



Self-rectifying resistive memory in passive crossbar arrays

Kanghyeok Jeon^{1,2,4}, Jeeson Kim^{2,4}, Jin Joo Ryu¹, Seung-Jong Yoo^{1,2}, Choongseok Song², Min Kyu Yang³, Doo Seok Jeong²  [✉] & Gun Hwan Kim¹  [✉]

Conventional computing architectures are poor suited to the unique workload demands of deep learning, which has led to a surge in interest in memory-centric computing. Herein, a trilayer ($\text{Hf}_{0.8}\text{Si}_{0.2}\text{O}_2/\text{Al}_2\text{O}_3/\text{Hf}_{0.5}\text{Si}_{0.5}\text{O}_2$)-based self-rectifying resistive memory cell (SRMC) that exhibits (i) large selectivity (ca. 10^4), (ii) two-bit operation, (iii) low read power (4 and 0.8 nW for low and high resistance states, respectively), (iv) read latency ($<10\ \mu\text{s}$), (v) excellent non-volatility (data retention $>10^4\ \text{s}$ at $85\ ^\circ\text{C}$), and (vi) complementary metal-oxide-semiconductor compatibility (maximum supply voltage $\leq 5\ \text{V}$) is introduced, which outperforms previously reported SRMCs. These characteristics render the SRMC highly suitable for the main memory for memory-centric computing which can improve deep learning acceleration. Furthermore, the low programming power (ca. 18 nW), latency (100 μs), and endurance ($>10^6$) highlight the energy-efficiency and highly reliable random-access memory of our SRMC. The feasible operation of individual SRMCs in passive crossbar arrays of different sizes (30×30 , 160×160 , and 320×320) is attributed to the large asymmetry and nonlinearity in the current-voltage behavior of the proposed SRMC, verifying its potential for application in large-scale and high-density non-volatile memory for memory-centric computing.

¹Division of Advanced Materials, Korea Research Institute of Chemical Technology (KRICT) 141 Gajeong-Ro, Yuseong-Gu, Daejeon, Republic of Korea. ²Division of Materials Science and Engineering, Hanyang University, Seoul, Republic of Korea. ³Intelligent Electronic Device Lab, Sahmyook University, Seoul, Republic of Korea. ⁴These authors contributed equally: Kanghyeok Jeon, Jeeson Kim. ✉email: dooseokj@hanyang.ac.kr; kimgh@kRICT.re.kr

Recent trends in computation highlight a shift from conventional computing towards memory-centric computing. In conventional computing the processors are central, and the data subject to processing are transferred to the processors from a separate memory unit. Memory-centric computing avoids this data transfer through the memory hierarchy by placing processing power near the memory domain^{1,2}. Examples of memory-centric computing include near-data-processing and in-memory processing (also known as processing-in-memory or computing-in-memory). A significant motivator for this shift is deep learning which requires immense memory capacity but simple data processing. Conventional computing exhibits significant shortcomings for this particular workload due to the enormous amount of data transferred between the separated memory and processors. These shortcomings include the memory wall, arising from the difference in performance between the processor and memory (processor > memory) and the consequent bottleneck in performance caused by the memory latency, and the immense power consumption over the buses between the processor and memory³. Specifically, the majority of the workload for deep learning results from the elementary operation for vector-matrix multiplication, which is a multiply-accumulate operation (one multiplication and one accumulation operation). Despite the simplicity of each operation, repetition over a massive matrix creates an enormous workload for the hardware because of its complexity $O(n^2)$. Notably, the trend in deep learning computing in recent years indicates an exponential increase in operation number; AlphaGo Zero in 2018 needed approximately 300,000 times the number of operations that were required for AlexNet in 2012. This trend is expected to continue. Therefore, employing memory-centric computing as a complementary approach or, more radically, an alternative approach to conventional computing is unavoidable if we wish to maintain sustainable progress in deep learning technologies.

Inference in deep learning only needs to read the weights in the memory, unlike training that needs to read and write the weights. Most of the workload for the hardware arises from inference rather than training as fully trained neural networks are only minimally re-trained and repeatedly applied to the given input data. Therefore, memory-centric computing for deep learning acceleration requires appropriate memories that have (i) large memory capacity, (ii) low latency in-memory read-out, (iii) low power consumption on memory read-out, (iv) non-volatility, and (v) complementary metal-oxide-semiconductor (CMOS) compatibility. Fast writing at low power is also desirable as a second priority. A common measure of hardware performance for inference is tera-operations per second per watt; therefore, requirements (ii) and (iii) directly improve the hardware performance. Requirement (i) is necessary because the state-of-the-art deep neural networks (DNNs) that can recognize real-world data are substantial in depth and unit number per layer. For instance, convolutional neural network (CNN)-based DNNs, such as, AlexNet⁴, VGGNet (specifically, VGG-19)⁵, GoogLeNet⁶, ResNet (specifically, ResNet-152)⁷, include approximately 60 M, 138 M, 4 M, and 60 M weights, respectively. When using a full precision float 32-bit format, the memory for the weights in a single model reaches 1.9 Gb, 4.4 Gb, 128 Mb, and 1.9 Gb, respectively. The memory should be sufficiently large to host these weights on-site to accelerate significantly the inference task. Requirement (iv) avoids loading the memory with massive amounts of weight data whenever it is rebooted. Compatibility with standard CMOS technologies (requirement (v)) is a critical criterion because the memory should be cointegrated with CMOS-based processing units.

Considering these requirements, there are several non-volatile memories which are regarded as potential storage-class memories

combining the advantages of main memories (random-access memory (RAM)) and data storage. They include ferroelectric RAM (FRAM)⁸, spin-torque-transfer RAM (STT-RAM)⁹, phase-change RAM (PcRAM)¹⁰, and resistive RAM (RRAM)¹¹. Thus far, these have not been commercialized as standalone storage-class memories due to a few shortcomings. For instance, FRAM and STT-RAM have an unavoidable limit to their memory capacity due to the use of transistors as bit-cell selectors and difficulty in fabrication. PcRAM may achieve high memory capacity using passive bit-cell selectors such as diodes and ovonic threshold switches; however, its high programming power and difficulty in multilevel programming preclude it from commercialization as storage-class memory. Similar to PcRAM, RRAM achieves high memory capacity and offers multilevel programming; however, its programming endurance is much lower than dynamic RAM and static RAM. Nevertheless, among these non-volatile memories, RRAM is the most likely candidate for the memory for memory-centric architecture for deep learning acceleration regarding requirements (i)–(iv) as the top priority. As a second priority, the advantages presented by RRAM in satisfying these requirements far outweigh its drawback of limited programming endurance. As a result, in-memory processors based on non-volatile memory frequently employ RRAM loaded with weight matrices^{12–16}.

RRAM offers feasible solutions to high memory capacity (requirement (i)) due to its multilevel operation and scalability down to the $4F^2$ design rule. Each memory bit-cell in RRAM is capable of multibit (n -bit; $n > 1$) representation using its 2^n resistance levels^{17,18}. This significantly increases its memory capacity. Moreover, RRAM can be realized in passive crossbar arrays (CAs) where each bit-cell is formed at an intersection of a row- and column-line, meeting the ultimate $4F^2$ design rule^{19–21}. Furthermore, the sneak current through unchosen cells, that leads to read-out errors should be considered^{22,23}. Staking a passive selector on the memory cell at a cross-point avoids read-out errors only if it is possible for the selector to address just the chosen bits with negligible interference, similar to transistors in active CAs. However, because a single terminal is used to turn on the selector, read, and program the memory cell, the independent operation of each of the two series elements may be challenging unless the selector is specifically tailored to the memory cell or vice versa. An alternative is to use self-rectifying memory cells (SRMCs) which are single memory cells, whose highly nonlinear and asymmetric current–voltage (I – V) behavior alone enables the current sensing amplifier to distinguish between chosen and unchosen cells^{24–30}. SRMCs have attracted significant attention because of their simplicity in bit-cell structure and thus potential compatibility with three-dimensional memory structure, enriching candidates for SRMCs, for example, $\text{NbO}_x/\text{TiO}_x/\text{NbO}_x$ ²⁶, $\text{TiO}_2/\text{HfO}_2$ ²⁹, $\text{Ta}_2\text{O}_5/\text{HfO}_2$ ²⁴, and Al-doped HfO_2 ²⁷. Albeit excellent in most aspects, each has shortcomings that hinder it from being a promising candidate for an SRMC.

In this paper, we propose an $\text{Hf}_{0.8}\text{Si}_{0.2}\text{O}_2/\text{Al}_2\text{O}_3/\text{Hf}_{0.5}\text{Si}_{0.5}\text{O}_2$ trilayer-based SRMC that accurately meets the requirements for the main memory in memory-centric computing. Our proposed device has high selectivity (ca. 10^4) and reliable 2-bit representation, which were verified in single cells in support of requirement (i), along with extremely low power consumption on a single read-out operation with 4 and 0.8 nW for low resistance state (LRS) and high resistance state (HRS), respectively, and latency of $<10 \mu\text{s}$ in a single read-out operation in support of requirements (ii) and (iii). Moreover we also demonstrate the excellent non-volatility (data retention $>10^4$ s at 85°C) in support of requirement (iv), and a programming pulse amplitude below 5 V, which is compatible with the CMOS voltage driving circuits in support of requirement (v). We summarize these features and

Table 1 Performance comparison between our SRMC and previous results.

