 devices), more energy-efficient crossbar arrays that conduct ~10^{14} inference operations per joule[4,8,34]. For training accelerators, it is necessary to consider the energy costs for switching. The direct electrical energy cost of writing is very low, on the order of 10^{-17} J for a $(100\text{-nm})^3$ scaled device (see supporting information). To account for the thermal energy needed to heat the chip to ~150°C, the energy cost per switching event is estimated to be ~10^{-14} J for scaled devices that switch in parallel (see Supporting information and Figure S12a), compared to ~10^{-12} for filamentary memristors[22]. When weight updates are sporadic, such as for inference, a simpler selector-free configuration can also be constructed (Supporting Figure S12c). We propose a device density of $8F^2$, where F is the feature size (Supporting Figure S12d).

Bulk-RRAM solves major challenges of filamentary-RRAM for analogue computing. By eliminating the dominant filament and instead storing information through the bulk $V_O^{\cdot\cdot}$ concentration, bulk-RRAM switches deterministically rather than stochastically by harnessing the defects' statistical ensemble behavior. The bulk-RRAM provides a generalizable approach towards enabling the predictable analogue memory element for enable energy-efficient neuromorphic

Author Manuscript

This article is protected by copyright. All rights reserved.

computing, and inspires the use of bulk point defects to control the stochasticity of nanoelectronic systems.

Experimental Section: The devices containing thin-film YSZ were fabricated by using optical photolithography and electron beam lithography to define the bottom Ti/Pt contacts on a Si/SiO_2 substrate, then sputtering consecutive layers of TiO_2 switching layer (DC-reactive), YSZ electrolyte (RF), TiO_{2-X} base layer (DC-reactive), and TiO_X (DC-reactive) top contact using shadow masks (Supporting Figure S2). All sputtering was conducted at room temperature. The bottom TiO_2 and YSZ layers were crystallized by annealing at 700°C (Supporting Figure S3), previously shown to be sufficient to crystallize sputtered thin-film YSZ[62] and TiO_2[46], but not high enough to support significant cation intermixing at the interface (Supporting Figures S4-5). The TiO_{2-X} base layer and top contact was not annealed. Energy dispersive spectroscopy suggests that the Y:Zr ratio is 12% ± 3% (Supporting Figure 4f), slightly lower than expected based on the target composition, but with sufficient defects to support oxygen vacancy conduction[63]. The model devices containing single-crystal YSZ were fabricated by sputtering TiO_2, annealing at 600°C to form anatase, and evaporating Pt contacts on opposites side of a single-crystal YSZ substrate. Oxygen vacancies were introduced into TiO_2 by reducing in a 2 bar H_2 environment at 400°C for 2 hr. Device testing was conducted using a Bio-logic SP-300 bipotentiostat or a National Instruments Data Acquisition Device (DAQ-6358) system controlled by the LabVIEW program. More details on device fabrication, measurements, physical modeling, and atomic force microscopy calculations can be found in the supporting information.

Supporting Information is available from the Wiley Online Library or from the author

Author Manuscript

This article is protected by copyright. All rights reserved.

Conflicts of Interest: The authors declare no financial conflicts of interest.

Acknowledgements: The work at Sandia National Laboratories was supported by the Laboratory-Directed Research and Development (LDRD) Programs, including the Harry Truman Fellowship Program (Y. Li). Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. The authors thank W. York, P. Wijeratne, M. Bonner, and S. Agarwal at Sandia National Laboratories for assistance with focused ion beam cross-section, X-ray diffraction, microfabrication, and Crossbar simulations; C. Baumer and D. Mueller at Forschungszentrum Juelich for discussions on oxygen vacancy transport; S. Ambrogio at IBM research for discussions on stochastic switching; K. Lim, D. Chen, and H. Li at Stanford University for diffraction analysis and electrochemical measurements; and J. Gold at Nanolabs (a Eurofins Company) for assistance with X-ray photoelectron spectroscopy measurements, analysis, and general discussion.

