Article

Thousands of conductance levels in memristors integrated on CMOS

https://doi.org/10.1038/s41586-023-05759-5

Received: 7 August 2022

Accepted: 25 January 2023

Published online: 29 March 2023

Check for updates

Mingyi Rao[1,2,5], Hao Tang[3,5], Jiangbin Wu[4,5], Wenhao Song[4,5], Max Zhang[1], Wenbo Yin[1], Ye Zhuo[4], Fatemeh Kiani[2], Benjamin Chen[2], Xiangqi Jiang[1], Hefei Liu[4], Hung-Yu Chen[4], Rivu Midya[2], Fan Ye[2], Hao Jiang[2], Zhongrui Wang[2], Mingche Wu[1], Miao Hu[1], Han Wang[4], Qiangfei Xia[1,2], Ning Ge[1], Ju Li[3] & J. Joshua Yang[1,2,4] ✉

Neural networks based on memristive devices[1–3] have the ability to improve throughput and energy efficiency for machine learning[4,5] and artificial intelligence[6], especially in edge applications[7–21]. Because training a neural network model from scratch is costly in terms of hardware resources, time and energy, it is impractical to do it individually on billions of memristive neural networks distributed at the edge. A practical approach would be to download the synaptic weights obtained from the cloud training and program them directly into memristors for the commercialization of edge applications. Some post-tuning in memristor conductance could be done afterwards or during applications to adapt to specific situations. Therefore, in neural network applications, memristors require high-precision programmability to guarantee uniform and accurate performance across a large number of memristive networks[22–28]. This requires many distinguishable conductance levels on each memristive device, not only laboratory-made devices but also devices fabricated in factories. Analog memristors with many conductance states also benefit other applications, such as neural network training, scientific computing and even 'mortal computing'[25,29,30]. Here we report 2,048 conductance levels achieved with memristors in fully integrated chips with 256 × 256 memristor arrays monolithically integrated on complementary metal–oxide–semiconductor (CMOS) circuits in a commercial foundry. We have identified the underlying physics that previously limited the number of conductance levels that could be achieved in memristors and developed electrical operation protocols to avoid such limitations. These results provide insights into the fundamental understanding of the microscopic picture of memristive switching as well as approaches to enable high-precision memristors for various applications.

Memristive-switching devices are known for their relatively large dynamical range of conductance, which can lead to a large number of discrete conductance levels. Different approaches have been developed to accurately program the devices[31]. However, only devices with fewer than 200 conductance levels have been reported so far[22,32]. There are no forbidden conductance states in the dynamical range of the device because a memristor is analog and can, in principle, achieve an infinite number of conductance levels. However, the fluctuation commonly observed at each conductance level (Fig. 1e) limits the number of distinguishable levels that can be achieved in a specific conductance range. We found that such fluctuation can be substantially suppressed, as shown in Fig. 1e,f, by applying appropriate electrical stimuli (called 'denoising' processes). Notably, this denoising process does not require any extra circuitry beyond the usual read-and-program circuits. We incorporated the denoising process into device-tuning algorithms and successfully programmed a memristor made in a standard commercial foundry (Fig. 1a–d) into 2,048 conductance levels (Fig. 1g), corresponding to a resolution of 11 bits. Conductive atomic force microscopy (C-AFM) was used to visualize the evolution of conduction channels during programming and denoising processes. We discovered that a normal switching operation (set or reset) always ends up with some incomplete conduction channels, which appear as islands or blurry edges along the main conduction channel and are less stable than the main conduction channel. First-principles calculations indicate that these incomplete channels are unstable phase boundaries with dopant levels in a range that is sensitive to nearby trapped charges, contributing to the large fluctuations of each conductance level. We showed, experimentally and theoretically, that an appropriate voltage in the denoising process either annihilates (weakens) or completes (enhances) these incomplete channels, resulting in a strong reduction in fluctuation and a

[1]TetraMem, Fremont, CA, USA. [2]Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA, USA. [3]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA. [5]These authors contributed equally: Mingyi Rao, Hao Tang, Jiangbin Wu, Wenhao Song ✉e-mail: jjoshuay@usc.edu