	Memory capacity		Read power (nW)		Read latency	Non-volatility Retention	CMOS compatibility Max. supply voltage
	Selectivity ($I @ V_{op} / I @ -1/3V_{op}$)	Multibit operation	LRS	HRS			
First priority							
This work	$\sim 10^4$	2 bits	4	0.8	$< 10 \mu\text{s}$	$> 10^4$ s (cumulative)	≤ 5 V
Haili et al. ²⁴	$\sim 10^4$	-	0.5	5×10^{-3}	-	$> 10^4$ s	-
Yoon et al. ²⁵	$\sim 10^4$	-	80	8×10^{-3}	-	$> 10^4$	-
Kim et al. ²⁶	$\sim 2 \times 10^4$	-	0.6	3×10^{-3}	-	$> 10^4$	-
Huang et al. ²⁷	$\sim 3 \times 10^2$	-	1.4	0.2	-	$> 10^4$	≤ 4 V
Zhou et al. ²⁸	$\sim 10^2$	2 bits	0.3	2×10^{-3}	-	$> 10^4$	≤ 4 V
Hsu et al. ²⁹	$\sim 10^3$	2 bits	2	2×10^3	-	$> 10^4$	-
Chou et al. ³⁰	$\sim 3 \times 10^4$	-	1.2	2×10^3	-	$> 10^4$	-
Second priority							
	Program power (nW)	Program latency	Program endurance	Test scale	Device structure		
This work	18	100 μs	$> 10^6$	30 \times 30, 160 \times 160, 320 \times 320 arrays	Ru/Hf _{0.8} Si _{0.2} O ₂ /Al ₂ O ₃ /Hf _{0.5} Si _{0.5} O ₂ /TiN		
Haili et al. ²⁴	6×10^3	-	-	-	Pt/Ta ₂ O ₅		
Yoon et al. ²⁵	100	-	$\sim 10^3$ (DC)	-	/HfO _{2-x} /Hf		
Kim et al. ²⁶	100	-	$\sim 5 \times 10^3$ (DC)	-	Pt/Ta ₂ O ₅		
Huang et al. ²⁷	1.3×10^3	1 μs	$\sim 10^5$	-	/HfO _{2-x}		
Zhou et al. ²⁸	4	10 ms	$\sim 5 \times 10^2$	-	/TiN		
Hsu et al. ²⁹	8×10^6	-	-	-	Pt/NbO _x		
Chou et al. ³⁰	6×10^3	-	20	36 bit array	/TiO _y /NbO _x		
					/TiN		
					TiN		
					/Al-HfO _x		
					/SiO ₂ /Si		
					Cu/Al ₂ O ₃ /aSi/Ta		
					Ni/TiO ₂		
					/HfO ₂ /Ni		
					Ta/TaO _x		
					/TiO ₂ /Ti		

compare with findings from previous studies in Table 1. The selectivity value of each reference was extracted from the current ratio between at the maximum programming voltage and its negative one-third voltage.

Results

Resistive switching operation of unit SRMCs. The proposed SRMC is based on an Hf_{0.8}Si_{0.2}O₂/Al₂O₃/Hf_{0.5}Si_{0.5}O₂ trilayer between a Ru top electrode (TE) and TiN bottom electrode (BE). The device fabrication procedure is detailed in the Methods Section. Figure 1a shows 90 *I*-*V* hysteresis loops at 85 °C for 30 different SRMCs of 2 \times 2 μm^2 area (three loops each). We chose a memory operation temperature of 85 °C which is the upper bound of the industrial temperature range (-40 – 85 °C). Notably, no electroforming was needed to activate the switching behavior. From the results, negligible cell-to-cell variability in *I*-*V* behavior even at the elevated temperature was first identified. With the measured bipolar switching (BS) characteristics, both set (from HRS to LRS) and reset (from LRS to HRS) switching events are gradual under positive and negative voltage, respectively. The *I*-*V* loops in Fig. 1a highlight large asymmetry in *I*-*V* behavior between positive and negative voltage and large nonlinearity in *I*-*V* on both sides and are eligible to be used as SRMCs. A voltage range that allows extremely low current (barely sufficient to distinguish between different resistance states) is referred to as an inhibit region; the large inhibit region (-0.8 – 0.6 V) of our SRMC is favorable to inhibit the sneak current through unselected cells in a passive CA. Additionally, given the gradual set switching

behavior, which is a self-compliance characteristic, no external current compliance is needed to protect the cell from a hard breakdown. This self-compliance characteristic is particularly desirable for passive CAs because a lack of transistors would otherwise limit current flow through memory cells.

We subsequently examined the BS of our SRMC in response to programming voltage pulses of different amplitudes (0–4.3 V) and widths (50 μs , 100 μs , and 1 ms). The measurement results in Fig. 1b indicate the onset of set switching at a positive voltage and a gradual reset behavior with the increase in reset voltage amplitude. The onset implies a threshold voltage for set switching, which enables non-destructive reading with a read-out voltage below this threshold voltage. Accordingly, we chose a read-out voltage of 2 V. The voltage pulses of 50 μs duration were insufficient to set the SRMC to a fully LRS, unlike the 100 μs and 1 ms duration cases (Fig. 1b), so we set the standard programming pulse width to 100 μs thereafter.

The high resistance, even in the LRS, causes a long RC time constant. This delimits the read-out speed significantly. We examined the read-out speed of the SRMC by applying a read-out pulse (2 V in amplitude and 5 μs in width) to five LRS SRMCs in parallel (Fig. 1c). It should be noted that we used five parallel SRMCs because the current level from a single SRMC is so small that it is barely measurable by an oscilloscope. The current plateau was reached after $\sim 3 \mu\text{s}$ and the same delay was shown during discharging. Therefore, the read-out latency is below 10 μs .

The key to non-volatility is data retention at real memory-operating temperatures. Hence, we tested the stability of the LRS

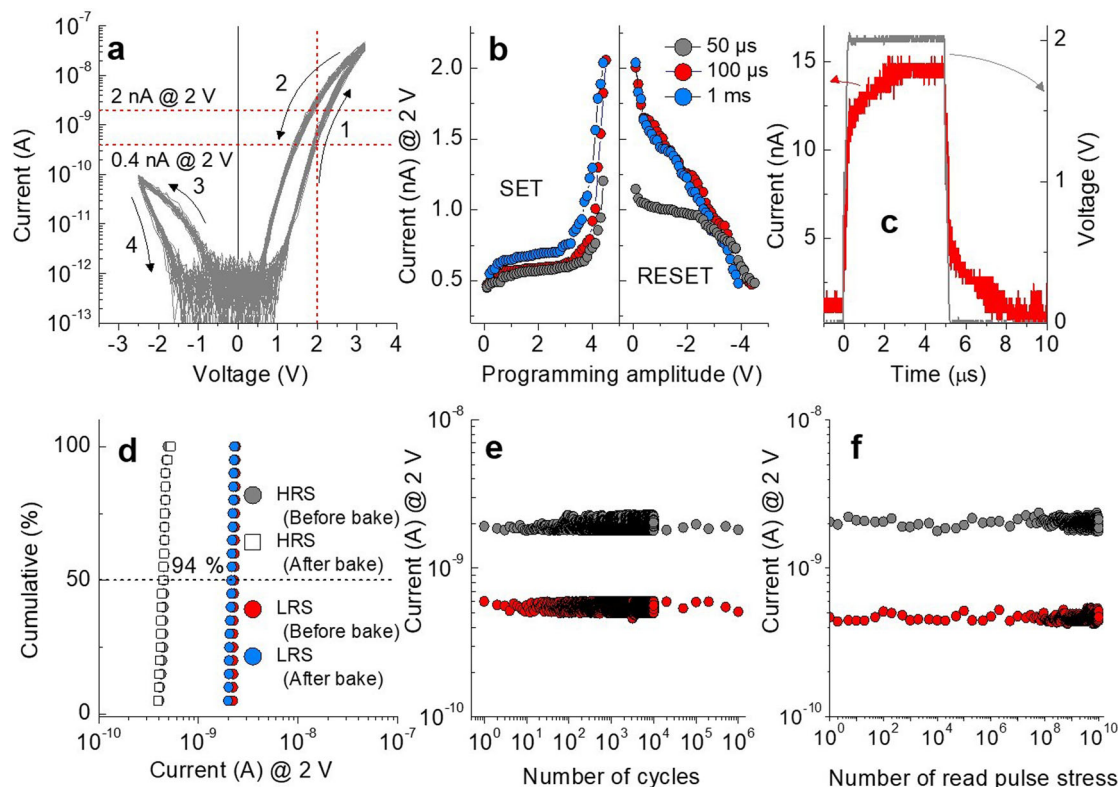


Fig. 1 Electrical characterization of the unit self-rectifying resistive memory cell (SRMC). **a** DC I - V characteristics of 30 SRMCs. Arrows indicate switching direction. Readable current margin verified at 2 V is 0.4–2 nA. **b** Resistance states programmed by varying amplitude of programming voltage pulse for three pulse widths (50 μ s, 100 μ s, and 1 ms). **c** Read-out current in response to read-out pulse (2 V and 5 μ s in amplitude and width, respectively). Current evaluated from voltage across 1 M Ω internal resistor of oscilloscope. **d** Memory retention of characteristic of 20 SRMCs in HRS and LRS as programmed and after baking (at 85 $^{\circ}$ C for 2 h). **e** Programming endurance of SRMC using 4.2 V/100 μ s set and -4.3 V/100 μ s reset pulses. **f** Read disturb characteristic of SRMC using repetitive reading pulse of 2 V/10 μ s. (gray and red circle for LRS and HRS, respectively).

for 20 SRMCs maintained at the elevated temperature of 85 $^{\circ}$ C for 2 h. The 20 SRMCs were first programmed to the HRS using identical reset voltage pulses and their currents were read at 2 V. They were subsequently programmed to the LRS using identical set voltage pulses and the currents were read at 2 V. The 20 SRMCs in the LRS were heated up to 85 $^{\circ}$ C for 2 h, followed by current read-out at 2 V. The results in Fig. 1d indicate the excellent data retention even at the elevated temperature and almost negligible cell-to-cell variability in BS operation. Additionally, retention measurement at a higher temperature (125 $^{\circ}$ C for 2 h) on a single cell was also performed to confirm the stable data non-volatility as shown in Supplementary Fig. 1. We also identified the programming endurance of the SRMC for up to 10^6 cycles, each with +4.2 V set and -4.3 V reset pulses (Fig. 1e). As elaborated in the Introduction section, because the number of *read* operation is much larger than that of writing operation in in-memory computing application, we examined read disturb characteristics of LRS and HRS by applying repetitive read pulses of 2 V. (10 μ s width) (Fig. 1f) Up to 10^{10} of reading operation, our SRMC showed stable non-volatility in each resistance states, which largely exceeds the 10^6 of endurance characteristic. (Fig. 1e)

Structural analyses of SRMC. Our SRMC is a vertical stack of Ru/ $\text{Hf}_{0.8}\text{Si}_{0.2}\text{O}_2/\text{Al}_2\text{O}_3/\text{Hf}_{0.5}\text{Si}_{0.5}\text{O}_2/\text{TiN}$ as confirmed by a cross-sectional high-resolution transmission electron microscope (HR-TEM) image (Fig. 2a). The upper $\text{Hf}_{0.8}\text{Si}_{0.2}\text{O}_2$ and lower $\text{Hf}_{0.5}\text{Si}_{0.5}\text{O}_2$ differ in chemical composition and are referred to as HSO¹ and HSO², respectively. HSO¹ and HSO² are separated by

a 1-nm-thick Al_2O_3 layer. These three layers are sandwiched between a Ru TE and TiN BE. Auger electron spectroscopy (AES) analyses on the SRMC consistently identify the stack structure as shown in Fig. 2b. Further analysis of the AES data indicates that HSO¹ and HSO² differ in chemical composition such that the cation-to-anion ratio is larger in HSO¹ than in HSO² (Fig. 2c). Additionally, we performed X-ray photoelectron spectroscopy analysis on our SRMC stack to compare the HSO¹ and HSO² layers (Fig. 2d–f). The two layers largely differ in the peak energy of an O1s spectrum; the spectrum for HSO¹ peaks at approximately 530.4 eV while that for HSO² peaks at \sim 530.0 eV. The higher peak energy in HSO¹ indicates a higher concentration of non-lattice oxygen than in HSO²^{31,32}. Rutherford Backscattering Spectroscopy (RBS) measurement results shown in Supplementary Fig. 2 indicate that the chemical composition of HSO¹ and HSO² is $\text{Hf}_{0.8}\text{Si}_{0.2}\text{O}_2$ and $\text{Hf}_{0.5}\text{Si}_{0.5}\text{O}_2$, respectively.

