CMOS-Compatible Electrochemical Synaptic Transistor Arrays for Deep Learning Accelerator

Jinsong Cui[1], Fufei An[1], Jiangchao Qian[1], Yuxuan Wu[1], Luke L. Sloan[1], Saran Pidaparthy[1], Jian-Min Zuo[1,2], and Qing Cao[1-5*]

[1]Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [2]Materials Research Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [3]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [4]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [5]Holonyak Micro and Nanotechnology Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, USA

**e-mail: qingcao2@illinois.edu*

Abstract: In-memory-computing architectures based on memristive crossbar arrays could offer higher computing efficiency than traditional hardware in deep learning. However, the core memory devices must be capable of performing high-speed and symmetric analogue programming with small variability. They should also be compatible with silicon technology and scalable to nanometre-size footprints. Here we report an electrochemical synaptic transistor that operates by shuffling protons between hydrogenated tungsten oxide channel and gate through a zirconium dioxide protonic electrolyte. These devices are compatible with complementary metal–oxide–semiconductor (CMOS) technology and can be scaled to lateral dimensions of 150×150 nm^2. They can be monolithically integrated with silicon transistors to form pseudo-crossbar arrays as chip-area and energy efficient deep learning accelerators in which parallel and symmetric programming of the channel conductance with small cycle-to-cycle variations are achieved via gate-voltage pulse. They can be programmed at frequencies approaching megahertz and offer endurance over 100 million read-write cycles.

Main

Advances in deep learning have allowed computers to outperform humans in a range of complicated tasks. These developments have though largely been achieved by increasing the depth and size of the neural network models used, which has exponentially increased the computational load and energy consumption needed to train and execute them[1]. In-memory-computing architectures based on crossbar arrays of non-volatile memory are of potential use as efficient deep learning accelerators[2]. They offer compatibility with the back-end-of-line (BEOL) integration with silicon logic circuits and excellent scalability, providing on-chip data storage with density and capacity high enough to eliminate off-chip memory access during data intensive computations, and capability to perform analogue matrix operations in parallel without incurring data movement. They could thus potentially offer higher computational efficiency and lower chip-area cost than microprocessors based on von Neumann architectures[3].

To successfully train deep neural networks using memristive arrays, the weight movement in the training algorithms needs to be accurately mapped into the characteristics of the memory hardware. The device conductance change used to represent the weight modulation should be symmetric for positive and negative stimulus, and be precise with small cycle-to-cycle variability. If not, the training accuracy will be dramatically affected[4-6]. This requirement, however, remains a challenge as in most memristive devices — including those based on phase-change materials (PCMs)[7], filament-forming metal oxides[8-10], and ferroelectric oxides[11] — conductance increase (potentiation) and decrease (depression) characteristics are highly asymmetric with large variability due to the underlying material and physical device mechanisms[3-5].

Metal-oxide memristors, for example, typically have an abrupt potentiation that is due to positive feedback between the filament growth and the electric field, and a large variability as the filament-formation process is intrinsically stochastic. Such non-ideal asymmetric programming can be compensated for by using advanced training algorithms, but it typically requires doubling

the chip area, and penalties in latency and energy consumption[12]. Recently, memristors with improved symmetry have been developed by employing complex switching media[13-16], but they still suffer from large variability, and many of them are not compatible with direct BEOL integration[15,16].

Electrochemical synaptic transistors were first used to construct artificial neural networks in 1960s[17], and have recently resurfaced as a candidate to realize deep learning accelerators based on in-memory-computing architectures[18]. As initially designed for neuromorphic computing rather than digital information storage, their channel conductance can be precisely modulated by the electrochemical intercalation reactions controlled by a bias applied to separate gate terminals. They thus provide multistate analogue programming with high symmetry and low variability compared to two-terminal memristive devices. However, current electrochemical random-access memory (ECRAM) demonstrations are limited by poor compatibility with silicon complementary metal–oxide–semiconductor (CMOS) technology, which must be monolithically integrated with the memristive arrays as part of the accelerator to provide peripheral functions. The electrolytes used in most ECRAM devices are either liquid or organic polymers[19-21]. Despite their good performance, it is difficult to incorporate these devices in circuits as scaled memory cells with the long-term stability and reliability required for electronic applications, and they are often only functional in a controlled environment. All-solid-state inorganic ECRAM designs have been developed. However, some use PdH_x, which suffer from their chemical instability towards air and moisture, as a protonic intercalant reservoir[22,23]. Many others use intercalation of lithium ions[24,25], which is not compatible with silicon-CMOS as lithium solid electrolytes are generally air-sensitive and lithium ions can readily diffuse across various oxides into the silicon lattice. These lithium ions change the silicon charge carrier concentration as shallow interstitial dopants. The others adopt oxygen ions as the intercalant[26-28], whose sluggish motion limits the device programming speed typically to above several milliseconds[29,30].

In this Article, we report an all-solid-state inorganic ECRAM that operates through the reversible insertion of protons into a H_xWO_3 channel from hydrogenated ZrO_2 electrolyte and

H_xWO_3 gate. These devices exhibit highly symmetric programming under gate-voltage pulses, low cycle-to-cycle and device-to-device spatiotemporal variability, reliable read-write operations above 100 million cycles, and energy consumption below femto-joule per transaction. Their channel conductance can be tuned over a wide range, from nano to micro-siemens, enabling the construction of large-size arrays with optimised power and performance[3]. The small radius of the proton intercalant leads to fast ionic dynamics for high-speed write-read programming with frequencies approaching one megahertz. WO_3 and ZrO_2 are compatible with silicon-CMOS technology and associated wafer-scale microfabrication techniques, which allows the ECRAM dimensions to be scaled down to 150×150 nm^2. With amorphous HfO_2 as both the interlayer dielectric and the proton-diffusion barrier, the ECRAMs can be fabricated above silicon circuits without affecting the performance of the underlying logic transistors. They are functional in harsh conditions such as high vacuum and elevated temperature, and their non-volatile conductance exhibits low drift and long retention. To demonstrate parallel operation and monolithic integration, we fabricate ECRAM pseudo-crossbar arrays addressed with integrated silicon selector transistors. These ECRAM-CMOS hybrid in-memory-computing accelerators could achieve training accuracy comparable to that of digital accelerators based on static random-access memory (SRAM), but under drastically reduced chip-area cost and energy consumption.

Device architecture and operations

A device schematic is shown in Fig. 1a, with its gate stack revealed in a cross-sectional scanning transmission electron microscopy (STEM) micrograph (Fig. 1b). The detailed fabrication process is described in Methods and Extended Data Fig. 1. We choose the stoichiometric amorphous WO_3 deposited by the reactive sputtering from WO_3 target as the ECRAM channel as it gives both low base conductance and long proton retention (See Supplementary Note 1). Compared to VO_2[31,32], the intercalation of protons into amorphous WO_3 does not involve the phase transformation of the crystal structure[33], ensuring faster ECRAM operations with better device endurance. A hydrogen-spillover process is then utilized to

introduce precisely controlled concentration of protons into the WO_3 on wafer scale (See Supplementary Note 2), followed by the deposition of an ultrathin film of ZrO_2 by atomic layer deposition (ALD) as the solid-state protonic gate electrolyte. The secondary-ion mass spectrometry (SIMS) depth profile indicates that substantial amount of protons are diffused into the ZrO_2 from the H-doped WO_3 channel during ALD (Fig. 1c), which help to passivate the surface and grain boundaries of ZrO_2 with hydroxyl groups, forming pathways for fast protonic conduction[34]. The gate is a H_xWO_3 and metal bilayer, where the H-rich H_xWO_3 serves as the proton reservoir and helps to minimize the device built-in potential (See Supplementary Note 3)[21]. Device channel conductance G can be programed through 64 (6-bit) discrete states with high-level of symmetry, small cycle-to-cycle variability, and a large dynamic range above 20 under voltage pulses (amplitude V_G=±4V, Fig. 1d), in contrary to those built with elemental metal gate which can only operate symmetrically under current pulses (Extended Data Fig. 2). The zoom-in view shows three continuous conductance states out of the total 64 and confirms that the CMOS-compatible protonic ECRAM prototype has low read noise with conductance down to nanosiemens regime (Fig. 1e). The gate current during write is merely around a few pico-amperes (Fig. 1f), corresponding to a low switching energy of about 10 pico-joule per step, which can be further reduced to below femto-joule with the adoption of up to 10^6 times faster write pulses down to microseconds. By adjusting the initial hydrogen concentration in the H_xWO_3 channel through varying parameters in the hydrogen spillover (See Supplementary Note 2), the device base conductance G_0 and the absolute conductance change ΔG per weight-update step can be effectively modulated. Fig. 1g illustrates the programing of a ECRAM built on a more hydrogen-rich H_xWO_3 channel, whose G_0 is increased from 1 to 200 nS with the average ΔG per step simultaneously increased by ~ 10 times, likely assisted by defect-defect interactions and the increase of proton concentration in the hydrogenated ZrO_2 electrolyte[20]. This tunability is critical for large-scale array implementations, where the memristive resistance needs to be optimized to minimize the thermal noise, the voltage drop in the interconnect metal lines, and the overhead on the peripheral circuits[3], and balance the trade-off between the absolute ΔG per step

and the on/off ratio. The channel conductance modulation also depends exponentially on the amplitude of the gate-voltage pulses, as higher gate bias increases the gate-current density and thus the total amount of injected protons per pulse (Fig. 1h). These ECRAM cells demonstrate good retention. The drift coefficient v, obtained by fitting the change of channel conductance as a function of time t to the power law of $G(t)=G_{t0}(t/t_0)^{-v}$ where G_{t0} is the initial conductance at time t_0 and v is the fitted drift coefficient, is ultra-low (Fig. 1i), especially compared to PCMs[35]. It shows that the slow self-discharging of the ECRAM under zero gate bias ensures that these discrete conductance levels are well preserved. States deviating further from G_0 exhibits a slight increase of the drift coefficient, likely caused by the larger gradient for proton diffusion. Such low drift coefficient over at least 10^3 seconds is more than enough to ensure that the accuracy will not be affected when ECRAM-based accelerators are used to train multilayer perceptrons and even convolutional neural networks with weight sharing in convolutional layers, where weights stored on ECRAM cells are continuously updated during back propagation[36]. The device retention is further improved if measured with zero gate current, which is consistent with the proven capability of electrochromic windows built on H_xWO_3 to retain their coloration density for years without external power supply[37].