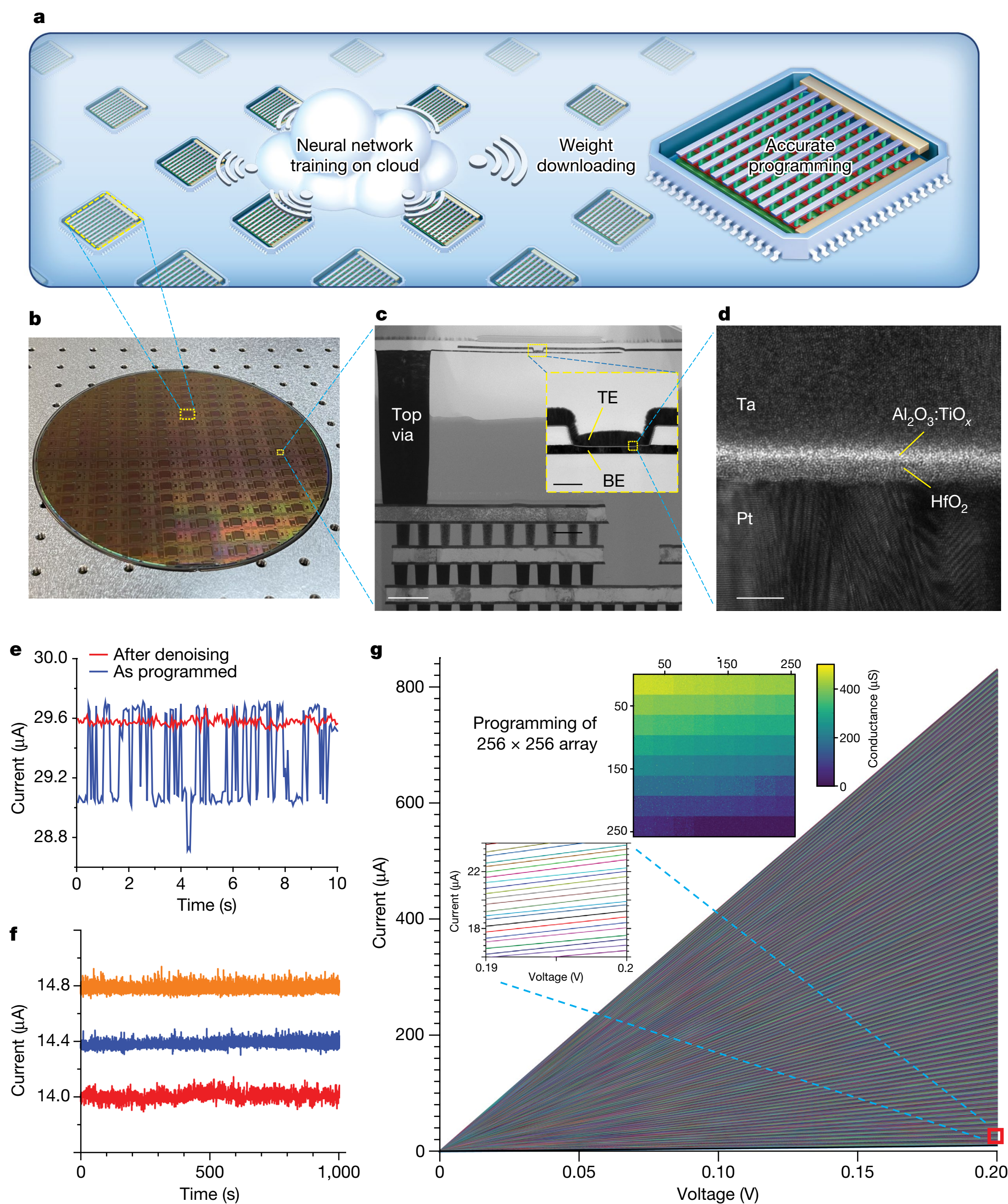

Fig. 1 | High-precision memristor for neuromorphic computing. a, Proposed scheme of the large-scale application of memristive neural networks for edge computing. Neural network training is performed in the cloud. The obtained weights are downloaded and accurately programmed into a massive number of memristor arrays distributed at the edge, which imposes high-precision requirements on memristive devices. **b**, An eight-inch wafer with memristors fabricated by a commercial semiconductor manufacturer. **c**, High-resolution transmission electron microscopy image of the cross-section view of a memristor. Pt and Ta serve as the bottom electrode (BE) and top electrode (TE), respectively. Scale bars, 1 μm and 100 nm (inset). **d**, Magnification of the memristor material stack. Scale bar, 5 nm. **e**, As-programmed (blue) and after-denoising (red) currents of a memristor are read by a constant voltage (0.2 V). The denoising process eliminated the large-amplitude RTN observed in the as-programmed state (see Methods). **f**, Magnification of three nearest-neighbour states after denoising. The current of each state was read by a constant voltage (0.2 V). No large-amplitude RTN was observed, and all of the states can be clearly distinguished. **g**, An individual memristor on the chip was tuned into 2,048 resistance levels by high-resolution off-chip driving circuitry, and each resistance level was read by a d.c. voltage sweeping from 0 to 0.2 V. The target resistance was set from 50 μS to 4,144 μS with a 2-μS interval between neighbouring levels. All readings at 0.2 V are less than 1 μS from the target conductance. Bottom inset, magnification of the resistance levels. Top inset, experimental results of an entire 256 × 256 array programmed by its 6-bit on-chip circuitry into 64 32 × 32 blocks, and each block is programmed into one of the 64 conductance levels. Each of the 256 × 256 memristors has been previously switched over one million cycles, demonstrating the high endurance and robustness of the devices.

substantial increase in memristor precision. The observed phenomena generally exist in a memristive-switching process with localized conduction channels, and the insights can be applied to most memristive systems for scientific understanding and technological applications.

Conductance levels and arrays on integrated chips

Memristors used in this study were fabricated on an eight-inch wafer by a commercial semiconductor manufacturer (Fig. 1b). Details about the