Received: ((will be filled in by the editorial staff))
Revised: ((will be filled in by the editorial staff))
Published online: ((will be filled in by the editorial staff))

References

This article is protected by copyright. All rights reserved.

Author Manuscript

[1] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. Le Gallo, K. Moon, J. Woo, H. Hwang, Y. Leblebici, G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. Le Gallo, K. Moon, J. Woo, H. Hwang, *Adv. Phys. X* **2017**, *2*, 89.

[2] D. Ielmini, H.-S. P. Wong, *Nat. Electron.* **2018**, *1*, 333.

[3] Q. Xia, J. J. Yang, *Nat. Mater.* **2019**, *18*, 309.

[4] M. J. Marinella, S. Agarwal, A. Hsia, R. Jacobs-gedrim, J. Niroula, S. J. Plimpton, E. I., C. D. J., *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2017**, *8*, 86.

[5] M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, Q. Xia, J. P. Strachan, *Adv. Mater.* **2018**, *30*, 1705914.

[6] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Davila, C. E. Graves, Z. Li, J. P. Strachan, P. Lin, Z. Wang, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, Q. Xia, *Nat. Electron.* **2018**, *1*, 52.

[7] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, W. D. Lu, *Nat. Electron.* **2019**, *2*, 290.

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

[9] P. Lin, C. Li, Z. Wang, Y. Li, H. Jiang, W. Song, M. Rao, Y. Zhuo, N. K. Upadhyay, M. Barnell, Q. Wu, J. J. Yang, Q. Xia, *Nat. Electron.* **2020**, *3*, 225.

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

[11] M. Lübben, F. Cüppers, J. Mohr, M. Von Witzleben, U. Breuer, R. Waser, C. Neumann, I. Valov, *Sci. Adv.* **2020**, *6*, eaaz9079.

[12] J. H. Yoon, J. Zhang, P. Lin, N. Upadhyay, P. Yan, Y. Liu, Q. Xia, J. J. Yang, *Adv. Mater.* **2020**, *32*, 1.

[13] S. Oh, T. Kim, M. Kwak, J. Song, J. Woo, *IEEE Electron Device Lett.* **2017**, *38*, 732.

[14] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. Nolfo, S. Sidler, M. Giordano, M. Bodini, C. P. F. Nathan, B. Killeen, C. Cheng, Y. Jaoudi, G. W. Burr, *Nature* **2018**, *558*, 60.

[15] K. Ding, J. Wang, Y. Zhou, H. Tian, L. L. Lu, R. Mazzarello, C. Jia, W. Zhang, F. Rao, E. Ma, *Science.* **2019**, *366*, 210.

[16] H. Yeon, P. Lin, C. Choi, S. Tan, Y. Park, D. Lee, J. Lee, F. Xu, B. Gao, H. Wu, H. Qian, Y. Nie, S. Kim, J. Kim, *Nat. Nanotechnol.* **2020**, DOI 10.1038/s41565-020-0694-5.

[17] R. Waser, M. Aono, *Nat. Mater.* **2007**, *6*, 833.

[18] A. Sawa, *Mater. Today* **2008**, *11*, 28.

This article is protected by copyright. All rights reserved.

Author Manuscript

[19] R. Waser, R. Dittmann, C. Staikov, K. Szot, *Adv. Mater.* **2009**, *21*, 2632.

[20] D. Kwon, K. M. Kim, J. H. Jang, J. M. Jeon, M. H. Lee, G. H. Kim, X. Li, G. Park, B. Lee, S. Han, M. Kim, C. S. Hwang, *Nat. Nanotechnol.* **2010**, *5*, 148.

[21] J. P. Strachan, M. D. Pickett, J. J. Yang, S. Aloni, A. L. David Kilcoyne, G. Medeiros-Ribeiro, R. Stanley Williams, *Adv. Mater.* **2010**, *22*, 3573.

[22] J. J. Yang, D. B. Strukov, D. R. Stewart, *Nat. Nanotechnol.* **2013**, *8*, 13.