Device operating speed, scalability, and endurance

High-speed programing of protonic ECRAMs was then demonstrated with fast gate-voltage pulses. With G_0 around 1 μS, a large ΔG of 3 μS can be induced by 32 potentiation-depression pulses with width down to 300 microseconds to afford an on/off ratio of 3 (Fig. 2a). Fig. 2b shows the reproducible cycling with 10 microsecond and 5 microsecond write pulses. Although the device dynamic range becomes smaller, the programming characteristics are still highly symmetric, and the ΔG after each pulse remains larger than 1 nS with low variability and read noise (Extended Data Fig. 3) to enable the storage of 8-bit of information within a single ECRAM cell. The effect of the off-state current can be eliminated with the addition of a dummy column in the crossbar architecture[38]. Compared to state-of-the-art ECRAMs employing oxygen ions as the intercalant[28], when operated under the 10 microsecond-wide write pulses to achieve

the same conductance-modulation ratio, our protonic ECRAMs require about 7 times lower number of weight-update steps under a ~40% lower gate-voltage amplitude, resulting from proton's higher ionic diffusivity. There is some upshift of the baseline G_0, indicating that more weight-update operations are required in the depression branch to bring the conductance back to its initial level. However, such asymmetry associated with the slightly different number of the intermediate states between the minimum and maximum conductance states will not degrade the training accuracy as predicted in simulation[5], and can be compensated by applying voltage pulses with slightly different amplitude or width for the potentiation *versus* depression operations. Since the conductance change correlates with the amount of charge, i.e., the number of protons, injected or extracted, ΔG, as well as the corresponding energy consumption per weight-update step, scales linearly with the pulse width (Fig. 2c).

In addition to the time required for the weight update, the ECRAM device speed is also limited by the read transients (See Extended Data Fig. 4 for the measurement setup). For read without write, the ECRAM channel behaves as a resistor capacitively coupled to the device capacitance. We monitored the current flowing across the source-drain electrodes, showing that the settling time t_{read} after the application of the read-voltage pulse was less than 500 nanoseconds as limited by the resistor-capacitor (*RC*) delay (Fig. 2d). For read post update, we observed a fast recovery current pulse followed by a slower decay eventually leading toward a stable sense current (Fig. 2e). This transient behavior is qualitatively similar to that of the ECRAMs employing lithium or oxygen ions as the intercalants, and it is limited by the dielectric relaxation processes and the charge transfer from the center of the channel to contacts on the edge. But the quantitative settling time $t_{\text{read-after-write}}$ for protonic ECRAMs is >100 times faster with proton's high ionic diffusivity[29,30].

Both t_{read} and $t_{\text{read-after-write}}$ are expected to diminish with scaling, which reduces both the gate capacitance and the required lateral diffusion distance of injected protons. In CMOS-compatible, all-inorganic protonic ECRAMs with their channel lateral dimensions scaled down to 150×150 nm^2 (Fig. 2f), which represent the smallest lateral ECRAMs ever fabricated, their symmetric and

linear programing characteristics are well preserved (Fig. 2g), showing the excellent device scalability. The absolute conductance values and the dynamic range of modulation are slightly degraded likely due to the impact from the contact resistance (Extended Data Fig. 5), and the larger read noise could be caused by the miniaturization of the device volume together with the increase of the channel surface-to-volume ratio, which reduce the averaging effect and increase the impact from traps at material interfaces. Even though the gate capacitance is significantly reduced with $>10^3$ times smaller device channel area, the t_{read} of the scaled ECRAMs is only marginally improved to 370 nanoseconds (Fig. 2h), indicating that the device capacitance is mainly limited by parasitics from, for example, the overlap regions between the source/drain and gate electrodes with the hydrogenated ZrO_2 sandwiched in between. However, the rate of the current stabilization in the read post update shows much more significant improvement, with the apparent $t_{read\text{-}after\text{-}write}$ less than 1 microsecond (Fig. 2i), showing the benefit of ECRAM channel-length scaling to mitigate the device read transients associated with charge/ion transport. To improve the device latency down to the desired 50–100 nanosecond regime, it is required to further suppress the device read transients and increase the ΔG per step in weight update, which can be accomplished by further scaling down the lateral length of the H_xWO_3 channel, optimizing the device structure to minimize the parasitic capacitance, modifying the composition and nanostructures of the ZrO_2-based electrolyte to increase its ionic conductivity[34], and reducing the gate stack (H_xWO_3 channel and the electrolyte) thickness (Supplementary Note 4).

In addition to their fast operations and excellent scalability, the CMOS-compatible protonic ECRAMs exhibited excellent endurance when modulated in air with fast voltage pulses. No degradation in terms of programming symmetry, cycle-to-cycle variability, baseline channel conductance, or conductance modulation was observed after 10^8 write-read operations, which correspond to over one million switches over the full on/off range (Fig. 2j). The moderate dynamic range of ECRAM corresponds to limited modulation of the proton concentration in the H_xWO_3 channel, which ensures that the devices will not degrade due to repetitive lattice expansion or ion trapping[39]. There could be slightly oxygen loss accompanying the proton

insertion and removal in the amorphous H_xWO_3 channel, but it is insignificant even after million times of operations driven by the gate pulses (Extended Data Fig. 6), as measured within the accuracy of TEM-energy dispersive X-ray spectroscopy (EDS). The variation of the maximum and minimum conductance after 32 potentiation and depression pulses is small with a relative standard deviation around 7%, which indicates a standard deviation normalized to the entire weight range around $7\%/\sqrt{32}=1\%$ for each weight-update cycle. With the 20 nm thick HfO_2 or Al_2O_3 passivation deposited by ALD at 120 ºC, the device operation was not affected even when measured in high vacuum (10^{-6} Torr, Fig. 2k). It is a significant advantage over previous protonic ECRAMs employing polymer/nanoporous electrolytes and/or PdH_x gate, which are only functional in a humidity-controlled, utmost medium vacuum, environment or forming gas[21-23,40,41]. The stable and symmetric programming characteristics were maintained even after annealing at 200 ºC, and the device can be successfully operated at 80–100 ºC without degrading the data retention or device endurance (Extended Data Fig. 7), which further verifies the reliability of our all-inorganic protonic ECRAM prototype in harsh environment. Similar as other memristive devices[40,42], the device conductance and conductance modulation systematically shift with the increase of temperature due to thermal excitation and activation, which needs to be compensated on the software level with the chip temperature continuously monitored by built-in temperature sensors. Finally, the performance of the all-inorganic protonic ECRAM is benchmarked with other BEOL-compatible analogue memory technologies, and shows clear advantages in symmetry, cycle-to-cycle variability, and energy consumption (See Supplementary Note 5).

ECRAM pseudo-crossbar arrays

All-inorganic protonic ECRAMs were monolithically integrated with silicon metal-oxide-semiconductor field-effect transistors (MOSFETs) (See Methods and Extended Data Fig. 8 for the process flow), in the form of 1-transistor-1-ECRAM cells (Fig. 3a), where the silicon transistor acted as the selector to solve the write-disturbance problem (Fig. 3b, there is negligible ECRAM conductance change with the selector transistor turned off). Note that ECRAM arrays

can potentially operate without selectors using the half-selection bias scheme[27,28], and here we mainly use the fabrication and characterization of 1-transistor-1-ECRAM cells to verify the CMOS compatibility. With the ALD HfO_2 deposited on top of the completed silicon MOSFETs and the metal interconnects as both the interlayer dielectric and proton diffusion barrier, the performance of these underlying silicon devices was not affected by the addition of the ECRAM layer on top (Extended Data Fig. 9). Beyond individual cells, an integrated pseudo-crossbar array was constructed (Fig. 3c,d). In the pseudo-crossbar array fabric (Fig. 3e), parallel row-by-row weight updates were successfully implemented, as in the programming patterns shown in Fig. 3f. The average weight-update in the selected cells was 210±20%, while those in the unselected rows were not disturbed with an average ΔG of -1±4%. After parallel programing, the pseudo-crossbar arrays can be used to parallelly execute the vector-matrix multiplication, which is the core and the most expensive computing operation in many deep-learning and image-processing algorithms. We used the color transformation as an example, where a transformation matrix is used to generate modified red (*R*), green (*G*), and blue (*B*) pixel data for recoloring (Fig. 3g). Each element of the transformation matrix was first mapped to the ECRAM array. The color of each pixel represented by an 8-bit number for each *R/G/B* channel was then converted to a voltage input vector delivered to the bit lines of the fabric. The current measured at the source lines for the weight sum was the result of the dot product between the input voltage vector and the analogue conductance matrix following the Kirchhoff's law, and it was finally encoded back to generate the modified *RGB* values (Fig. 3h). Despite the non-idealities associated with the device variability and conductance drift, the color transformation result was comparable to what performed by the software with a narrow error band (Fig. 3i).