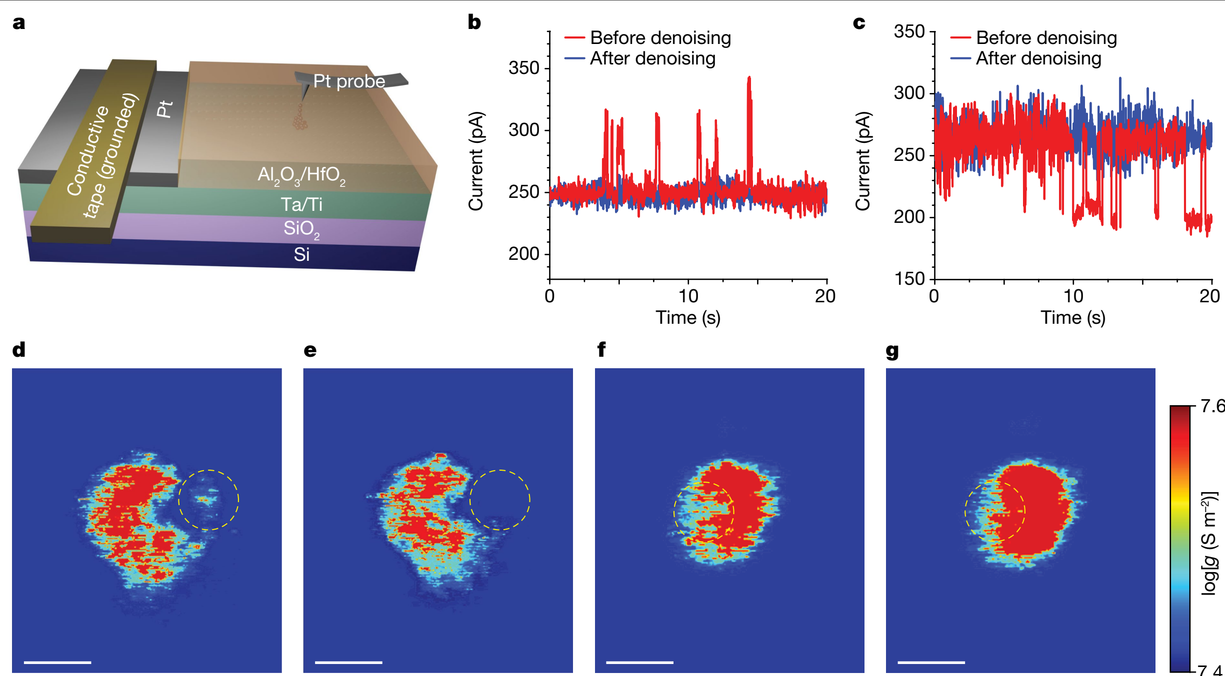

Fig. 2 | Direct observation of the evolution of conduction channels in the denoising process using C-AFM. a, Schematic of the customized memristor structure and C-AFM testing set-up. A C-AFM probe was used as the top electrode in the customized device. Because Ta easily oxidizes in air and is not a practical probe material, a Pt probe was used. This Pt probe had the same purpose as that of the bottom Pt electrode of the standard memristor that we used. To maintain the material stack of a standard memristor, the customized memristor has a reversed structure. **b**, Current readings at 0.1 V before (red) and after (blue) a denoising process using a subthreshold reset voltage. **c**, Current readings at 0.1 V before (red) and after (blue) a denoising process using a subthreshold set voltage. **d**, Conductance map measured by C-AFM scanning corresponding to the before-denoising state (red) in **b**. **e**, Conductance map corresponding to the after-denoising state (blue) in **b**. **f**, Conductance map measured by C-AFM scanning corresponding to the before-denoising state (red) in **c**. **g**, Conductance map corresponding to the after-denoising state (blue) in **c**. The dashed yellow circles in **d**–**g** highlight the changes observed before and after the denoising process. Scale bars, 10 nm.

fabrication process are provided in the Methods. Cross-section views of a memristor are shown in Fig. 1c, and the crucial resistive switching layers are magnified in Fig. 1d. The elemental image produced by electron energy-loss spectroscopy is shown in Supplementary Fig. 1. The device, which consists of a Pt bottom electrode, a Ti/Ta top electrode and a HfO_2/Al_2O_3 bilayer, was fabricated in a 240-nm via above the CMOS peripheral circuitry. The Al_2O_3 and Ti layers are designed to be thin (<1 nm) so that they seem as a mixed layer rather than two separate continuous layers. When the bottom electrode is grounded, the device can be switched by applying either a sufficiently positive voltage (for set) or a negative voltage (for reset) to the top electrode. The fluctuation level (characterized by the standard deviation of a measured current under a constant voltage) after a set or a reset operation is distributed in a wide range (Supplementary Fig. 2). The result indicates that an as-programmed state typically has large fluctuations. This considerably limits the applications of memristors, but is a characteristic of memristive materials more generally[33–36]. The data also show that a set operation tends to induce a larger fluctuation in an as-programmed state than does a reset operation. Such reading fluctuations mainly consist of random telegraph noise (RTN), which typically has step-like transitions between two or more current levels at random time points under a constant reading voltage. Such RTN generally exists in memristors. Even fluctuations that do not seem step-like may in fact be made of a RTN[37], which can be shown only when the measurement sampling rate is higher than the RTN frequency, as shown in Supplementary Fig. 3. It has been demonstrated previously by simulations that memristor RTN may be caused by charges occasionally trapping into certain defects and blocking conduction channels because of Coulomb screening[34,38]. However, experiments that directly link trapped charges, conduction channel(s) and RTN, and how to remove it, are missing. Although this is a critical issue for memristors in general, it has been unclear how to reduce the RTN in memristors. These experiments are important not only for understanding the physical origin of memristor RTN but also for revealing the entire microscopy picture of memristive switching and providing possible solutions to high-precision memristors.