Resistive switching mechanism and current behavior of SRMC. Regarding current transport in our SRMC, the current in both the LRS and HRS scales with a device area in the wide range 0.0484–100 μm^2 is plotted in Supplementary Fig. 3. This indicates interface-type switching as opposed to localized switching³³; the whole device area is responsible for the switching by modulating the interfacial electronic energy barrier in a non-volatile manner^{34,35}. This is consistent with the fact that our SRMC did not require an electroforming process, which is known to introduce conducting filaments^{11,36}. In this regard, the largely asymmetric I - V behavior may be due to the use of asymmetric metal electrodes and thus asymmetric interfacial barrier heights.

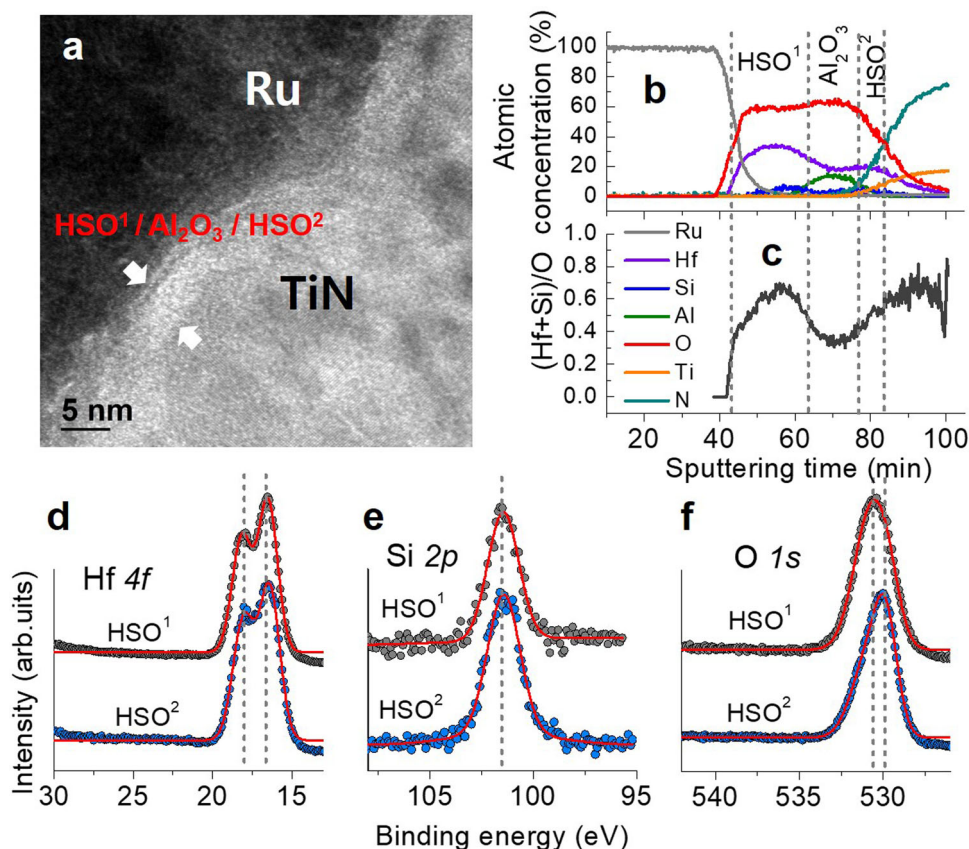


Fig. 2 Microstructural and chemical analyses. **a** Cross-sectional high-resolution transmission electron microscope image of our SRMC. **b** Depth profile of the elements, which were measured by Auger electron microscopy. **c** Atomic ratio (Hf+Si)/O along depth of SRMC. X-ray photoelectron spectra of **d** Hf4f, **e** Si2p, and **f** O1s emission for HSO¹ and HSO².

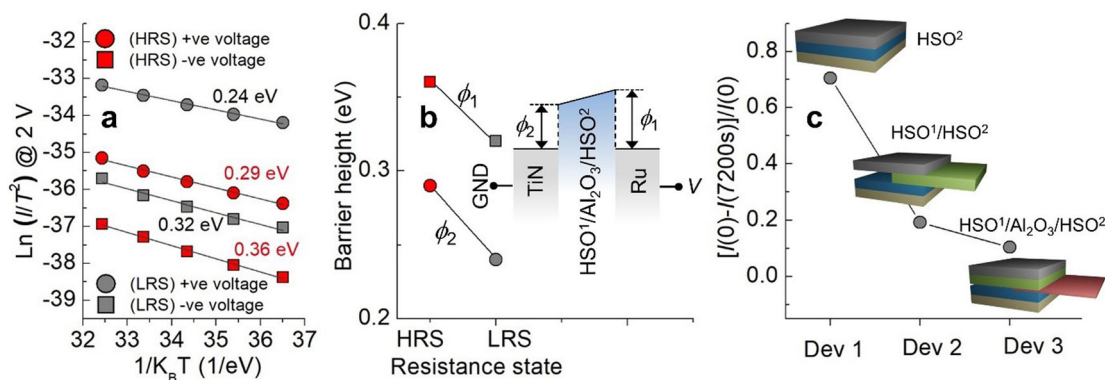


Fig. 3 Current behavior in temperature domain emission equation and data retention for SRMC devices. **a** Fitting Schottky emission equation to current measured at various temperatures (45–85 °C) and ±2 V for HRS and LRS. **b** Estimated barrier heights (ϕ_1 and ϕ_2) indicated in inset for HRS and LRS. **c** Data retention for the proposed SRMC (Dev 3) at 85 °C compared with Dev 1 (Ru/HSO²/TiN) and Dev 2 (Ru/HSO¹/HSO²/TiN). The as-programmed current level and current level at 7200 s are denoted by $I(0)$ and $I(7200\text{ s})$, respectively.

The asymmetry in the interfacial barrier height was acquired by fitting the Schottky emission equation^{37,38} to the experimental I – V data at different temperatures (45–85 °C). Here, the assumption was that the interfacial barrier at the cathode dictates the overall current transport through the SRMC such that the barrier limits the injection current level. The measured data on the $\ln(I/I^2)$ and reciprocal $k_B T$ plane indicate good linearity, where T and k_B are absolute temperature and Boltzmann’s constant, respectively (Fig. 3a). This analysis yields a barrier height pair, at the top and bottom interfaces, for the HRS and LRS. For both states, electron injection from the TE (i.e., under negative voltage), encounters a

higher Schottky barrier than the injection from the BE (i.e., under positive voltage), implying asymmetry in the barrier height due to asymmetry in the electrode presently in use (Fig. 3b).

The change in barrier height upon switching may be attributed to oxygen vacancy redistribution by the applied programming voltage^{39,40}. Oxygen vacancies are redistributed in response to the direction of a programming field by electronic drift, resulting in the polarization of space charge. However, one should consider the high depolarization field built up during the programming period, which takes effect immediately after the removal of the programming field⁴⁰. This presents a significant challenge for

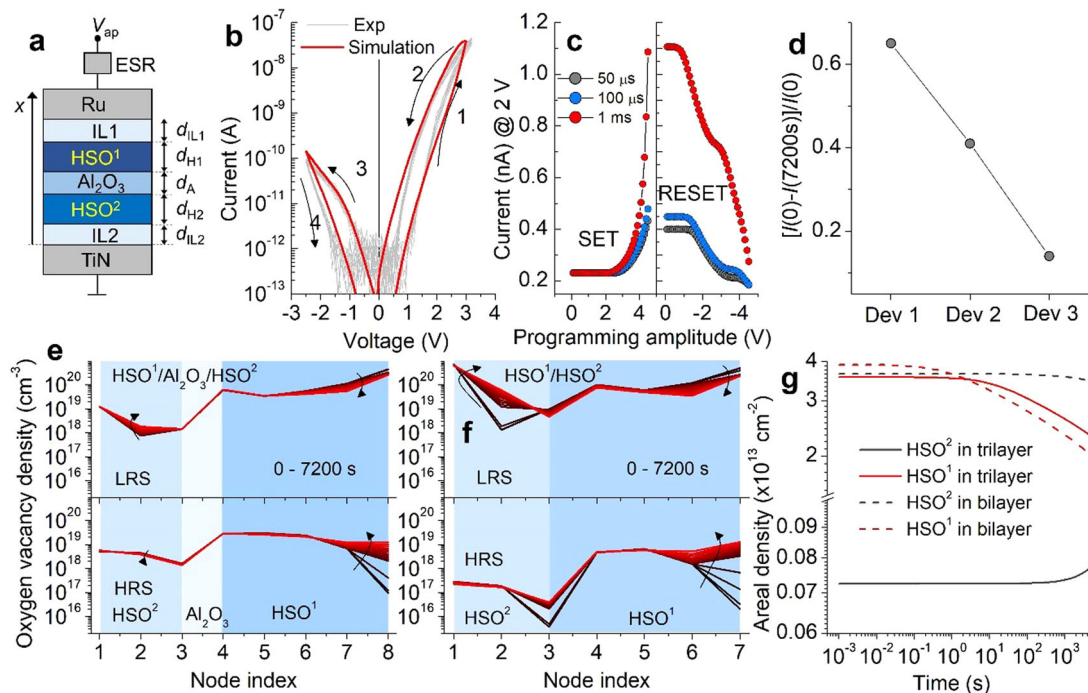


Fig. 4 Resistive switching simulation. **a** One-dimensional configuration of the SRMC for simulation. **b** Simulated I - V loop (quasi-static behavior) in comparison with experimental data. **c** Simulated switching behaviors in response to voltage pulses of different widths and amplitudes. **d** Simulated LRS retention for the HSO²-only cell (Dev 1), HSO¹/HSO² cell (Dev 2), and HSO¹/Al₂O₃/HSO² SRMC (Dev 3). **e, f** Simulated oxygen vacancy distributions in the trilayer SRMC and HSO¹/HSO² cell in the LRS (upper panel) and HRS (lower panel). The change of the distribution in each state was monitored in the time range (0–7200 s). **g** Retention of areal density of oxygen vacancies in the LRS in each layer of the trilayer and bilayer cells.