With the functionality of ECRAM pseudo-crossbar arrays verified in experiment, we then simulated the performance of accelerators built on ECRAM arrays integrated with silicon-CMOS peripheral circuits[43]. Based on the experimentally measured dynamic range, non-linearity, symmetry, read-write voltage and speed, cycle-to-cycle weight-update variation, and device-to-device variations (Extended Data Fig. 10) of CMOS-compatible, all-inorganic protonic

ECRAMs (See Supplementary Note 6), the modeling, which considers all the device non-idealities, indicates that such in-memory-computing accelerators can achieve similar level of accuracy (86-93%, the small degradation in accuracy for devices operated with faster pulses is caused by their smaller on/off ratio) in learning the classification of the Modified National Institute of Standards and Technology (MNIST) dataset compared to the SRAM-based digital accelerators or the software (94%), as enabled by ECRAMs' good linearity and symmetry, low variability, and their large number of available states (Fig. 3j). Meanwhile, the overall energy consumption, including contributions from the ECRAM array, the selector transistors, and the periphery circuit, is up to two times lower than SRAM-based accelerators (Fig. 3k), and the chip-area cost, with the analogue and high-resistance ECRAMs monolithically integrated on top of the silicon circuits, is reduced by over 10 times. The latency is 50 times longer due to the much slower operation of our current ECRAM prototypes compared to SRAM (Fig. 3l), but it is competitive against accelerators built on PCM and metal-oxide memristors (Supplementary Note 7)[43]. ECRAM accelerators can also be used to train the convolutional neural networks, achieving similar level of accuracy (~85%) as benchmarked against their SRAM-based counterparts or software (Fig. 3m), but superior to those employing other emerging non-volatile memories[44].

Conclusions

We have reported a CMOS-compatible, all-inorganic protonic ECRAM technology that offers highly symmetric switching characteristics, low cycle-to-cycle variability, high read-write pulse endurance, and good retention with low drift. Our approach uses a hydrogenated ZrO_2 as a protonic electrolyte to ensure CMOS-compatibility and provides high diffusivity of protons for fast device operation approaching megahertz. It adopts a symmetric gate stack composed of H_xWO_3 channel and gate to enable symmetric voltage programming. The smallest lateral ECRAMs have an active device area of $150\times150nm^2$, verifying excellent device scalability. We fabricated a pseudo-crossbar array composed of ECRAM synaptic memory cells monolithically integrated on top of silicon selector transistors, which can perform both parallel analogue programming and vector-matrix multiplication. The results illustrate the technological potential

of our electrochemical synaptic transistor arrays in energy- and cost-efficient in-memory computing accelerators for deep learning. The key remaining challenge is to further reduce the device speed down to 50–100 nanoseconds through both scaling and improving the solid-state protonic electrolyte.

Methods

Fabrication of CMOS-compatible all-inorganic protonic ECRAMs. The process flow is schematically illustrated in Extended Data Figure 1. 40 nm of HfO_2 was first deposited as a proton-diffusion barrier and etching-stop layer by ALD (Veeco Nanotech ALD System) on a silicon substrate covered with 90 nm thermal oxide, using Tetrakis(dimethylamido)-hafnium and water as precursors at 200 ºC. A photolithography step was then performed to define the source-drain electrode patterns into the photoresist (AZ 5214, exposure dose was 200mJ·cm^{-2}), and an electron-beam evaporation (Temescal) was used to deposit 2 nm Cr/ 40 nm Pt, followed by lift-off in acetone to form the source-drain contacts. 80 nm of WO_3 was grown by reactive radio-frequency magnetron sputtering (Kurt J. Lesker PVD 75) of WO_3 target (Kurt J. Lesker) by means of a plasma composed of argon as the carrier gas and oxygen as the reactive gas. After pumping down the sputtering-deposition chamber to high vacuum (5×10^{-7} Torr), pure oxygen and argon were introduced with their relative flow rates adjusted to the desired ratio with the help of the mass-flow controllers to a stable chamber pressure of 1.5 millitorr. During deposition, the substrate was held at 300 ºC and the radio-frequency forward input power was maintained at 140 W. After WO_3 deposition to the desired thickness of 50–100 nm, another photolithography step was performed to define the device channel footprint into the photoresist. A subsequent SF_6 reactive-ion etching (RIE, Oxford Mixed ICP-RIE system) removed the WO_3 in the unprotected area to stop on the HfO_2 or Pt metal surfaces. The inductively coupled plasma (ICP) power was 1,000 W, the RIE power was 100 W, the flow rate of SF_6 was 20 s.c.c.m., and the chamber pressure was maintained at 5 millitorr. A thin, i.e., 10 nm thick, Aluminum film was then blanketly deposited covering the whole substrate by sputtering (AJA Orion-8 Magnetron

Sputtering System). The deposited Al film was etched by soaking the substrate in 4 $mol \cdot L^{-1}$ HCl aqueous solution, and part of the hydrogen generated was incorporated into the WO_3 channel as proton intercalants through hydrogen spillover. After further adjusting the proton concentration in H_xWO_3 by annealing at 150 °C in air, the substrate was immediately transferred to an ALD chamber, where 15 nm of ZrO_2 was deposited using Tetrakis-(dimethylamino)-zirconium (Strem Chemicals) and water vapor as precursors at 120 °C. The low ALD deposition temperature and the adoption of water as precursor are critical to maintain the proton content within the ZrO_2 to enhance its ionic conductivity[45,46]. Photolithography was used again to pattern the gate electrode, which has overlap with the source-drain contacts, into the photoresist. 20 nm of WO_3 was deposited again by the radio-frequency reactive sputtering with the substrate held at room temperature, and then converted to H_xWO_3 by means of the same hydrogen spillover process as described above. Electron-beam evaporation was used to deposit 2 nm Cr/40 nm Au followed by lift-off of the photoresist in acetone to complete the gate electrode stack composed of H_xWO_3/Cr/Au trilayer. Another 20 nm of HfO_2 or Al_2O_3 was deposited by ALD as a passivation layer to improve the device stability and endurance. A final lithography step was performed to expose the probing pads for the source-drain electrodes, where the ZrO_2 and HfO_2/Al_2O_3 dielectrics on top were etched by HF to complete the device fabrication flow.

Fabrication of nanoscale ECRAMs. To fabricate ECRAMs with their device channel length and channel width scaled down to sub-micron regime, electron-beam lithography (EBL), instead of photolithography, was used to define the device source-drain contacts with 150 nm channel length, the device isolation pattern for the metal hard mask with width down to 150 nm, and the composite H_xWO_3/metal gate, using 6% ethyl lactate (EL-6) and polymethyl methacrylate (PMMA 950k A3, Kayaku Advanced Materials) bilayer as the photoresist. The exposure dose was 750 $\mu C \cdot cm^{-2}$ and the developer was methyl isobutyl ketone (MIBK) 1:2 diluted with isopropyl alcohol (IPA). In the device isolation step, 100 nm Au hardmask defined by liftoff was used instead of the photoresist, which was stripped by soaking the substrate in commercial gold etchant (Transene TFA) after the SF_6 RIE.

Monolithic integration of all-inorganic protonic ECRAMs with silicon MOSFETs. The process, which is schematically illustrated in Extended Data Figure 8, started from fabricating the silicon MOSFETs on a silicon-on-insulator (SOI) substrate (Soitec, 70 nm lightly *p*-doped device-layer silicon on 2 μm buried oxide). A layer of thermal oxide was first grown by dry oxidation at 1050 ºC to a thickness of about 90 nm. Photolithography was performed to define patterns of the heavily doped source-drain regions of the transistors into the photoresist (AZ-5214), followed by removing the SiO_2 in the exposed area with buffered oxide etchant. After stripping the photoresist in acetone, a film of phosphorus-containing spin-on-dopant (Filmtronics P509) was blanketly deposited by spin casting at 3,000 rpm for 30 seconds followed by a soft bake at 110 ºC for 3 min. Annealing at 850 ºC for 20 minutes in a three-zone tube furnace (Lindberg) with N_2 (2 $L \cdot min^{-1}$) and O_2 (1 $L \cdot min^{-1}$) flow caused the phosphorus to diffuse from the spin-on-dopant into the underlying silicon to form the heavily *n*-doped source/drain contact regions. After cooling down to room temperature, the wafer was immersed in HF to remove both the spin-on-dopant and the thermal oxide mask, followed by piranha cleaning to remove the residual phosphorus oxide. The device islands were then patterned by photolithography, and the silicon in the exposed area was removed by timed ICP-RIE (20 mtorr, 20 s.c.c.m. CF_4 flow, 300 W ICP power, and 100 W RIE power for 90 seconds) stopping on the buried oxide of the SOI wafer. After removing the photoresist by acetone and cleaning the surface by piranha solution, the source-drain contact electrodes were patterned by the third photolithography step, followed by the electron-beam evaporation of 2 nm Cr/40 nm Au and lift-off in acetone. 40 nm of HfO_2 was then deposited as the high-κ gate dielectric by ALD, using Tetrakis(dimethylamido)-hafnium and water as precursors at 200 ºC. Another photolithography and the lift-off scheme were subsequently performed again to define the metal gate electrodes composed of 2 nm Cr and 40 nm Au to complete the silicon MOSFET fabrication. Afterwards, another 40 nm HfO_2 was deposited by ALD at 200 ºC, serving as both the interlayer dielectric and the proton-diffusion barrier. The ECRAM layer was then fabricated on top using the process described above. The vias for interlayer interconnects were exposed by photolithography, with ICP-RIE utilized to

etch through the ZrO_2 gate electrolyte of the ECRAM, the HfO_2 interlayer dielectric, and the HfO_2 gate oxide of the silicon MOSFET to stop on the metal contact (5 mtorr, 10 s.c.c.m. CHF_3 flow and 5 s.c.c.m. Ar flow, 100 W ICP power, and 40 W RIE power). A final lithography and lift-off were performed to pattern the interlayer interconnects connecting the source electrodes of the silicon selector transistors to the gate electrodes of the ECRAMs.