We discovered that the fluctuation level could be greatly reduced by applying small voltage pulses with optimized amplitude and width. An example is given in Fig. 1e, in which an as-programmed state with a considerable fluctuation (blue) was stabilized into a low-fluctuation state (red) by denoising pulses. Using a three-level feedback algorithm devised to denoise, as shown in Supplementary Fig. 4, a single memristor was tuned into 2,048 conductance states between 50 and 4,144 μS, with a 2-μS interval between every two neighbouring states. All states were read by a voltage sweeping from 0 to 0.2 V, as shown in Fig. 1g. The bottom inset to Fig. 1g shows magnification of the current–voltage curves, which show the well-distinguishable states and the marked linearity of each state. Three nearest-neighbour states after denoising are shown in Fig. 1f, in which a constant voltage of 0.2 V reads each state for 1,000 s. The current fluctuation of every state is within 0.4 μA, corresponding to 2 μS in conductance. No significant overlap was observed in the neighbouring states. A magnification of the measurement at high-conductance states is shown in Supplementary Fig. 5. Memristors from multiple chips of an 8-inch wafer were measured, demonstrating considerable programming uniformity across the entire wafer, as shown in Supplementary Fig. 6. We further used the denoising process in the array-level programming of an entire 256 × 256 array using the on-chip circuitry. The experimentally programmed patterns are shown in Fig. 1g (top inset) and Supplementary Fig. 7. For demonstrations using the on-chip circuitry, the programming precision was limited by the precision of the on-chip analog-to-digital conversion peripheral circuitry,