data retention and thus non-volatility. The outstanding data retention in our SRMC may be achieved by the use of a separate oxygen reservoir (HSO¹) by an oxygen-blocking layer (Al₂O₃). Figure 2c shows that it is conceivable that the HSO¹ layer may have a higher oxygen vacancy concentration than the HSO² layer, serving as an oxygen reservoir, which creates excellent data retention. Data that support this hypothesis are presented in Fig. 3c; unlike the Ru/HSO²/TiN stack, single-layer-based Ru/HSO²/TiN stack creates a severe retention issue, identifying a current decrease by 70% at 85 °C. However, the key to high data retention in our SRMC lies not only in the oxygen reservoir (HSO¹) but also in the 1-nm-thick Al₂O₃ layer between HSO¹ and HSO². A comparison between the Ru/HSO¹/Al₂O₃/HSO²/TiN and Ru/HSO¹/HSO²/TiN SRMCs shows further improvement in data retention by inserting the Al₂O₃ layer between the HSO¹ and HSO² layers (Fig. 3c). Therefore, it is believed that the thin Al₂O₃ layer hinders the depolarization of space charge (oxygen vacancy)^{41,42}.

Modeling of switching in SRMCs. To identify the role of each layer in our SRMC, we modeled our SRMC from scratch regarding oxygen vacancy dynamics in response to the applied voltage. A schematic of the one-dimensional SRMC configuration is shown in Fig. 4a. We considered the trilayer as a mixed ionic-electronic conductor with free electrons and oxygen vacancies as mobile electronic and ionic defects. The oxygen vacancy redistribution in time and space domains was fully addressed using the Fick's second law. We used the quasi-static approximation to consider electron distribution in the SRMC given the large difference in diffusion coefficient between oxygen vacancy and electron. The defining features of the model are as follows.

- The electron transport is thermally activated such that the interfacial electron transport conforms to the Schottky

emission and the bulk electron transport to the band conduction rather than localized conduction.

- HSO¹ is given a lower reference state chemical potential μ_i^0 for oxygen vacancy than HSO² to allow HSO¹ to hold a larger oxygen vacancy density than HSO² to be consistent with the experimental data in Fig. 2.
- The Al₂O₃ layer is given a lower oxygen vacancy diffusion coefficient than HSO¹ and HSO² by one order of magnitude.
- The Ru TE works as an oxygen vacancy source.
- An interfacial layer (IL) is placed at each interface, which may work as the Helmholtz layer^{39,40}.
- The breakdown of the first-order approximation (FOA) of the drift-diffusion equation is considered, which is likely the case when the internal electric field exceeds a few 10s MV/cm as for our SRMC.
- The experimentally observed asymmetry in I - V is realized by using asymmetric electrodes with different work functions ($W_{\text{Ru}} > W_{\text{TiN}}$).

The calculation procedure is elaborated in the Methods section.

In our model, the dc electronic current in the steady-state is dictated by the electronic injection current at the cathode, which conforms to the Schottky emission. That is, the injection current is determined by the interfacial electric field that lowers the Schottky barrier height (SBH) by image force. The redistribution of oxygen vacancies by programming voltage alters the interfacial electric field at both interfaces because the Debye length for the oxygen vacancy density considered is larger than the thickness of our trilayer. The relation between the interfacial electric fields and oxygen vacancy distribution is best explained using Poisson's equation.

$$\frac{dE}{dx} = \frac{q\rho(x)}{\epsilon_r\epsilon_0} \approx \frac{qC_{V_0}(x)}{\epsilon_r\epsilon_0},$$

where the space charge density ρ is approximated to the oxygen

vacancy density c_{V_0} because the free electron density is much lower than the vacancy density due to the large bandgap in the trilayer. The dielectric constant and vacuum permittivity are denoted by ϵ_r and ϵ_0 , respectively. That is, the trilayer is fully depleted. Solving the differential equation for the electric field at the bottom interface $E(0)$ or the top interface $E(d)$ yields the following equations.

$$\begin{cases} E(0) = -\frac{V_{ap}}{d} - \frac{q}{\epsilon_r \epsilon_0 d} \int_0^d \int_0^x c_{V_0}(x') dx' dx \\ E(d) = -\frac{V_{ap}}{d} + \frac{q}{\epsilon_r \epsilon_0} \int_0^d (c_{V_0}(x') - \frac{1}{d} \int_0^x c_{V_0}(x') dx') dx \end{cases}, \quad (1)$$

where d denotes the total thickness of the trilayer. From these equations, it is obvious that the change in vacancy distribution alters both interfacial electric fields. The key to non-volatile switching is that (i) set and reset switching operations cause $\Delta c_{V_0,t=0}$ ($= c_{V_0,t=0}^{LRS} - c_{V_0,t=0}^{HRS}$) sufficiently large to change $E(0)$ and $E(d)$ and (ii) the programmed distribution should be retained, $\Delta c_{V_0,t=0} \approx \Delta c_{V_0,t=\infty}$.

We took into account the breakdown of the FOA of the drift-diffusion equation in that the oxygen vacancy migration velocity exponentially increases with the electrochemical potential gradient^{43–45}. The breakdown of the FOA may be the substrate for the voltage-time dilemma⁴⁶.

As boundary conditions, the TiN/HSO² bottom interface forms oxygen-blocking contact while the Ru/HSO¹ top interface allows oxygen vacancies to move through the interface (non-blocking contact) with a constant vacancy density on the Ru side of the interface. The parameters used in our modeling are listed in Supplementary Table 1, including several key parameters, e.g., vacancy diffusion coefficients in HSO¹, HSO², and Al₂O₃^{47,48}.

The response of our model to quasi-static staircase voltage sweep (0.25 V/s) is plotted in Fig. 4b. The simulated I - V loop is well consistent with the experimental data, identifying good reproducibility of experimental data in a quasi-static voltage domain. Subsequently, we tested the response of our model to voltage pulses of different widths (50 μ s, 100 μ s, and 1 ms) and amplitude (0.1–4.5 V). The results are shown in Fig. 4c. Similar to the experimental results, the set switching behavior represents the onset of switching at ~ 3 V, so that setting read-out voltage to 2 V was allowed as for the experimental measurements. The reset switching behavior (particularly, with 1 ms reset pulses) indicates a gradual change in resistance, in agreement with the experimental data.

The excellent LRS retention for our SRMC was successfully reproduced using our model as shown in Fig. 4d. We also modeled the other cells, HSO²-only cell (Dev 1) and HSO¹/HSO² cell (Dev 2) to identify their LRS retention characteristics. Note that for the two cells we used the same parameters as the SRMC model except their thicknesses. The comparison in Fig. 4d highlights the excellent LRS retention of our SRMC model in support of the experimental data.

As such, the key to data retention is the time-dependent redistribution of oxygen vacancies in each state. Our model simulation allows us to monitor the evolution of oxygen vacancy distribution at any time step. We set and reset the model and kept track of vacancy distribution for 7200 s. The monitored results are plotted in Fig. 4e; the upper and lower panels show the distributions for the LRS and HRS, respectively. The distributions indicate (i) the lower vacancy density in HSO² than HSO¹ in both resistance states, (ii) small change in vacancy density in both states over time, i.e., small $c_{V_0,t=0} - c_{V_0,t=7200}$ in both layers, and (iii) small difference in vacancy density between LRS and HRS, i.e., small $\Delta c_{V_0,t=0}$ in HSO². Considering the indication (i), the resistance state of our model is mainly dictated by the oxygen vacancy density in HSO¹ rather than HSO². This is because the

interfacial electric fields are mainly determined by large space charge density as shown in Eq. (1); integrating the oxygen vacancy density over HSO² is much smaller than over HSO¹. The indication (ii) is the direct cause of the excellent LRS retention.

The indication (iii) is caused by the Al₂O₃ oxygen-blocking layer. The low diffusion coefficient of Al₂O₃ hinders oxygen vacancies from entering into (leaving from) HSO² during set (reset) switching. This can be seen in comparison with the HSO¹/HSO² cell whose oxygen vacancy distributions in both states are plotted in Fig. 4f. Figure 4f identifies that the lack of the oxygen-blocking layer allows a large number of oxygen vacancies to enter into (leave from) HSO², unlike the trilayer. Thus, the role of the Al₂O₃ oxygen-blocking layer in switching is to confine the active switching region to HSO¹. The better LRS retention for the trilayer than HSO¹/HSO² is understood in terms of the confined switching to HSO¹. According to Eq. (1), the larger the oxygen vacancy density, the larger the interfacial electric field evolves; the larger electric field in the vicinity of the top interface $E(d)$ drives the more oxygen vacancies back to the source (Ru). The unconfined cell (HSO¹/HSO²) holds the more oxygen vacancies over than the trilayer, and the larger $E(d)$ releases the more oxygen vacancies to the vacancy source. To show this, we evaluated the areal density of oxygen vacancies in each layer for the trilayer and HSO¹/HSO² bilayer. The areal density was calculated by integrating the vacancy density over the HSO¹ or HSO² region. The results are shown in Fig. 4g, which identifies the larger decay in oxygen vacancy density in HSO² in the bilayer than the trilayer.