Instrumentation. The cross-sectional high-angle annular dark field (HAADF) STEM images and the associated elemental configurations were obtained using the FEI Talos F200X G2 STEM equipped with four-crystal EDS system (FEI Super-X). All STEM-EDS data were collected for more than 30 minutes with a 10 microseconds dwell time and ~200 pA probe current, at an accelerating voltage of 200 kV. The sample was prepared using the Thermo Scios2 dual-beam focused-ion beam to a thickness around 50 nm under 5 kV. The SIMS elemental depth profile was obtained using the Phi TRIFT III time-of-flight SIMS system with the material removal performed using low-energy Cs^+ ion source. SEM micrographs were acquired using a Hitachi S4800 microscope. The device electrical characterizations were performed either in ambient or under high vacuum at the desired temperature using a manual probe station (LakeShore Cryotronics CRX–6.5K) connected with a semiconductor parameter analyzer (Keysight B1500A) equipped with integrated high-resolution source-measurement units (Keysight B1517A) and waveform generator/fast-measurement unit (Keysight B1530A). The AZ 5214E photoresist was patterned with a Heidelberg MLA150 aligner, and the EL6/PMMA photoresist was patterned with a Elionix ELS-G150 150 keV EBL system. The *Neurosim* simulations were performed on the Illinois Campus Cluster HAL[47]. The X-ray diffraction (XRD) spectrum was recorded using the Bruker D8 Advanced XRD system. The composition of deposited tungsten oxides was determined by Rutherford backscattering spectrometry (RBS) using the NEC Pelletron accelerator equipped with RBS chamber. Optical transmittances of WO_3 and H_xWO_3 were recorded using an Agilent Cary 5000 ultraviolet-visible absorbance spectrophotometer. XPS spectra of WO_3 and H_xWO_3 were acquired on a PHI Versa Probe III instrument with a

monochromatic Al Kα (1486.6 eV) source. The pass energy was 55 eV and spectra were referenced to C1*s* peak (adventitious carbon) at 284.8 eV.

Data availability: All data are available in the main text or the supplementary materials. All information and materials can be requested from the corresponding author.

Supplementary Information is linked to the online version of the paper at **www.nature.com/natelectron/**.

Acknowledgements This work was supported by the US National Science Foundation grant 1950182 (Q.C.) and 2139185 (Q.C. and J.-M.Z.). Significant aspects of the material characterizations and device fabrications were performed using the shared user facilities of the University of Illinois Materials Research Laboratory and Holonyak Micro and Nanotechnology Laboratory. This work also made use of the Illinois Campus Cluster HAL, utilizes resources supported by the National Science Foundation's Major Research Instrumentation program grant 1725729 and the University of Illinois at Urbana-Champaign. We thank Dr. Timothy Spila for assistance with RBS and SIMS measurements, and Dr. John Baltrus for performing the XPS measurements.

Author Contributions Q.C. conceived and designed the experiments. J.-S.C., F.-F.A., Y.-X.W., and L.L.S. performed the experiments. J.-C. Q., S.P. and J.-M.Z. performed the STEM/EDS analysis. Q.C. and J.-S.C. wrote the manuscript. All authors discussed the results (all of which are reported in the main text and supplement) and commented on the manuscript.

Competing Interests: J.-S.C. and Q.C. are inventors on a provisional patent application entitled "Solid-state electrochemical random access memory (ECRAM) and methods of making and operating a solid-state ECRAM" (63/434,627) submitted by the Board of Trustees of the University of Illinois.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Q.C. (email: qingcao2@illinois.edu).

References

1 Anthony, L. F. W., Kanding, B. & Selvan, R. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv*, 2007.03051 **[cs.CY]** (2020).

2 Xia, Q. & Yang, J. J. Memristive crossbar arrays for brain-inspired computing. *Nat. Mater.* **18**, 309-323 (2019).

3 Haensch, W., Gokmen, T. & Puri, R. The next generation of deep learning hardware: Analog computing. *Proc. IEEE* **107**, 108-122 (2019).

4 Yu, S. Neuro-inspired computing with emerging nonvolatile memorys. *Proc. IEEE* **106**, 260-285 (2018).

5 Xi, Y. *et al.* In-memory learning with analog resistive switching memory: A review and perspective. *Proc. IEEE* **109**, 14-42 (2021).

6 Gokmen, T. & Vlasov, Y. Acceleration of deep neural network training with resistive cross-point devices: Design considerations. *Front. Neurosci.* **10**, 333 (2016).

7 Ambrogio, S. *et al.* Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **558**, 60-67 (2018).

8 Prezioso, M. *et al.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61-64 (2015).

9 Yao, P. *et al.* Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641-646 (2020).

10 Wan, W. *et al.* A compute-in-memory chip based on resistive random-access memory. *Nature* **608**, 504-512 (2022).

11 Dutta, S. *et al.* Monolithic 3D integration of high endurance multi-bit ferroelectric FET for accelerating compute-in-memory. *IEDM Tech. Digest*, 36.34.31-36.34.34 (2020).

12 Gokmen, T. & Haensch, W. Algorithm for training neural networks on resistive device arrays. *Front. Neurosci.* **14**, 103 (2020).

13 Woo, J. *et al.* Improved synaptic behavior under identical pulses using AlO_x/HfO_2 bilayer RRAM array for neuromorphic systems. *IEEE Electron Device Lett.* **37**, 994-997 (2016).

14 Wu, W. *et al.* A methodology to improve linearity of analog RRAM for neuromorphic computing. *VLSI Symp. Tech. Digest*, 103-104 (2018).

15 Choi, S. *et al.* SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations. *Nat. Mater.* **17**, 335-340 (2018).

16 Mou, X. *et al.* Analog memristive synapse based on topotactic phase transition for high-performance neuromorphic computing and neural network pruning. *Sci. Adv.* **7**, eabh0648 (2021).

17 Widrow, B. *Generalization and information storage in networks of adaline neurons*. 435-461 (Spartan Books, 1962).

18 Fuller, E. J. *et al.* Redox transistors for neuromorphic computing. *IBM J. Res. Dev.* **63**, 9:1-9:9 (2019).

19 Gkoupidenis, P., Schaefer, N., Garlan, B. & Malliaras, G. G. Neuromorphic functions in PEDOT:PSS organic electrochemical transistors. *Adv. Mater.* **27**, 7176-7180 (2015).

20 Yao, X. *et al.* Protonic solid-state electrochemical synapse for physical neural networks. *Nat. Commun.* **11**, 3134 (2020).

21 Fuller, E. J. *et al.* Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing. *Science* **364**, 570-574 (2019).

22 Onen, M., Emond, N., Li, J., Yildiz, B. & del Alamo, J. A. CMOS-compatible protonic programmable resistor based on phosphosilicate glass electrolyte for analog deep learning. *Nano Lett.* **21**, 6111-6116 (2021).

23 Onen, M. *et al.* Nanosecond protonic programmable resistors for analog deep learning. *Science* **377**, 539-543 (2022).

24 Fuller, E. J. *et al.* Li-ion synaptic transistor for low power analog computing. *Adv. Mater.* **29**, 1604310 (2017).

25 Tang, J. *et al.* ECRAM as scalable synaptic cell for high-speed, low-power neuromorphic computing. *IEDM Tech. Digest*, 13.11.11-13.11.14 (2018).

26 Shi, J., Ha, S. D., Zhou, Y., Schoofs, F. & Ramanathan, S. A correlated nickelate synaptic transistor. *Nat. Commun.* **4**, 2676 (2013).

27 Kim, S. *et al.* Metal-oxide based, CMOS-compatible ECRAM for deep learning accelerator. *IEDM Tech. Digest*, 35.37.31-35.37.34 (2019).

28 Lee, C., Choi, W., Kwak, M., Kim, S. & Hwang, H. Excellent synapse characteristics of 50 nm vertical transistor with WO_x channel for high density neuromorphic system. *VLSI Symp. Tech. Digest*, 1-2 (2021).

29 Solomon, P. M. *et al.* Transient investigation of metal-oxide based, CMOS-compatible ECRAM. *IEEE Int. Reliab. Phys. Symp. Proc.*, 1-7 (2021).

30 Bishop, D. *et al.* Time-resolved conductance in electrochemical systems for neuromorphic computing. *2018 International Conference on Solid State Devices and Materials (SSDM)*, 23-24 (2018).

31 Jo, M. *et al.* Gate-induced massive and reversible phase transition of VO_2 channels ssing solid-state proton electrolytes. *Adv. Funct. Mater.* **28**, 1802003 (2018).

32 Oh, C. *et al.* Deep proton insertion assisted by oxygen vacancies for long-term memory in VO_2 synaptic transistor. *Adv. Electron. Mater.* **7**, 2000802 (2021).

33 Leng, X. *et al.* Insulator to metal transition in WO_3 induced by electrolyte gating. *npj Quantum Mater.* **2**, 35 (2017).

34 Meng, Y. *et al.* Review: Recent progress in low-temperature proton-conducting ceramics. *J. Mater. Sci.* **54**, 9291-9312 (2019).

35 Ding, K. *et al.* Phase-change heterostructure enables ultralow noise and drift for memory operation. *Science* **366**, 210-215 (2019).

36 Li, Y. *et al.* Capacitor-based cross-point array for analog neural network with record symmetry and linearity. *VLSI Symp. Tech. Digest*, 25-26 (2018).

37 Zhang, J.-G., Benson, D., Tracy, C. E., Webb, J. & Deb, S. *Self-bleaching mechanism of electrochromic WO_3 films*. Vol. 2017 104-112 (SPIE, 1993).