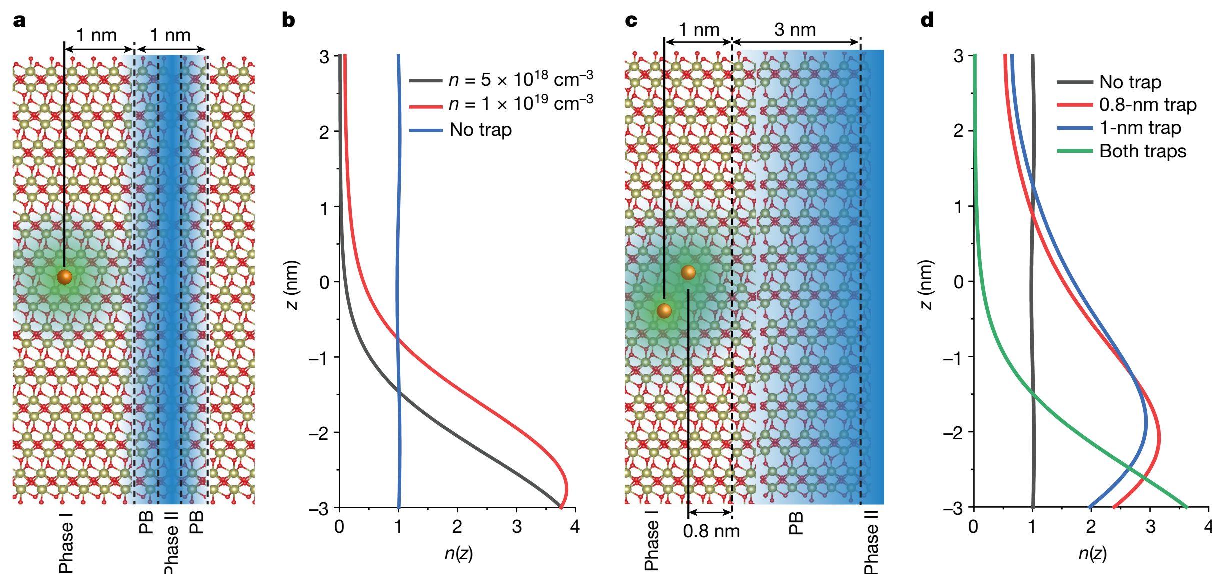

Fig. 3 | Trapped-charge-induced conductance change in incomplete conduction channels. a, The RTN-responsible defect (orange) is 1 nm away from an island-like conduction channel (blue). The channel is formed by a conductive phase region (phase II) and the phase boundary (PB) region. **b**, The transport electron wavefunction corresponding to **a**, where z denotes the position of the channel along the electron transport direction (from −3 nm to 3 nm), and $n(z)$ shows the normalized integration of the transport electron wavefunction on the plane perpendicular to the z direction, which indicates the electrical conduction at each z position. The black and red curves are $n(z)$ when the carrier density in the channel is 5×10^{18} cm^{-3} or 1×10^{19} cm^{-3} with one electron trapped at the defect, respectively, and the blue line is $n(z)$ with no electron trapped. **c**, Two defects (orange) are positioned away from a channel that is attached to the main conduction channel. The PB region is 3 nm in width. **d**, The transport electron wavefunction corresponding to **c**. The red and blue lines correspond to $n(z)$ when one electron is trapped in the defect 0.8 nm and 1 nm away from the channel, respectively, and the green and black lines correspond to $n(z)$ when both or none of the defects have trapped electrons. The carrier density in the channel for the simulation is 5×10^{18} cm^{-3}.

which was 6-bit (64 levels) in this design. The testing set-up and the schematic of the driving circuits are shown in Supplementary Fig. 8. The extra system cost caused by the denoising process is estimated in Supplementary Information Section 9. Because a relatively smaller voltage is needed for denoising than is required for typical set or reset programming, the extra energy consumption is only a small fraction of the energy needed for programming. Further studies show that the denoising operation can also reduce RTN in other material stacks, for example, a TaO_x-based memristor, as shown in Supplementary Fig. 10. Because reading noise has been observed in various resistive switching materials, the results indicate that the denoising step is an important, potentially essential, process for the training of memristive neural networks because unstable readings lead to incorrect outputs from the neural networks, and these cannot be compensated by adaptive in situ training.

Conduction channel evolution in denoising processes