[23] M. Prezioso, B. D. Hoskins, G. C. Adam, K. K. Likharev, D. B. Strukov, *Nature* **2015**, *521*, 61.

[24] S. Kumar, J. P. Strachan, R. S. Williams, *Nature* **2017**, *548*, 318.

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

[26] J. Lee, W. D. Lu, *Adv. Mater.* **2018**, *30*, 1702770.

[27] S. Kim, S. Choi, W. Lu, *ACS Nano* **2014**, *8*, 2369.

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

[29] S. Menzel, M. Waters, A. Marchewka, U. Böttger, R. Dittmann, R. Waser, *Adv. Funct. Mater.* **2011**, *21*, 4487.

[30] D. Ielmini, F. Nardi, C. Cagli, *Appl. Phys. Lett.* **2010**, *96*, 053503.

[31] S. Yu, X. Guan, H. S. P. Wong, *Tech. Dig. - Int. Electron Devices Meet. IEDM* **2011**, 17.3.1.

[32] S. Ambrogio, S. Balatti, V. Mccaffrey, D. C. Wang, D. Ielmini, *IEEE Trans. Electron Devices* **2015**, *62*, 3812.

[33] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, D. Ielmini, *IEEE Trans. Electron Devices* **2014**, *61*, 2912.

[34] T. Gokmen, Y. Vlasov, *Front. Neurosci.* **2016**, *10*, 333.

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

[36] S. Agarwal, S. J. Plimpton, D. R. Hughart, A. H. Hsia, I. Richter, J. A. Cox, C. D. James, M. J. Marinella, in *Int. Jt. Conf. Neural Networks*, **2016**, pp. 929–938.

[37] S. H. Jo, K. Kim, W. Lu, *Nano Lett.* **2009**, *9*, 496.

[38] Y. Li, W. C. Chueh, *Annu. Rev. Mater. Res.* **2018**, *48*, 137.

[39] R. A. De Souza, *Adv. Funct. Mater.* **2015**, *25*, 6326.

This article is protected by copyright. All rights reserved.

Author Manuscript

[40] J. Shi, S. D. Ha, Y. Zhou, F. Schoofs, S. Ramanathan, *Nat. Commun.* **2013**, *4*, 2676.

[41] E. J. Fuller, F. El Gabaly, F. Léonard, S. Agarwal, S. J. Plimpton, R. B. Jacobs-Gedrim, C. D. James, M. J. Marinella, A. A. Talin, *Adv. Mater.* **2017**, *29*, 1604310.

[42] Y. van de Burgt, E. Lubberman, E. J. Fuller, S. T. Keene, G. C. Faria, S. Agarwal, M. J. Marinella, A. Alec Talin, A. Salleo, *Nat. Mater.* **2017**, *16*, 414.

[43] J. Tang, D. Bishop, S. Kim, M. Copel, T. Gokmen, T. Todorov, S. Shin, K. Lee, P. Solomon, K. Chan, W. Haensch, J. Rozen, in *2018 Int. Electron Devices Meet.*, IEEE, San Francisco, **2018**, pp. 18:292-18:295.

[44] M. T. Sharbati, Y. Du, J. Torres, N. D. Ardolino, M. Yun, F. Xiong, *Adv. Mater.* **2018**, *30*, 1802353.

[45] 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.

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

[47] K. Moon, S. Lim, J. Park, C. Sung, S. Oh, J. Woo, J. Lee, H. Hwang, *Faraday Discuss.* **2019**, *213*, 421.

[48] H. Huang, C. Ge, Q. Zhang, C. Liu, J. Du, J. Li, C. Wang, L. Gu, G. Yang, K. Jin, *Adv. Funct. Mater.* **2019**, *29*, 1902702.

[49] A. Melianas, T. J. Quill, G. LeCroy, Y. Tuchman, H. v. Loo, S. T. Keene, A. Giovannitti, H. R. Lee, I. P. Maria, I. McCulloch, A. Salleo, *Sci. Adv.* **2020**, *6*, eabb2958.