Two-bit operation of SRMCs. The capability of multilevel programming was identified for four resistance levels: one HRS and three LRSs (L1, L2, and L3). The available resistance state ranges from 5 to 1 G Ω at a read-out voltage of 2 V, corresponding to 0.4–2 nA (Fig. 1b). The range was equally divided into four resistance bits, each with one of the four resistance states (0.4–0.8 nA for HRS, 0.8–1.2 nA for L1, 1.2–1.6 nA for L2, and 1.6–2 nA for L3). Each range (0.4 nA in width) was sub-divided into the available current range (0.16 nA) and forbidden range (0.24 nA) to reduce the state overlap probability (SOP) between neighboring states. We applied the incremental step pulse programming (ISPP)/error check and correction (ECC) method^{17,18} considering the available current range (0.16 nA in width) which is separated from neighboring states by the forbidden range (0.24 nA in width). The multiple resistance levels were programmed using two distinct schemes: (i) erase-and-program and (ii) erase-free schemes. The former fully erases the SRMC from the LRS before each reprogramming, whereas the latter programs a new LRS directly from its current state without the full erase process. The erase-free scheme reduces the time complexity in multilevel programming because the erase process is omitted. These methods are shown in detail in previous studies¹⁷.

To address the reliability of multilevel programming, 5 SRMCs were programmed into the four distinct resistance states at 85 $^{\circ}$ C using each programming protocol. Figure 5a identifies the multilevel operation of the five SRMCs (indexed #1–#5) using the erase-and-program scheme. Each subplot in Fig. 5a shows switching between the HRS and one of the three LRSs (L1, L2, and L3) over 50 cycles for one of the five SRMCs (#1–#5). The erase-free scheme was also applied to another set of five SRMCs subject to switching between L1 and L2, L2 and L3, and L1 and L3 over 50 cycles at 85 $^{\circ}$ C (Fig. 5b). During ISPP/ECC, the pulse amplitude required to program one of the three LRSs varied. We plotted the cumulative distribution of the pulse amplitudes for L1, L2, and L3 in Fig. 5c. The four measured resistance states were statistically analyzed to identify the SOP between the states as

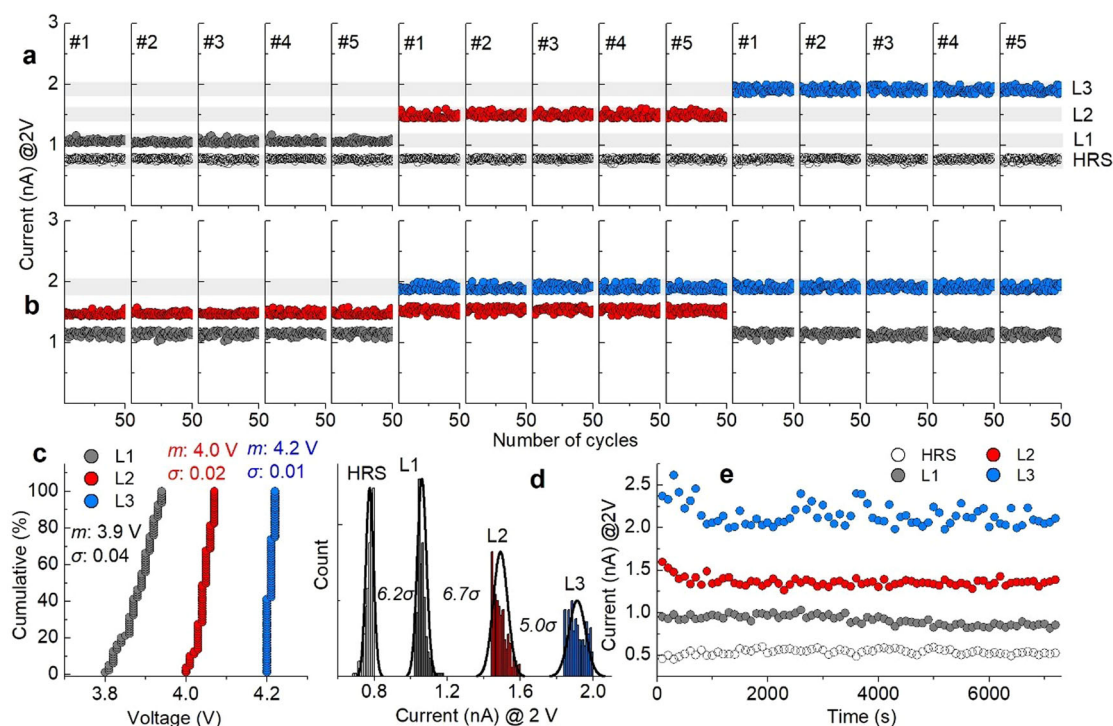


Fig. 5 Two-bit states of SRMC. Two-bit states programmed using **a** erase-and-program scheme and **b** erase-free scheme on five SRMCs (indexed #1–#5). **c** Cumulative distribution of amplitudes of two-bit programming pulses. Average amplitude and standard deviation denoted by m and σ . **d** SOP between two-bit states. **e** Retention of two-bit states at 85 °C.

shown in Fig. 5d. The minimum distance in current between neighboring states arises from L2 and L3, which are separated by 5.0σ in a Gaussian distribution. This implies a $2.86 \times 10^{-5}\%$ probability of errors in a multibit read-out. Additionally, retention of the 2-bit data is an important concern. We addressed the retention by monitoring the four resistance states at 85 °C for 2 h, yielding the profiles of read currents in Fig. 5e. The data indicate a barely noticeable change in the current level for the four states even at the elevated temperature.

SRMCs in a passive crossbar array. We fabricated a 30×30 CA of the SRMCs, each of which was fully addressable. The layouts of the CA and morphology of a single SRMC are shown in Fig. 6a, b, respectively. To address a single cell, we applied an operation voltage (V_{op}) to the cell column-line (biasing line) with the row-line (ground line) being grounded. The current was measured on the ground line. Additionally, the other column- and row-lines were pulled up to the column- and row-inhibit voltage, respectively, to suppress the sneak current. We considered two biasing schemes: half-biasing (Scheme 1) and one-third biasing (Scheme 2). When addressing a cell, Scheme 1 pulls up the biasing line and half pulls up the column- and row-inhibit-biasing lines, whereas Scheme 2 pulls up the biasing line, pulls up the column-inhibit-biasing lines one-third, and pulls up the row-inhibit-biasing lines two-thirds. Schemes 1 and 2 are summarized in Table 2.

One hundred different SRMCs in the 30×30 CA were randomly chosen to characterize their I - V loops using Schemes 1 and 2, illustrated in Fig. 6c, d, respectively. The aim was twofold: the identification of disturbance from the unchosen cells and cell-to-cell variability of switching behavior. For each scheme, three I - V loops for each of 100 cells, i.e., 300 loops in aggregate, are shown in Fig. 6e, f. First, a comparison between Fig. 6e (6f) and Fig. 1a identifies negligible effects of the 899 parallel SRMCs on the selected SRMC, leveraging their self-rectifying and highly

nonlinear I - V characteristics. Second, the appended I - V loops show negligible variability. The variability was evaluated by statistical analysis on the read-out currents (at 2 V) of the 100 cells (Fig. 6g, h for Schemes 1 and 2, respectively). The distributions for both schemes highlight good cell-to-cell uniformity in the 30×30 CA, and thus no overlap between HRS and LRS.

The excellent uniformity in the I - V loop is observed; this is due, in part, to the lack of electroforming, which is otherwise likely to endow each cell with uncontrollable random variability. Additionally, the current level in the inhibit region is comparable to the open circuit current level, implying extremely low current in the inhibit region, which is desirable when the CA size becomes large. In this CA configuration, each of the SRMCs is classified as (i) a selected cell between the biasing and ground lines, (ii) an unselected cell either between the column-inhibit-biasing and ground lines or between the biasing and row-inhibit-biasing lines, or (iii) an unselected cell between the column- and row-inhibit-biasing lines. The last two groups are named unselected groups 1 and 2, respectively. For each scheme, the theoretical voltage across an SRMC (voltage on a column-line minus voltage on a row-line) in each group is indicated in Fig. 6c, d. As shown in Fig. 6c, Scheme 1 allows half the pull-up voltage across the unselected group 1 cells (blue-filled circle) and zero voltage across the unselected group 2 cells (black-filled circle). Scheme 2 applies a positive one-third of the pull-up voltage (blue-filled circle) and negative one-third of the pull-up voltage (black-filled circle) to the unselected group 1 and 2 cells, respectively (Fig. 6d). Given a square CA ($N \times N$), the number of cells in the unselected groups 1 and 2 is proportional to N and N^2 , respectively. Thus, the larger the array, the more dominantly the unselected group 2 cells contribute to the resistance parallel to the selected cell's resistance. Although the effect of the unselected group 2 cells is barely noticeable in the 30×30 CA, it is likely to take effect in larger CAs, which is a challenge.

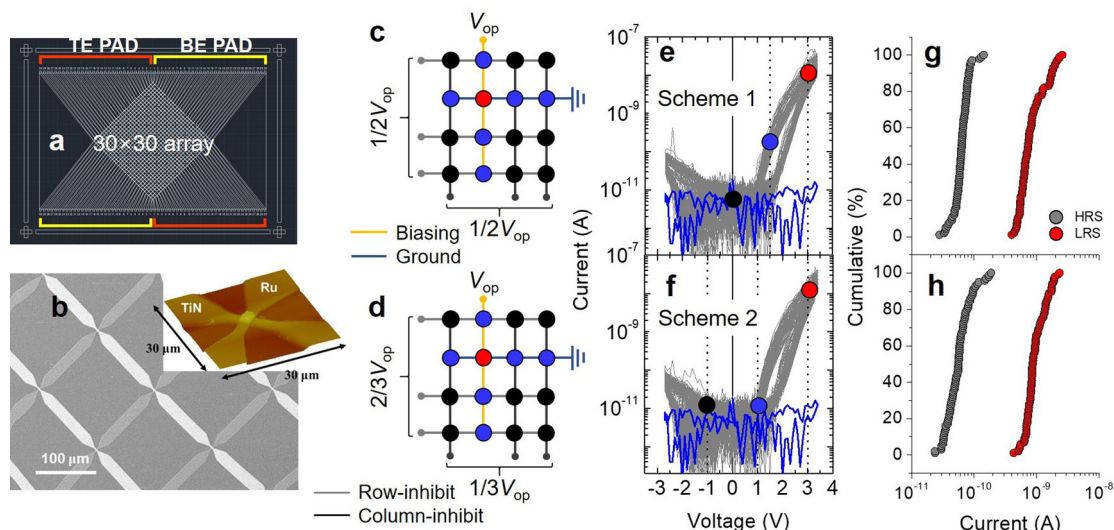


Fig. 6 30 × 30 CA of SRMCs. **a** Top view of CA layout. **b** Scanning electron microscope image of the array. The inset shows an atomic force microscope image of a unit SRMC. Schematic of **c** Scheme 1 and **d** Scheme 2. Appended *I*–*V* loops of 100 randomly chosen SRMCs (three loops for each SRMC), which were measured using **e** Scheme 1 and **f** Scheme 2. For $V_{op} = 3$ V, voltage across selected cell (red-filled circle), unselected group 1 cell (blue-filled circle), and unselected group 2 cell (black-filled circle) is indicated. Open circuit current was also plotted (blue line). The currents read using Scheme 1 and Scheme 2 on the 100 SRMCs are shown in the distributions in **g** and **h**, respectively. Scheme 1: the mean current m and standard deviation σ for HRS and LRS are $(6.5 \times 10^{-11}$ A, $1.8 \times 10^{-11})$ and $(9.3 \times 10^{-10}$ A, $5.1 \times 10^{-10})$, respectively. Scheme 2: $(6.2 \times 10^{-11}$ A, $3.2 \times 10^{-11})$ and $(1.0 \times 10^{-9}$ A, $4.1 \times 10^{-10})$ for HRS and LRS, respectively.