38 Chen, P., Peng, X. & Yu, S. NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **37**, 3067-3080 (2018).

39 Wen, R.-T., Granqvist, C. G. & Niklasson, G. A. Eliminating degradation and uncovering ion-trapping dynamics in electrochromic WO_3 thin films. *Nat. Mater.* **14**, 996-1001 (2015).

40 Melianas, A. *et al.* Temperature-resilient solid-state organic artificial synapses for neuromorphic computing. *Sci. Adv.* **6**, eabb2958 (2020).

41 Katase, T., Onozato, T., Hirono, M., Mizuno, T. & Ohta, H. A transparent electrochromic metal-insulator switching device with three-terminal transistor geometry. *Sci. Rep.* **6**, 25819 (2016).

42 Walczyk, C. *et al.* Impact of temperature on the resistive switching behavior of embedded HfO_2-based RRAM devices. *IEEE Trans. Electron Devices* **58**, 3124-3131 (2011).

43 Chen, P., Peng, X. & Yu, S. NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures. *IEDM Tech. Digest*, 6.1.1-6.1.4 (2017).

44 Peng, X., Huang, S., Jiang, H., Lu, A. & Yu, S. DNN+NeuroSim V2.0: An end-to-end benchmarking framework for compute-in-memory accelerators for on-chip training. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **40**, 2306-2319 (2021).

45 Zhou, Y. *et al.* Enhanced transmittance modulation of ITO/NiO_x/ZrO_2:H/WO_3/ITO electrochromic devices. *Ionics* **22**, 25-32 (2016).

46 Park, J. S. *et al.* Evidence of proton transport in atomic layer deposited Yttria-stabilized zirconia films. *Chem. Mater.* **22**, 5366-5370 (2010).

47 Kindratenko, V. *et al.* in *Practice and Experience in Advanced Research Computing (PEARC' 20).* 41–48 (Association for Computing Machinery).

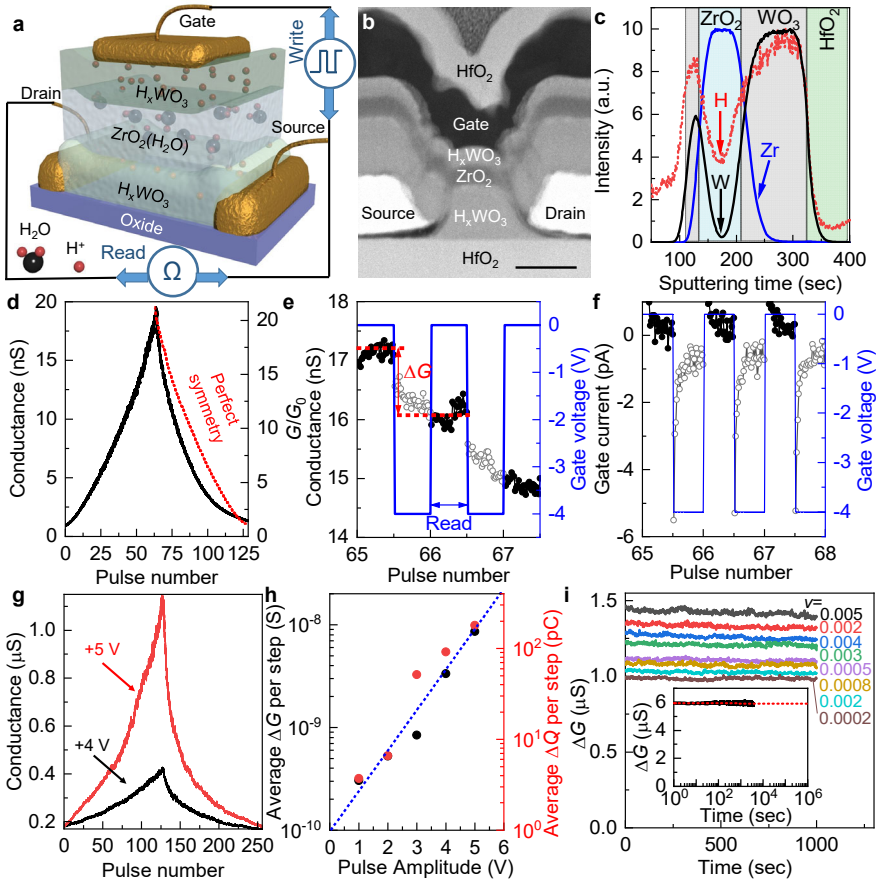


Figure 1. Structure and electrical characteristics of CMOS-compatible, all-inorganic protonic ECRAMs. **a,** Schematic illustrating the structure and read/write operations of the CMOS-compatible, all-inorganic protonic ECRAM prototype. **b,** Cross-sectional STEM micrograph showing the device gate stack. Scale bar: 50 nm. **c,** SIMS depth profiles for W (black), Zr (blue), and H (red) across the ECRAM device gate stack. Background color serves as visual guide to mark different layers including HfO_2 (green), H_xWO_3 (grey), and ZrO_2 (blue). The material interfaces are defined at positions where the major element (W and Zr for H_xWO_3 and ZrO_2 layers, respectively) composition has reached to 50% of the plateau. **d,** Programing of the CMOS-compatible, all-inorganic protonic ECRAM (channel length L_{ch}=10 μm and width W=3 μm) with gate-voltage pulses (64 potentiation then 64 depression pulses with amplitude of ±4 V and width of 3 sec). Red dotted line is the mirror image of the potentiation curve. The absolute device

channel conductance and the relative modulation with regard to the baseline conductance (G/G_0) are used for the left and right y-axis, respectively. **e-f,** Zoom-in views of the programing characteristics showing the readout of discrete conductance states with sufficient noise margin (part **e**, source-drain bias for read was 0.1 V) and the gate current during the weight-update (part **f**), driven by gate voltage pulses (blue, right axis). Red dotted lines serve as visual guide to mark the average channel conductance in each state. Current values measured during read and write operations are represented with solid and hollow symbols, respectively. **g,** Programing of the ECRAM fabricated on a more hydrogen-rich H_xWO_3 channel with voltage pulses of identical width of 3 sec but different amplitude of ±4 V (black) and ±5 V (red), respectively. **h,** Average ΔG per gate programing pulse (black, left axis) and the total charge injected by the gate current (ΔQ, red, right axis) as a function of the pulse amplitude. Blue dotted line represents the linear fitting to the data. **i,** Retention of selected analog states (0.1 V read) under zero gate bias in ambient and the corresponding drift coefficient (v). Inset: Retention with the gate floating. Red dotted line represents the fitting to the power decay function of $G(t)=G_{t0}(t/t_0)^{-v}$.

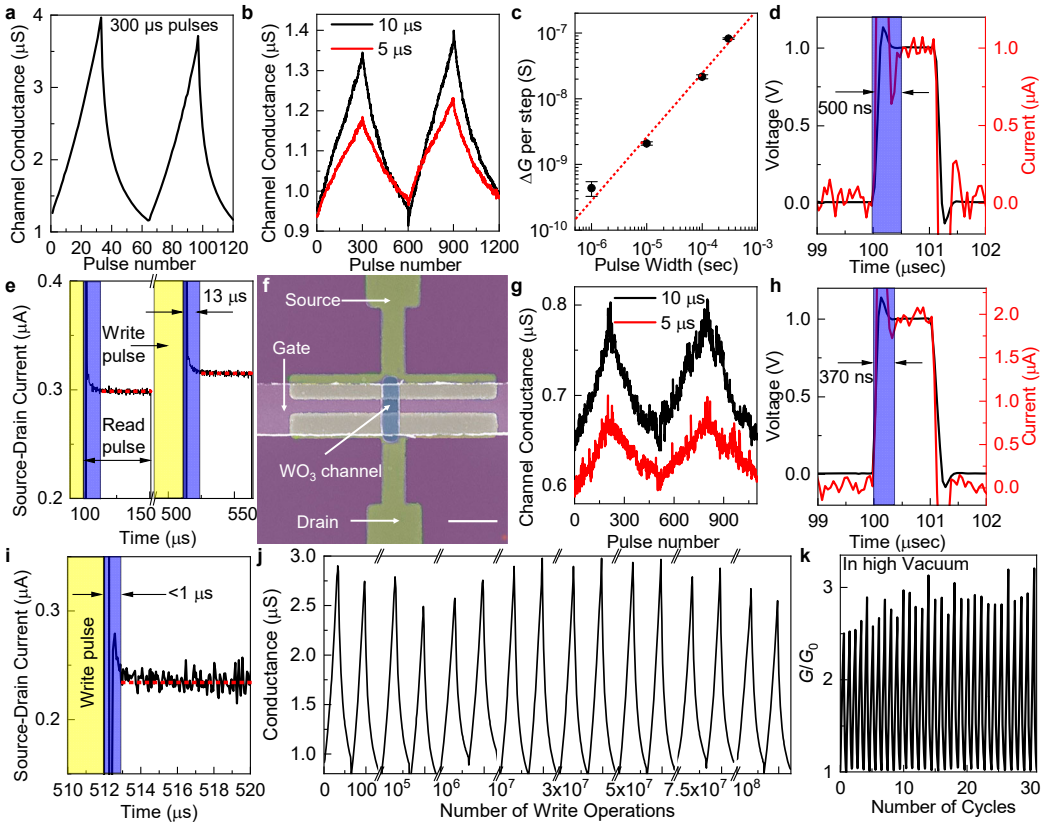


Figure 2. Speed, scaling, and endurance of CMOS-compatible, all-inorganic protonic ECRAMs. **a,** Programing of a micron-scale (L_{ch}=10 μm, W=3 μm) CMOS-compatible, all-inorganic protonic ECRAM with 300 μsec gate-voltage pulses (±4 V). **b,** Reproducible and highly symmetric programming with 10 μsec (black) and 5 μsec (red) write pulses (amplitude=±4 V). **c,** Double logarithmic scale plot showing the average ΔG per weight-update step as a function of the pulse width with the same pulse amplitude of 4 V. Red dotted line represents the linear fitting to the data. Error bars represent the standard deviation. **d,** Measured transient variation of the sense current (red, right axis) during the ECRAM read operation performed with a voltage pulse (black, left axis) applied on the drain electrode and the source/gate grounded. The blue shading highlights the settling time t_{read}. **e,** Recovery waveform of the sense current of read post write pulses. The yellow shadings indicate the write pulses of 300 μs, and the blue shadings highlight the settling time $t_{read\text{-}after\text{-}write}$ required to reach steady states