[50] K. Szot, W. Speier, G. Bihlmayer, R. Waser, *Nat. Mater.* **2006**, *5*, 312.

[51] A. J. Bard, L. R. Faulkner, *Electrochemical Methods: Fundamentals and Applications*, John Wiley And Sons, **2001**.

[52] B. J. Morgan, G. W. Watson, *Phys. Rev. B* **2009**, *80*, 233102.

[53] C. Ahamer, A. K. Opitz, G. M. Rupp, J. Fleig, *J. Electrochem. Soc.* **2017**, *164*, F790.

[54] S. Kim, T. Todorov, M. Onen, T. Gokmen, D. Bishop, P. Solomon, K.-T. Lee, M. Copel, D. B. Farmer, J. A. Ott, T. Ando, H. Miyazoe, V. Narayanan, J. Rozen, in *2019 Int. Electron Devices Meet.*, **2019**, pp. 847–850.

[55] J. Jamnik, J. Maier, *Phys. Chem. Chem. Phys.* **2001**, *3*, 1668.

[56] D. Chen, S. R. Bishop, H. L. Tuller, *Adv. Funct. Mater.* **2013**, *23*, 2168.

[57] W. C. Chueh, S. M. Haile, *Phys. Chem. Chem. Phys.* **2009**, *11*, 8144.

[58] E. J. Fuller, Y. Li, C. Bennet, S. T. Keene, A. Melianas, S. Agrawal, M. J. Marinella, A. Salleo, A. A. Talin, *IBM J. Res. Dev.* **2019**, *63*, 9.

This article is protected by copyright. All rights reserved.

Author Manuscript

[59] C.-H. Jan, F. Al-amoody, H.-Y. Chang, T. Chang, Y.-W. Chen, N. Dias, W. Hafez, D. Ingerly, M. Jang, E. Karl, S. K.-Y. Shi, K. Komeyli, H. Kilambi, A. Kumar, K. Byon, C.-G. Lee, J. Lee, T. Leo, P.-C. Liu, N. Nidhi, R. Olac-vaw, C. Petersburg, K. Phoa, C. Prasad, C. Quincy, R. Ramaswamy, T. Rana, L. Rockford, A. Subramaniam, C. Tsai, P. Vandervoorn, L. Yang, A. Zainuddin, P. Bai, in *2015 Symp. VLSI Technol. (VLSI Technol.*, **2015**, pp. T12–T13.

[60] J. C. Gonzalez-Rosillo, R. Ortega-Hernandez, B. Arndt, M. Coll, R. Dittmann, X. Obradors, J. Suñe, T. Puig, *Adv. Electron. Mater.* **2019**, *5*, 1800629.

[61] E. Herrmann, A. Rush, T. Bailey, R. Jha, *IEEE Electron Device Lett.* **2018**, *39*, 500.

[62] A. F. Jankowski, J. P. Haye, *J. Vac. Sci. Technol. A* **1995**, *13*, 658.

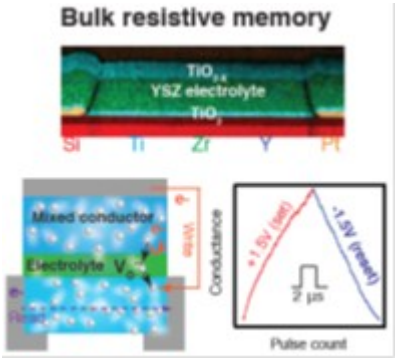
[63] R. Krishnamurthy, Y.-G. Yoon, D. J. Srolovitz, R. Car, *J. Am. Ceram. Soc.* **2005**, *87*, 1821.

Author Manuscript

This article is protected by copyright. All rights reserved.

TOC Text:

This work presents a resistive memory cell based on the electrochemical migration of oxygen vacancies for in-memory neuromorphic computing. By using the average behavior of all oxygen vacancies to store analogue information states, this cell overcomes stochastic and unpredictable switching plaguing filament-forming memristors, and instead achieves linear, predictable, and deterministic switching.



Author Manuscript

This article is protected by copyright. All rights reserved.