Table 2 Voltage across cells over a CA for Schemes 1–4.

	Scheme 1	Scheme 2	Scheme 3	Scheme 4
Biasing line voltage	V_{op}	V_{op}	V_{op}	V_{op}
Row-inhibit-line voltage	$1/2V_{op}$	$2/3V_{op}$	$1/3V_{op}$	$1/3V_{op}$
Column-inhibit-line voltage	$1/2V_{op}$	$1/3V_{op}$	$2/3V_{op}$	$1/3V_{op}$
Voltage across a selected cell	V_{op}	V_{op}	V_{op}	V_{op}
Voltage across an unselected group 1 cell	$1/2V_{op}$	$1/3V_{op}$	$2/3V_{op}$	$2/3V_{op}$
Voltage across an unselected group 2 cell	0	$-1/3V_{op}$	$1/3V_{op}$	0

To address this challenge, we investigated the *I*–*V* behavior of a predefined selected SRMC embedded in a 160×160 (~25 kb) and 320×320 (~100 kb) CA. The layouts of the CAs are shown in Supplementary Fig. 4. Note that these arrays were not random-accessible because the number of lines exceeds the number of currently available probes. Instead, we programmed the unselected group 2 cells simultaneously to LRS by leaving all row-lines (except the signal row-line) and all column-lines (except the signal column-line) common. The measurement configuration is depicted in Supplementary Fig. 5. The detail of the measurement is written in the Methods section. Schemes 1 and 2 also applied to the large CAs. For both schemes, the voltage across the selected SRMC and unselected groups 1 and 2 cells is indicated on the *I*–*V* loop taken from Fig. 1a in Fig. 7a, b. Two additional biasing schemes (Schemes 3 and 4) were considered for comparison. Scheme 3 applies a one-third V_{op} and two-thirds V_{op} to the row- and column-inhibit lines, respectively. Therefore, the voltage across the unselected group 1 and 2 cells is two-thirds V_{op} and one-third V_{op} , respectively (Fig. 7c). Scheme 4 applies a one-third V_{op} to both the row- and column-inhibit lines, allowing one-third or two-thirds V_{op} across the unselected group 1 cell and zero voltage across the unselected group 2 cell (Fig. 7d). The details of Schemes 3 and 4 are summarized in Table 2.

To examine the sneak current from the unselected groups 1 and 2 cells, we attempted to program all unselected groups 1 and

2 cells into their LRS and subsequently examined the *I*–*V* characteristics of the selected cell. The measured *I*–*V* loops for the selected cell embedded in the 160×160 CA are shown in Fig. 7e. The different biasing schemes caused a negligible difference in the *I*–*V* loops of the selected cell. This confirms that the self-rectifying and nonlinear *I*–*V* behavior of the SRMC maintains a sufficiently low current through the unselected cells to enable the true current to be read through the selected cell.

The 320×320 CA allows the sneak current to vary the ground line current more obviously than the smaller CAs, yielding more obviously distinct *I*–*V* loops depending on the voltage-application scheme (Fig. 7f). Scheme 2 yields a lower current than Scheme 1 over the whole voltage range, whereas the largest current level was yielded by Schemes 3 and 4. This is because Scheme 2 applies the lowest voltage to unselected group 1 cells, which share the same row-line as the selected cell. Nevertheless, the switching behavior of the selected cell indicates two distinct states despite the sneak current in this seemingly worst-case conductance distribution.

Two-bit operation of SRMCs in the CAs was examined successfully. Owing to the random-accessibility of the 30×30 CA, five SRMCs were chosen randomly and subject to the two-bit operation, resulting in readable four states (Supplementary Fig. 6) as for the single cells. Additionally, we identified the two-bit operation of the predefined SRMC in the 320×320 CA, yielding clearly distinct four states (Supplementary Fig. 7).

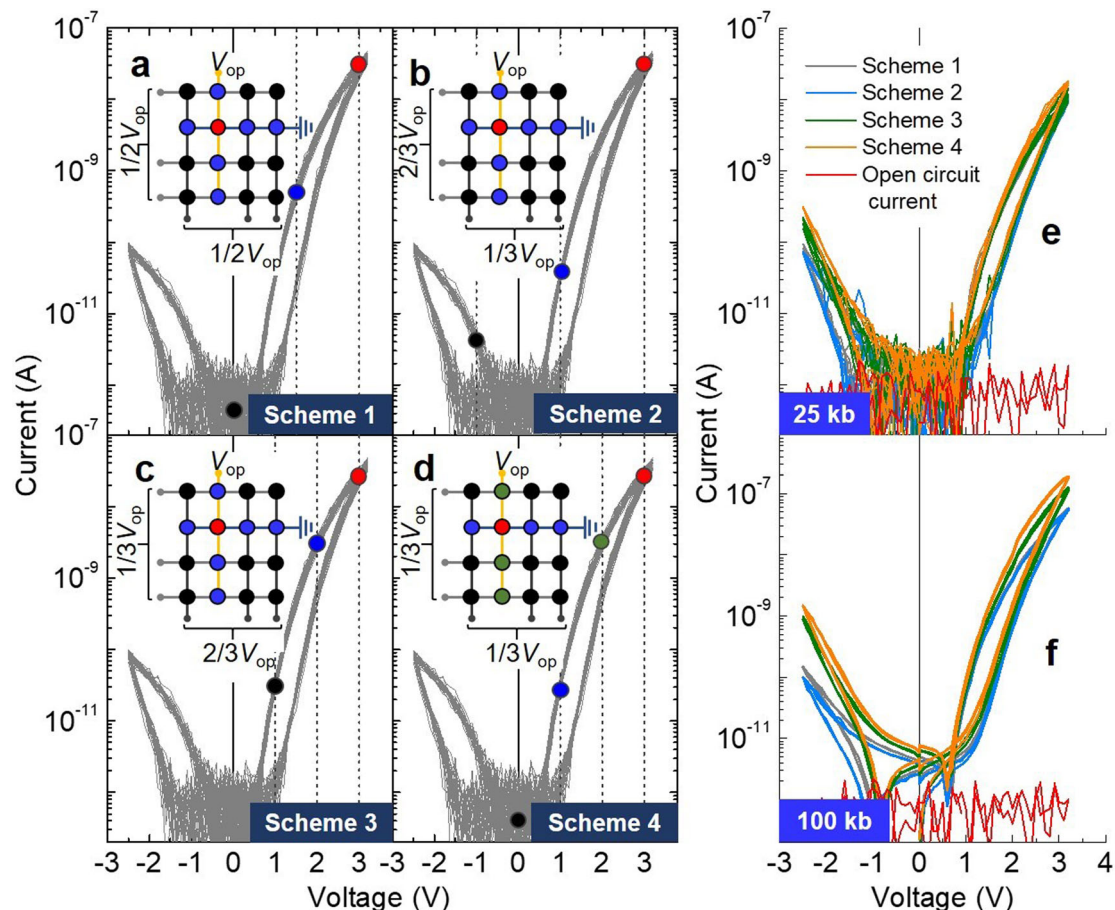


Fig. 7 160×160 and 320×320 CAs of SRMCs. **a-d** Illustrations of Schemes 1-4 and voltage across different cells indicated by different colors. *I-V* loops of selected cell that was embedded in **e** 160×160 and **f** 320×320 CA.

Vector-matrix multiplication acceleration using the 30×30 crossbar array. Finally, we identified the feasible acceleration of vector-matrix multiplication ($\mathbf{w} \times \mathbf{x}$; $\mathbf{w} \in \mathbb{Z}^{30 \times 30}$, $\mathbf{x} \in \mathbb{Z}^{30}$) by reducing the computational complexity to $O(n)$. To this end, we aimed to calculate a dot product $\mathbf{w}[i, :] \cdot \mathbf{x}$ at one cycle, where $\mathbf{w}[i, :]$ denotes the i th row of matrix \mathbf{w} . We restricted the elements w of matrix \mathbf{w} to 2-bit integers ($w \in \{0, 1, 2, 3\}$) and the elements x to 1-bit integers ($x \in \{0, 1\}$). The matrix \mathbf{w} was transposed and mapped onto our 30×30 SRMC CA (conductance of each cell $\in \{\text{HRS}, L1, L2, L3\}$). The vector \mathbf{x} with 1-bit integer elements was encoded as a voltage array \mathbf{V}_{ap} ($V_{\text{ap}} \in \{0, 2V\}$) and applied to the 30 row-lines of the CA (Fig. 8a). The current measured at the i th column-line was the intermediate result of the dot product $\mathbf{w}[i, :] \cdot \mathbf{x}$. As depicted in Fig. 8b, we addressed one column at one cycle by pulling down the chosen column-line to the ground while inhibit voltages (V_{inhibit}) were applied to the rest of column-lines ($2/3V_{\text{ap}}$), so that we reduced the complexity to $O(n)$, which is otherwise $O(n^2)$.