whose averages are marked with red dashed lines. **f,** False-colored scanning-electron microscopy (SEM) micrograph of a nanometer-scale (L_{ch}=150 nm, W=150 nm) CMOS-compatible, all-inorganic protonic ECRAM. Scale bar: 500 nm. **g,** Reproducible and symmetric switching characteristics of the scaled (L_{ch}=W=150 nm) ECRAM modulated with 10 μsec (black) and 5 μsec (red) write pulses (amplitude=±4 V). **h,** Waveforms of the read voltage (black, left axis) and the sense current (red, right axis) of the scaled ECRAM to extract its t_{read} as highlighted by the blue shading. **i,** Time-resolved source-drain current of the scaled ECRAM in a read performed right after a write pulse (yellow shadings), showing a faster settling time (blue shading) to steady state (red dashed line). **j,** Endurance test for 10^8 write-read pulses, showing no device degradation with the intermediate switching cycles plotted after 10^5, 10^6, 10^7, 3×10^7, 5×10^7, 7.5×10^7, and 10^8 pulses. **k,** Cycling the ECRAM passivated with HfO_2 between its high and low-conductance states in high vacuum of 10^{-6} Torr.

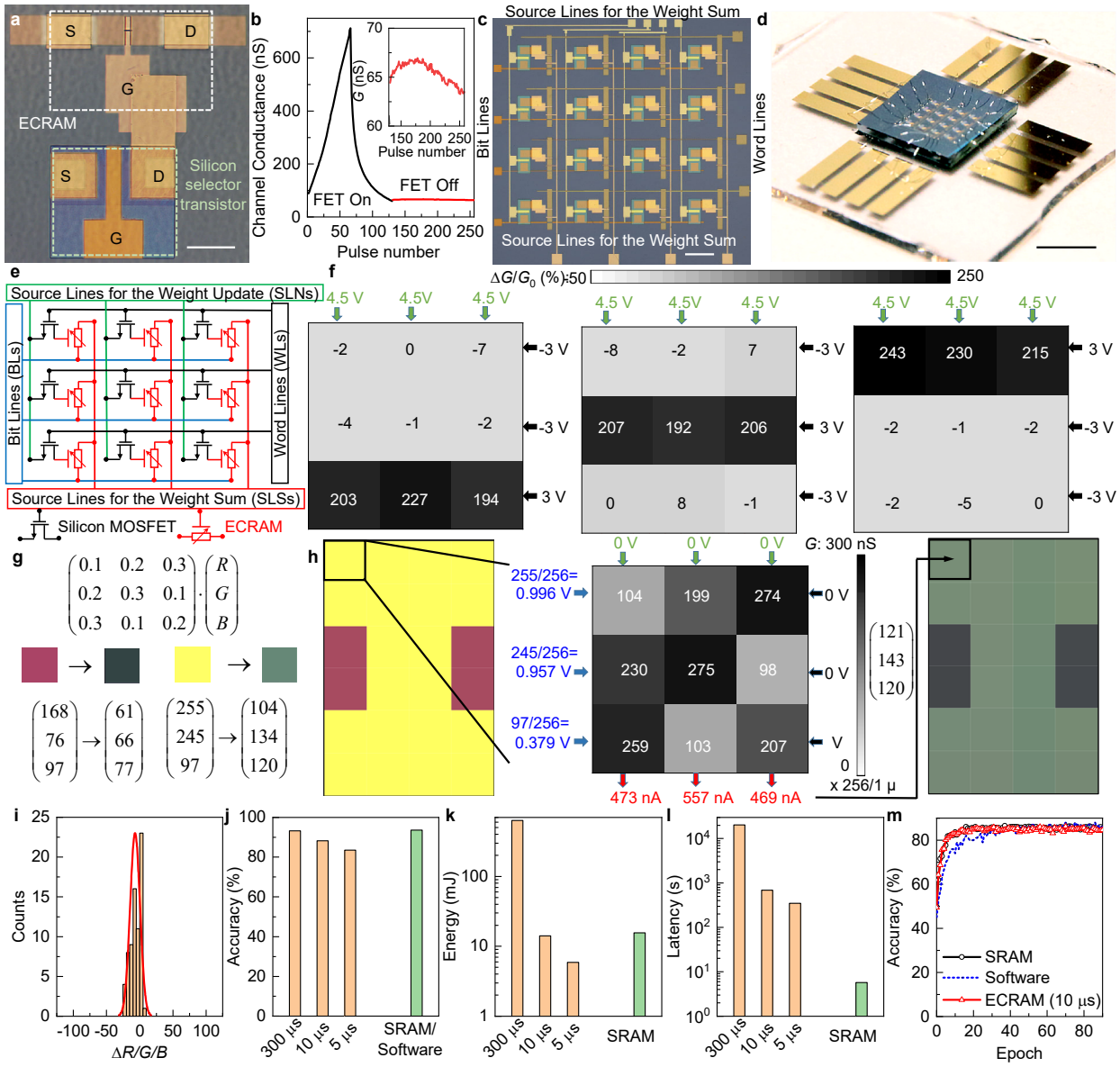
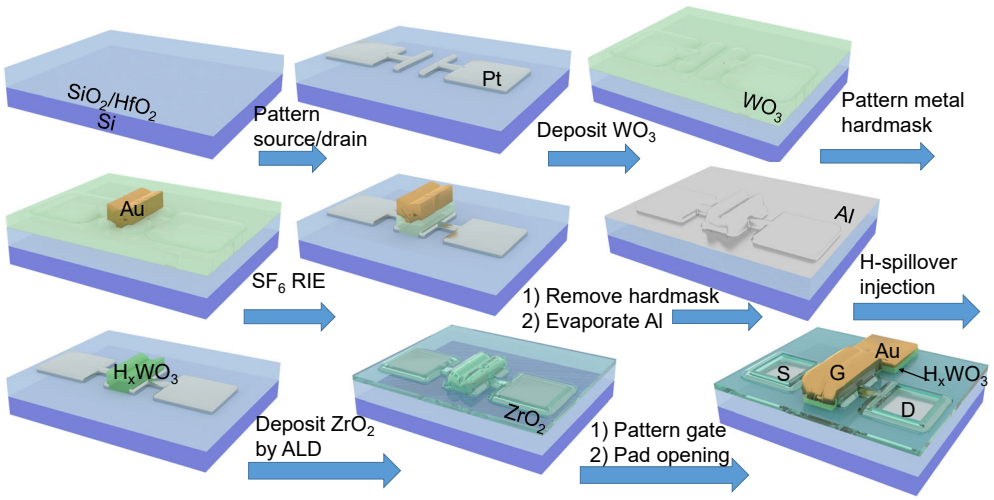
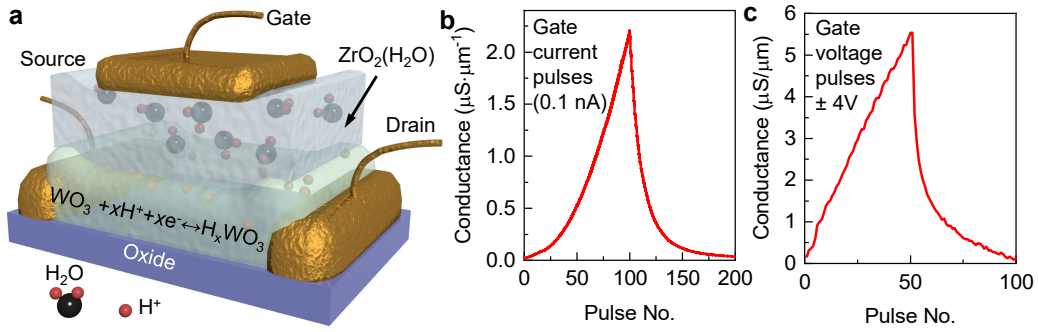


Figure 3. Monolithic integration of all-inorganic protonic ECRAMs with silicon transistors, and the operations of their pseudo-crossbar arrays as in-memory-computing accelerators. **a,** Optical image of a memory cell consisting of a ECRAM and a silicon MOSFET, which are fabricated on different layers. The white and green dashed lines serve as visual guide to mark the boundary of the ECRAM and the silicon selector transistor, respectively. Scale bar: 200 μm. **b,** Programing of the ECRAM cell by voltage-pulse trains composed of 64 potentiation followed by 64 depression steps, with the selector transistor turned either on (black) or off (red, inset). **c-d,** Optical