We chose four random matrices (w_1, w_2, w_3 , and w_4) of different sparsities (0, 25, 51, and 55%, respectively). The percentage of each integer (0, 1, 2, 3) in each matrix is shown in Fig. 8c. The chosen matrices were mapped onto four 30×30 CAs such that the individual cells of the CAs were randomly accessed and programmed to the correct conductance states using Scheme 2. The programmed conductance map for each matrix is shown in Fig. 8d-g. The conductance of each SRMC was individually read out at a read-out voltage of 2 V to acquire the maps. We then performed the dot product $\mathbf{w}[i, :] \cdot \mathbf{x}$ for each i at

one cycle with vector \mathbf{x} of ones, i.e., $\mathbf{x} = [1, 1, \dots, 1]$. The vector-matrix multiplication operation for each matrix thus consumes 30 column-line-addressing cycles, yielding a current vector \mathbf{j} ($\in \mathbb{R}^{30}$) as the intermediate product (Fig. 8d-g). The measured current at each column-line is almost identical to the current value extrapolated from each cell current in the same column, indicating marginal disturbance from the unselected cells. For the multiplication with four matrices (w_1, w_2, w_3 , and w_4), the CA domain consumes powers of 4.22, 3.44, 2.83, and 2.69 μW , respectively. The considered multiplication is the worst case in terms of power consumption because of the extremely dense vector \mathbf{x} (of ones).

To output the final product \mathbf{z} ($\mathbf{z} = \mathbf{w} \times \mathbf{x}$; $\mathbf{z} \in \mathbb{Z}^{30}$), current from the i th column j_i for all i needs to be encoded as a binary number, which subsequently enters into the near-memory digital domain for additional processing. A common method is to convert the summed current to voltage and subsequently to quantize the converted voltage using an analog-to-digital converter (ADC)⁴⁹. Alternatively, the summed current can directly be converted to a binary value using a current sense amplifier (CSA) with multiple reference currents that are iteratively compared with the summed current⁵⁰. In either way, the important consideration is twofold: (i) energy consumption and (ii) bit-width of the product \mathbf{z} . Regarding power consumption, ADCs are well known to consume a considerable amount of energy inasmuch as the total energy consumption of an RRAM-based inference accelerator is dominated by the ADCs¹³. An alternative method using a CSA⁵⁰ keeps the static current from the chosen line flowing while the

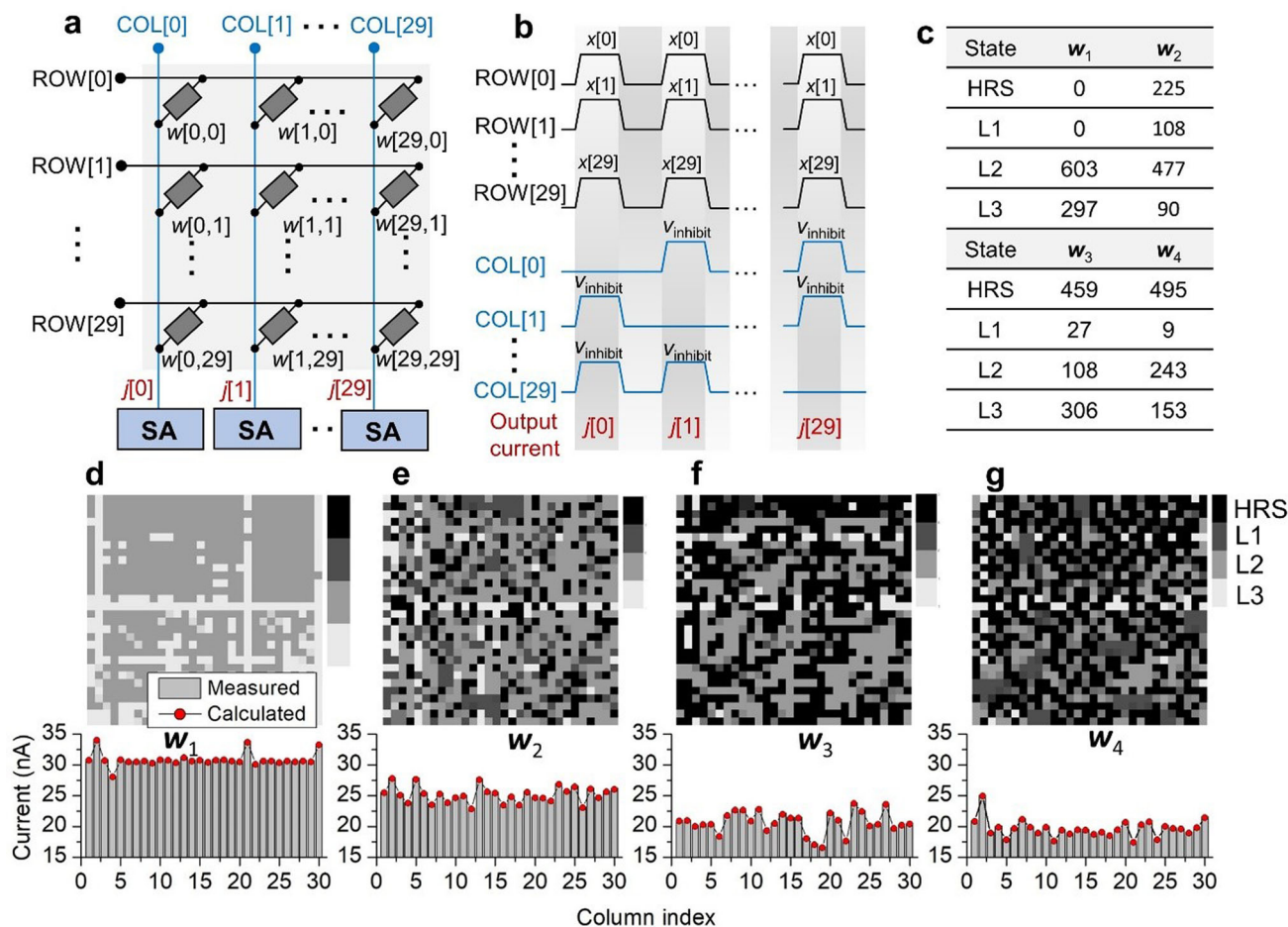


Fig. 8 Acceleration of vector-matrix multiplication using the 30 × 30 CA. **a** Configuration of a 30 × 30 matrix w mapped onto a CA of the same size. Vector x is encoded as voltage signals ('0' = 0 V, '1' = 2 V) and enters into the row-lines (ROW[0]-ROW[29]). The resulting current vector j as an intermediate product enters into sense amplifiers (SAs) to be quantized. **b** Schematic timing diagrams of row- and column-line signals. The inhibit voltages applied to unchosen column-lines are denoted by $V_{inhibit}$. **c** Statistics of four states (HRS, L1, L2, L3) in four random matrices (w_1 - w_4). **d-g** (upper panel) Conductance maps of the four random matrices (w_1 - w_4) and (lower panel) measured current vectors j for the four matrices. We considered a vector x of ones. The measurement results are compared with the calculated current vectors using the measured currents on individual cells.

summed current being converted iteratively, causing additional energy consumption. The bit-width should be chosen carefully to avoid the performance, i.e., inference, degradation by the quantization bit-width. As shown in quantized neural networks such as DoReFa-Net⁵¹, the resolution, i.e., bit-width, of activations more critically determines the inference accuracy than that of weights. The activation resolution is dictated by the bit-width of the output z . Therefore, the bit-width of the product z is an important consideration in the design of summed current-encoding circuits.

Regarding multibit factor x , time-division multiplexing is a desirable method by encoding the vector x as shown in Fig. 9. Because of the nonlinear I - V behavior in the LRS of our SRMCs, encoding a factor as input voltage amplitude is unsuitable unlike linear I - V cases^{14,52}. The l -bit elements $x[i]$ are time-division multiplexed from the least significant bits (LSBs) to the most significant bits (MSBs) and are applied to the row-lines at one column-line addressing cycle for the dot product $w[i, :] \cdot x$. Thus, each dot product cycle includes l sub-cycles. The output current at each sub-cycle is encoded as a binary value and subsequently multiplied by 2^{k-1} , where k denotes the digit corresponding to the sub-cycle. The results are finally summed to output the dot product $w[i, :] \cdot x$.

Discussion

We proposed an SRMC based on a $Hf_{0.8}Si_{0.2}O_2/Al_2O_3/Hf_{0.5}Si_{0.5}O_2$ trilayer stack, which highlights large selectivity ($\sim 10^4$), two-bit operation, low read power (4 nW for LRS and 0.8 nW for HRS), read latency ($< 10 \mu s$), excellent data retention ($> 10^4 s$ at 85 °C), and CMOS compatibility (maximum supply voltage $\leq 5 V$). Particularly, the large selectivity due to the high asymmetry and nonlinearity in the I - V behavior potentially supports high-density passive CAs of the SRMCs, which is one of the key elements to memory-centric computing in support of deep learning acceleration. Feasibility was identified in 30×30 , 160×160 , and 320×320 arrays of our SRMCs. The I - V behavior of an isolated SRMC was reproduced well in the arrays without significant effects on the unselected cells. These excellent characteristics may be attributed to nonfilamentary switching, i.e., switching on the grounds of Schottky barrier modulation at the cathode, which is homogeneous over the device area. A common issue of such nonfilamentary switching is data retention due to the rapid depolarization of point defects, which was overcome by using the engineered trilayer switching stack in this study. Furthermore, the low programming power (ca. 18 nW), latency (100 μs), and endurance ($> 10^6$) highlight the energy-efficiency and highly reliable random-access memory of our SRMC.

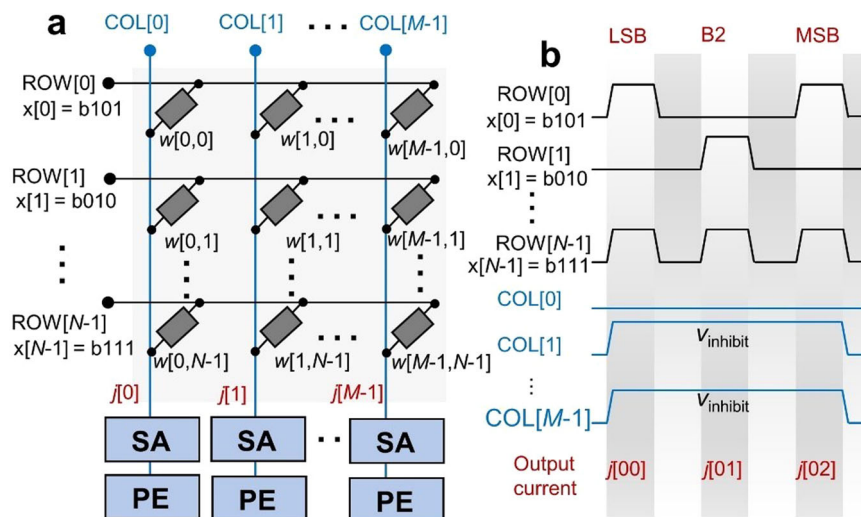


Fig. 9 Acceleration of multibit vector-matrix multiplication. **a** Configuration of a mapped weight matrix w ($M \times N$) and multibit vector x (here, 3-bit). Elements $x[i]$ are time-multiplexed, so that the multiplication delay is proportional to the bit-width of elements $x[i]$. **b** Timing diagrams of the signals to calculate $w[:,i] \cdot x$ for a given i with multibit elements including $x[0]$ ($=b101$), $x[1]$ ($=b010$), and $x[29]$ ($=b111$). The resulting currents in the three time divisions ($j[00]$, $j[01]$, and $j[02]$) are first quantized by the SAs and subsequently multiplied by 1 , 2^1 , and 2^2 , respectively, and summed in the processing elements (PEs).