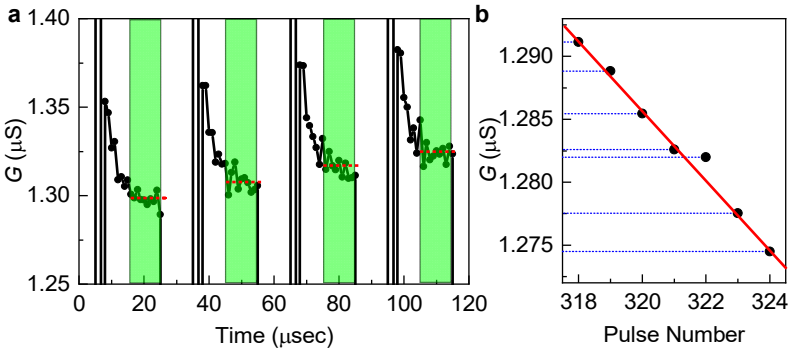
images of the fabricated ECRAM-cell array (part **c**, scale bar: 500 μm) and the bonded chip (part **d**, scale bar: 5 mm). **e,** Circuit diagram of the 3 by 3 ECRAM pseudo-crossbar array. **f,** Parallel row-by-row programing of the ECRAM array, with the normalized conductance modulation ($\Delta G/G$) indicated by grayscale. The voltages applied on the word lines (WLs) and the source lines for the weight update (SLNs) are shown in black and green, respectively. **g,** Color transformation performed by the dot product between the input *RGB* vector and the transformation matrix. **h,** Color transformation of a 4-pixel by 6-pixel image of letter I performed in parallel using the ECRAM array. Taken the top-left pixel as example, the encoded voltage inputs to the bit lines (BLs) are shown in blue, and the current outputs at the source lines for the weight sum (SLSs) measured in experiment are shown in red. The channel conductance *G* of each ECRAM after programing is illustrated in grayscale. **i,** Histogram showing the deviation of the modified *RGB* values calculated by the ECRAM array from those determined by software. **j-l,** Achievable accuracy (part **j**), energy consumption (part **k**), and latency (part **l**) for the ECRAM-CMOS hybrid in-memory-computing accelerator to learn the classification of the MNIST dataset using different write-pulse widths (orange), as benchmarked against the SRAM-based digital accelerator or software (green). **m,** Simulated accuracy of ECRAM (red, operated with 10 μsec gate-voltage pulses) and SRAM-based (black) accelerators to learn the classification of the CIFAR (Canadian Institute for Advanced Research)-10 dataset with the VGG (Visual Geometry Group)-8 network as a function of the training epochs performed. Results from learning in software (blue dotted line) are also displayed for comparison.



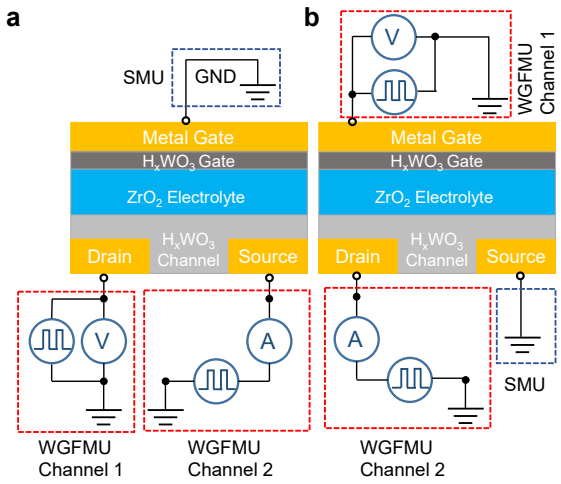
Extended Data Figure 1| Schematics illustrating the process flow to fabricate the CMOS-compatible, all-inorganic protonic ECRAM.



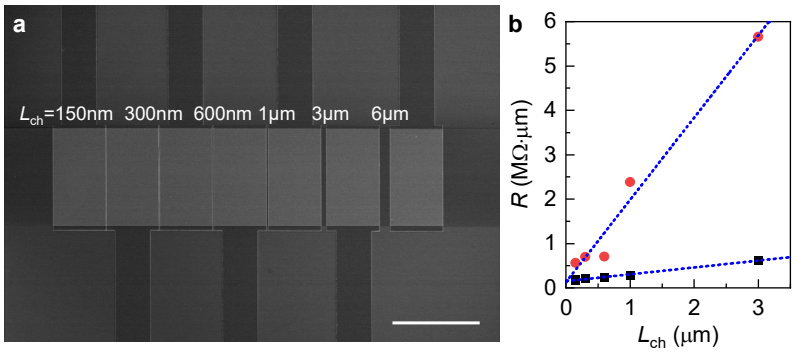
Extended Data Figure 2| All-inorganic protonic ECRAM with elemental metal gate. **a,** Schematic of the protonic ECRAM employing only Cr/Au as the gate electrode. **b,** Symmetric programing of the device channel conductance under gate-current pulses (100 potentiation then 100 depression with the gate-current pulse amplitude of 100 pA and width of 3 sec). **c,** Asymmetric programing of the same device but under gate-voltage pulses (50 potentiation then 50 depression with the gate-voltage pulse amplitude of ±4 V and width of 3 s), caused by the non-zero built-in open-circuit potential.



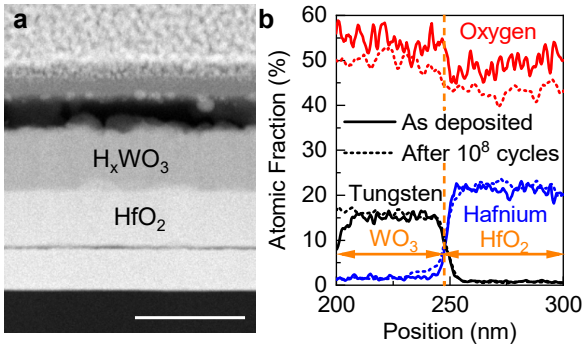
Extended Data Figure 3| Read noise and cycle-to-cycle variability of ECRAM operated with 10 μsec programming gate-voltage pulses. **a,** Zoom-in view of the programing characteristics showing the readout of four discrete conductance states with low read noise during the 10 μsec sensing period shaded in green, after 10 μsec settling time, out of the 300 conductance states shown in Fig. 2b. Red dotted lines serve as visual guide to mark the average channel conductance in each state. **b**, Zoom-in view of the read-out conductance (black circle) after the corresponding weight-update pulses. The red solid line is the expected conductance value obtained through fitting the depression curve. The small deviation of the experimental data to the expectation indicates small cycle-to-cycle variation. Blue dotted lines serve as visual guide to mark the measured conductance values on y axis.



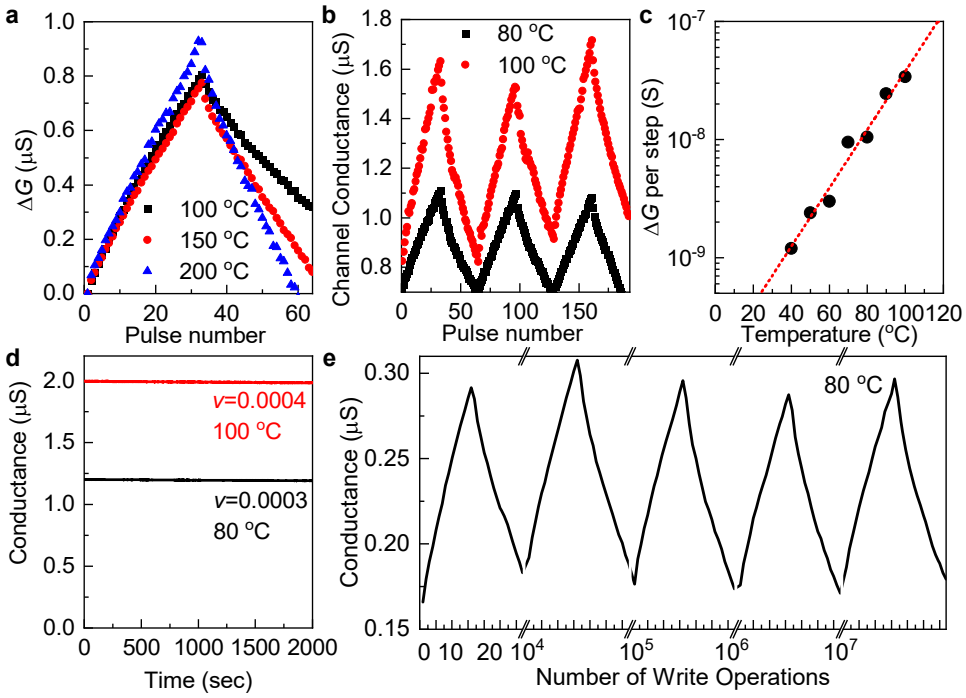
Extended Data Figure 4| ECRAM pulse-measurement setups to extract the read transients. **a,** Measurement configuration to extract t_{read}. WGFMUs are connected to the ECRAM source and drain terminals. The gate is grounded. During measurement, the voltage pulse is applied on the drain side and the other channel of WGFMU reads the current signal. **b,** Measurement configuration to perform the microsecond pulse write operations and extract $t_{read\text{-}after\text{-}write}$. WGFMUs are connected to the ECRAM gate and drain terminals. The source is grounded. During measurement, the voltage pulse is applied on the gate terminal for the weight-update and the other channel of WGFMU applies a small voltage pulse to reads the current flowing across the channel afterwards. SMU: source-measurement unit (Keysight B1517A). WGFMU: waveform generator/fast-measurement unit (Keysight B1530A). GND: ground.



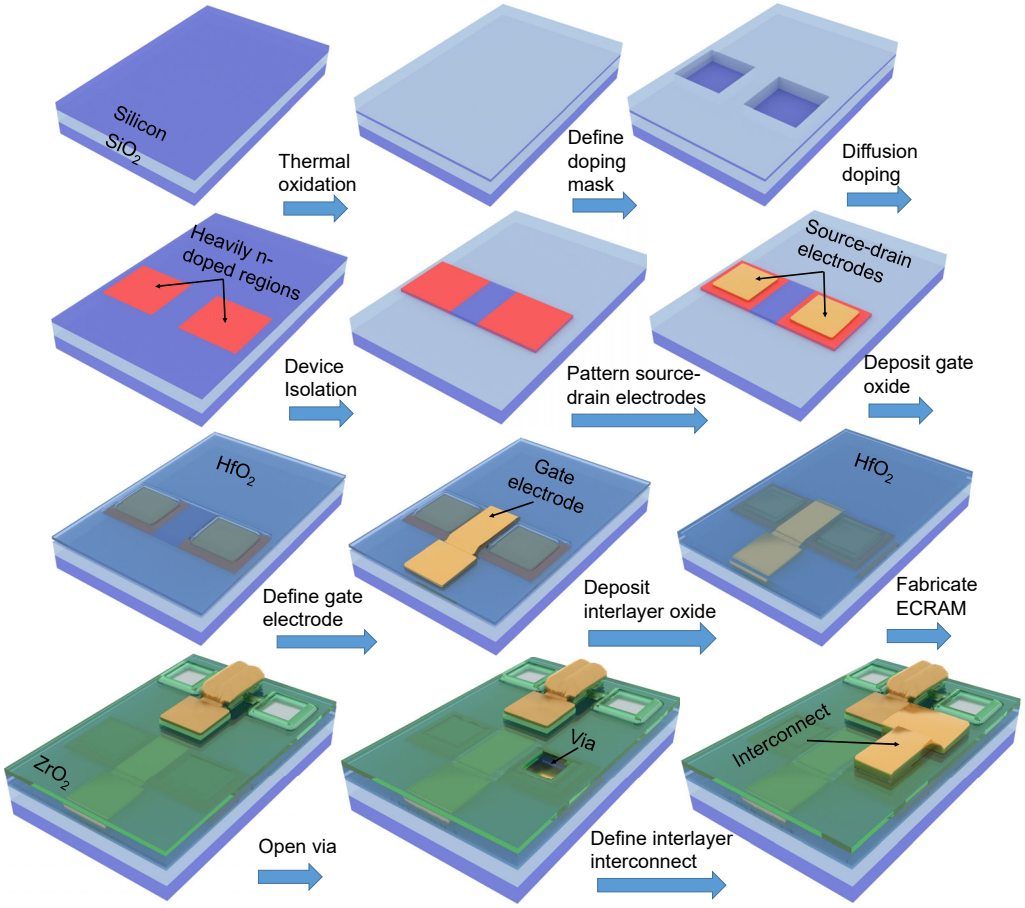
Extended Data Figure 5| Contact resistance of ECRAMs based on H_xWO_3. **a**, SEM micrograph showing the transmission-line structure fabricated with L_{ch} varied from 6 μm down to 150 nm but identical W of 60 μm. Scale bar: 50 μm. **b**, Width normalized device resistance R as a function of L_{ch} for H_xWO_3 with Pt contact switched between high (180 MΩ·sq^{-1}) and low (150 kΩ·sq^{-1}) sheet resistance corresponding to the operating dynamic range of ECRAM, showing a degradation of the device on/off ratio with the scaling of the L_{ch}.



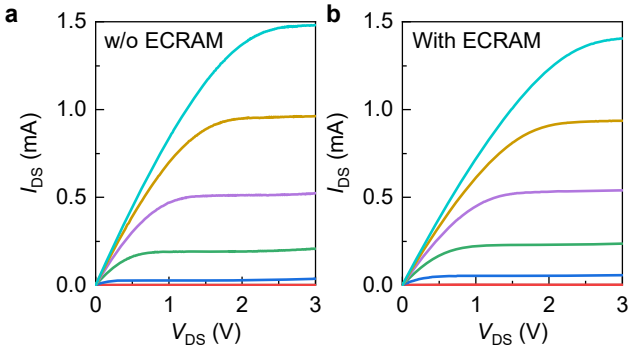
Extended Data Figure 6| H_xWO_3 channel after the endurance test. **a,** STEM micrograph showing the gate stack of the ECRAM after operation with 100 million read-write cycles. Scale bar: 100 nm. **b**, The depth profiles showing the atomic fractions of W (black), Hf (blue), and O (red) before (solid lines) and after (dotted lines) the 10^8 cycle endurance test, as measured by energy dispersive X-ray spectroscopy. The organ dashed line serves as a visual guide to mark the interface between WO_3 and HfO_2.



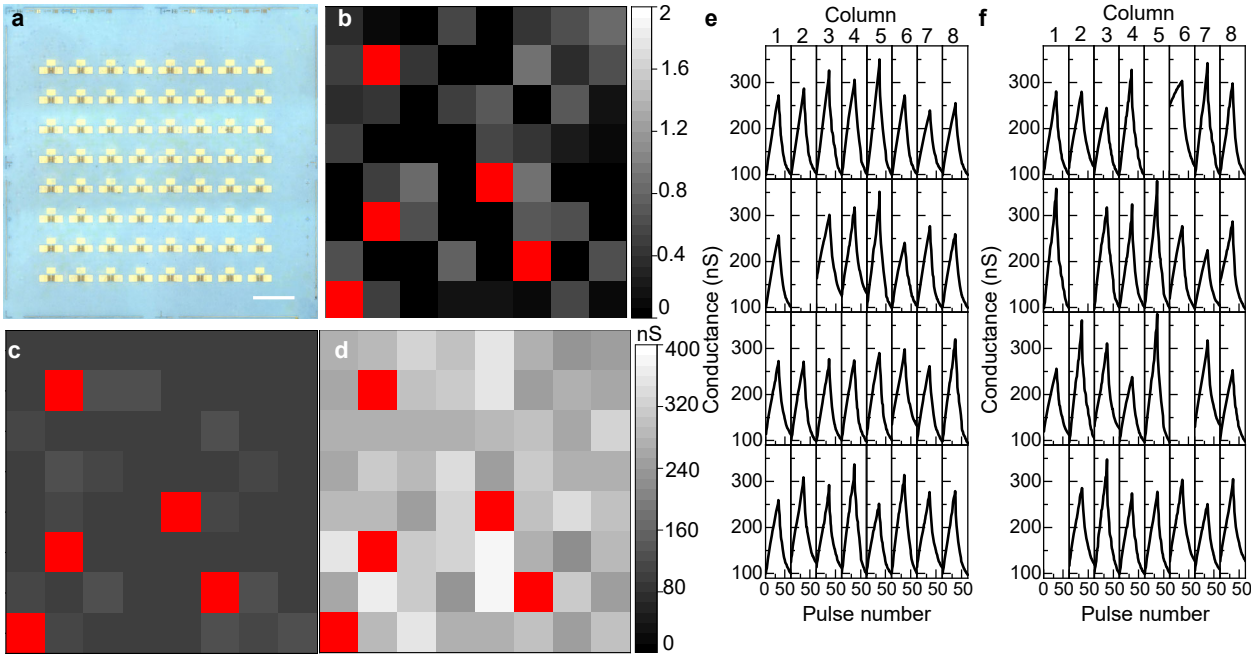
Extended Data Figure 7| Operation and reliability of all-inorganic protonic ECRAMs under elevated temperature. **a,** Programing of an all-inorganic protonic ECRAM with Al_2O_3 passivation measured at the same temperature after being annealed at 100 °C (black), 150 °C (red), and 200 °C (blue), respectively. **b**, Programing an all-inorganic protonic ECRAM at 80 °C (black) and 100 °C (red). G_0 and ΔG increased due to the thermal excitation of additional carriers in the H_xWO_3 channel and the increase of proton diffusivity, respectively. **c**, Logarithmic scale plot showing the average ΔG per weight-update step as a function of the device operating temperature. Red dotted line represents the linear fitting to the data. **d**, Retention of the fully on-state conductance of the ECRAM cell at 80 °C (black) and 100 °C (red), and the corresponding drift coefficient (v). **e**, ECRAM endurance test at 80 °C for 10^7 write-read pulses, showing no device degradation with the intermediate switching cycles plotted after 10^4, 10^5, 10^6, and 10^7 pulses.



Extended Data Figure 8| Schematics illustrating the process flow to fabricate a 1-transistor-1-ECRAM memory cell, consisted of an all-inorganic protonic ECRAM and a silicon MOSFET in different layers monolithically integrated together on a SOI wafer substrate.



Extended Data Figure 9| Current-voltage characteristics of the silicon MOSFET selector measured before (part a) and after (part b) fabricating the protonic ECRAM layer on top with 40 nm HfO_2 as the interlayer dielectric. V_{DS}: source-drain bias; I_{DS}: source-drain current.



Extended Data Figure 10| Device-to-device variation extracted from 8×8 ECRAM array. **a**, Optical micrograph showing the completed 8 by 8 ECRAM array. Scale bar: 500 μm. **b-d**, Spatial mapping of the non-linearity (part **b**, standard deviation around 10%), minimum conductance (part **c**, standard deviation about 12%), and maximum conductance (part **d**, standard deviation about 13%) of the fabricated ECRAMs. Red indicated failed devices resulting from lithography or material defects. **e-f**, Collections of the programing characteristics of ECRAMs in row 1-4 (from top to bottom, part **e**) and row 5-8 (part **f**) of the array, respectively.

References

45 Zhou, Y. *et al.* Enhanced transmittance modulation of ITO/NiO_x/ZrO_2:H/WO_3/ITO electrochromic devices. *Ionics* **22**, 25-32 (2016).

46 Park, J. S. *et al.* Evidence of proton transport in atomic layer deposited Yttria-stabilized zirconia films. *Chem. Mater.* **22**, 5366-5370 (2010).

47 Kindratenko, V. *et al.* in *Practice and Experience in Advanced Research Computing (PEARC' 20).* 41–48 (Association for Computing Machinery).