Ideally, CAs may achieve the ultimate complexity $O(1)$ of vector-matrix multiplication beyond the complexity $O(n)$ by addressing all column-lines at one cycle. The basic premise is that all non-ideal factors, e.g., sneak current and line resistance effects, are excluded. The sneak current effect may be marginal because all bit-cells are supposed to be non-negatively biased when all column-lines are simultaneously addressed, i.e., grounded. However, the effect of finite line resistance is significant. The finite line resistance causes the inhomogeneous distribution of bit-cell voltages over the cells on the same row-line such that the further a bit-cell from the row-line contact, the lower voltage is applied across the bit-cell. Further, this effect is boosted when the bit-cells on the same row-line allow simultaneous current flow, which is the case of all column-line addressing.

Additionally, simultaneously addressing all column-lines requires one CSA and following logic circuit per column line, whereas addressing one column-line at a cycle allows one CSA to be shared among a group of column-lines through time-division multiplexing. This additional peripheral circuit-area overhead can be prohibitive in large-scale CAs. Therefore, the complexity reduction to $O(1)$ may be realized only when these challenges are overcome.

Methods

Device fabrication. A 200-nm-thick TiN layer was sputtered on a SiO_2/Si substrate and patterned to a shape of crossbar-type BE. The TiN BE was patterned by a conventional photolithography and dry-etching process by an inductively coupled plasma reactive ion etching (ICP-RIE). An ICP power of 200 W and a substrate bias power of 20 W were maintained during TiN etching. During the dry-etching process, the reactant gas flow rates of Ar and Cl_2 were maintained at 5 standard cubic centimeters per minute (sccm) and 30 sccm, respectively. Furthermore, the process temperature was maintained at 25 °C by a water-circulation cooling system. The observed etching rate was ~ 70 nm/min. The residual photo-resist (PR) on the patterned TiN BE was removed by an acetone etchant and cleaned sequentially with isopropyl alcohol and deionized water. The 1-nm-thickness HSO^1 and 2-nm-thickness HSO^2 thin films layer were then deposited by traveling wave-type ALD at 250 °C using a tetrakis-ethylmethyldiamido hafnium (TEMA-Hf) and bis(diethylamino)silane (BDEAS) precursor, respectively, and H_2O and O_2 plasma as a source of Hf and Si oxidant, respectively. To form each of HSO^1 and HSO^2 layers, the super-cycle ALD processes with HfO_2 and SiO_2 layers were used. During the deposition of HSO^1 and HSO^2 thin film, the ALD cycle ratios of 1:3 and 1:1 for $\text{SiO}_2/\text{HfO}_2$ were set, respectively. Between the HSO^1 and HSO^2 layers, the Al_2O_3 thin film layer was deposited by traveling wave-type ALD at 150 °C using a trimethyl aluminum (TMA) and H_2O as a source of Al and oxidant, respectively.

Subsequently, the crossbar-type TE pattern was formed by photolithography and then the 100-nm-thick Ru layer was deposited by DC magnetron sputtering. Finally, through the conventional lift-off process, the Ru/ $\text{HSO}^1/\text{Al}_2\text{O}_3/\text{HSO}^2/\text{TiN}$ stacked device was fabricated. All of the unit and CA devices have identical fabrication processes.

Structural analysis of SRMC device. The sample for TEM analysis was prepared by a focused ion beam (FIB, Helios NanoLabTM by FEI) operation. HR-TEM (Tecna G2 F30 S-TWIN by FEI) analysis was then performed to obtain a cross-sectional view of the Ru/ $\text{HSO}^1/\text{Al}_2\text{O}_3/\text{HSO}^2/\text{TiN}$ stacked RS device. The XPS analysis was performed to examine the chemical binding status of the HSO^1 and HSO^2 layer with an X-ray photoelectron spectrometer (XPS, Thermo Fisher Scientific Inc.) using an Al $K\alpha$ source with a spot size of 400 μm and energy step size of 0.1 eV. The samples for XPS analysis were prepared using blanket-type HSO thin films on a TiN substrate. It should be noted that because the thicknesses of HSO^1 and HSO^2 in the device are very thin, additional samples were prepared for XPS analysis. The elemental depth profile was obtained using AES (ULVAC-PHI 700, coaxial full CMA type analyzer, 10 kV/10 nA of electron beam energy) measurements. During the AES measurement, a sputtering rate of ~ 0.2 $\text{\AA}/\text{s}$ was maintained. RBS (National Electrostatics Corp.) analysis for qualified chemical composition of our active layers was performed with separately prepared HSO^1 and HSO^2 . (50-nm-thick each on SiO_2/Si substrate) AFM (Digital Instruments Dimension 3000, Veeco Science) analysis was conducted to observe the CA device morphology and determine the exact feature sizes of the RS device. The CAs were observed using a scanning electron microscope (SEM, JEOL JSM-6700F).

Electrical measurements. The resistive switching characteristic of the device was measured using an HP4145B semiconductor parameter analyzer in the I - V sweep mode. Measuring temperature was controlled by a hot stage using a temperature controller. The pulse-based electrical measurements were conducted using an HP4145B, arbitrary function generator (Agilent 81150 A), oscilloscope (MSOX3024T, Tektronix), and electromechanical radiofrequency electric-circuit switching box. Throughout the measurement processes, the voltage was biased to the Ru TE, while the TiN BE was electrically grounded. The resistance (or current) values of the programming and erasing were verified at 2 V using the SPA. Measuring the 2-bit RS operation, ISPP/ECC algorithms were performed using two convertible electrical circuits composed of [AFG-RS device-OSC] and [SPA-RS device], respectively. These two types of electrical circuits were approached alternately by the electromechanical RF electrical circuit switching boxes. In the random-access operation (Fig. 6), the 30×30 sized matrix switching zig was additionally equipped in the previous electrical circuit. All electrical measurements were performed using a LabVIEWTM-based control program.

Modeling of resistive switching dynamics. We modeled the resistive switching behaviors of our SRMC regarding oxygen vacancy dynamics in response to the applied voltage. The one-dimensional model configuration considered is shown in Fig. 4a. The SRMC consists of five layers, $\text{HSO}^1/\text{Al}_2\text{O}_3/\text{HSO}^2$ plus two interfacial dipole layers between HSO^1 and TE (IL_1) and HSO^2 and BE (IL_2).

41. Traoré, B. et al. On the origin of low-resistance state retention failure in HfO₂-based RRAM and impact of doping/alloying. *IEEE Trans. Electron Devices* **62**, 4029–4036 (2015).
42. Traoré, B. et al. Microscopic understanding of the low resistance state retention in HfO₂ and HfAlO based RRAM. *IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2014).
43. Jeong, D. S., Kin, I., Lee, T. S., Lee, W. S. & Lee, K. S. Electric-field-enhanced ionic diffusivity in electrolytes: a model study. *J. Korean Phys. Soc.* **61**, 913–919 (2012).
44. Meuffels, P. & Schroeder, H. Comment on “Exponential ionic drift: fast switching and low volatility of thin-film memristors” by D.B. Strukov and R.S. Williams in *Appl. Phys. A* (2009) 94: 515–519. *Appl. Phys. A* **105**, 65–67 (2011).
45. Noman, M., Jiang, W., Salvador, P. A., Skowronski, M. & Bain, J. A. Computational investigations into the operating window for memristive devices based on homogeneous ionic motion. *Appl. Phys. A* **102**, 877–883 (2011).
46. Waser, R., Dittmann, R., Staikov, G. & Szot, K. Redox-based resistive switching memories—nanoionic mechanisms, prospects, and challenges. *Adv. Mater.* **21**, 2632–2663 (2009).
47. Sufi, Z., Hemanth, J., Lisa, F. E. & Devendra, G. Measurement of oxygen diffusion in nanometer scale HfO₂ gate dielectric films. *Appl. Phys. Lett.* **98**, 152903 (2011).
48. Nakamura, R. et al. Diffusion of oxygen in amorphous Al₂O₃, Ta₂O₅, and Nb₂O₅. *J. Appl. Phys.* **116**, 033504 (2014).
49. Walden, R. H. Analogue-to-digital converter survey and analysis. *IEEE J. Sel. Areas Commun.* **17**, 539–550 (1999).
50. Chen, W. H. et al. CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors. *Nat. Electron.* **2**, 420–428 (2019).
51. Zhou, S., et al. DoReFa-Net: training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv: 1606.06160v3* (2018).
52. Hu, M et al. Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication. *53rd ACM/EDAC/IEEE Design Automation Conference (DAC)* (IEEE, 2016).

Acknowledgements

G.H.K. would like to acknowledge a Korea Research Institute of Chemical Technology grant (Grant no. SS2021-20; Development of smart chemical materials for IoT devices). This work was partly supported by a research grant from the National Research Foundation of Korea under Grant no. NRF-2019R1C1C1009810. This research was also supported by the Ministry of Trade, Industry & Energy (grant number 20012002) and Korea Semiconductor Research Consortium program for the development of future semiconductor devices.

Author contributions

K.J. performed the device fabrication and electrical characterization. J.K. and C.S. conducted the array characterization. J.J.R. and S.-J.Y. conducted the thin film deposition and chemical composition analysis. M.K.Y. presented the technical discussion with regard to the electrical and materials characteristics. D.S.J. performed the SRMC modeling and characterization. D.S.J. and G.H.K. supervised all experiments and compiled the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-23180-2>.

Correspondence and requests for materials should be addressed to D.S.J. or G.H.K.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021