Deciphering the underlying reason for the above results is essential for finding a reliable solution to the problem of unstable conductance states and understanding the dynamic process of memristive switching. Visualizing the evolution of conduction channels during electrical operations is informative for this purpose[39–42]. We used C-AFM to precisely locate the active conduction channel(s) and scan all of the surrounding regions. Details of the measurement are provided in the Methods and Supplementary Fig. 11. A customized device was fabricated for the C-AFM measurements. A schematic of its structure is shown in Fig. 2a. To use the Pt-coated C-AFM tip as the top electrode, the device was designed to have a reversed structure compared with that of the standard device shown in Fig. 1d. By grounding the bottom electrode and applying a voltage to the top electrode, the device can be operated as our standard device with opposite voltage polarities—that is, a positive voltage tends to reset the device, and a negative voltage tends to set the device. Denoising operations were also successfully performed by C-AFM, as shown in Fig. 2b,c. The conductance scanning results corresponding to the reading results of Fig. 2b are shown before (Fig. 2d) and after (Fig. 2e) denoising, and those for the reading results of Fig. 2c are shown in Fig. 2f,g. A comparison of the conductance maps in Fig. 2d,e reveals that the main part of the conduction channel (the 'complete' channel) remains nearly the same whereas the positive denoising voltage annihilates an island-like channel (the 'incomplete' channel). By contrast, the negative denoising voltage (Fig. 2f,g) reduces the noise by removing the current dips in Fig. 2c. These results indicate that the conductance of an RTN-rich state can be divided into two parts: the base conductance provided by complete channels and the RTN provided by incomplete channels. These incomplete channels had formed together with complete channels but were smaller in size. Such incomplete channels were also observed in $SrTiO_3$-based resistive switching devices[43]. A memristor can be denoised by eliminating incomplete channels (by either removing or completing them). Incomplete channels are more sensitive to voltage stimuli compared with complete channels, which makes it possible to tune the former without affecting the latter by using appropriate electrical stimuli. Further studies suggest that this is a general mechanism and can also be performed in other material stacks (Supplementary Fig. 12). It should be noted that the seemingly isolated island(s) may or may not be electrically connected with the main conduction channel beneath the surface. However, this does not change the denoising mechanisms or operation protocols.

Switching and denoising mechanisms

To understand the mechanism of denoising, we studied the microscopic origin of RTN in memristors. An important question is whether RTN is induced by an atomic effect or electronic effect. As shown in Supplementary Fig. 13, incomplete channels are consistently observed in a C-AFM scan whenever RTN is observed. Once incomplete channels are eliminated, the RTN disappears. This indicates that RTN is associated

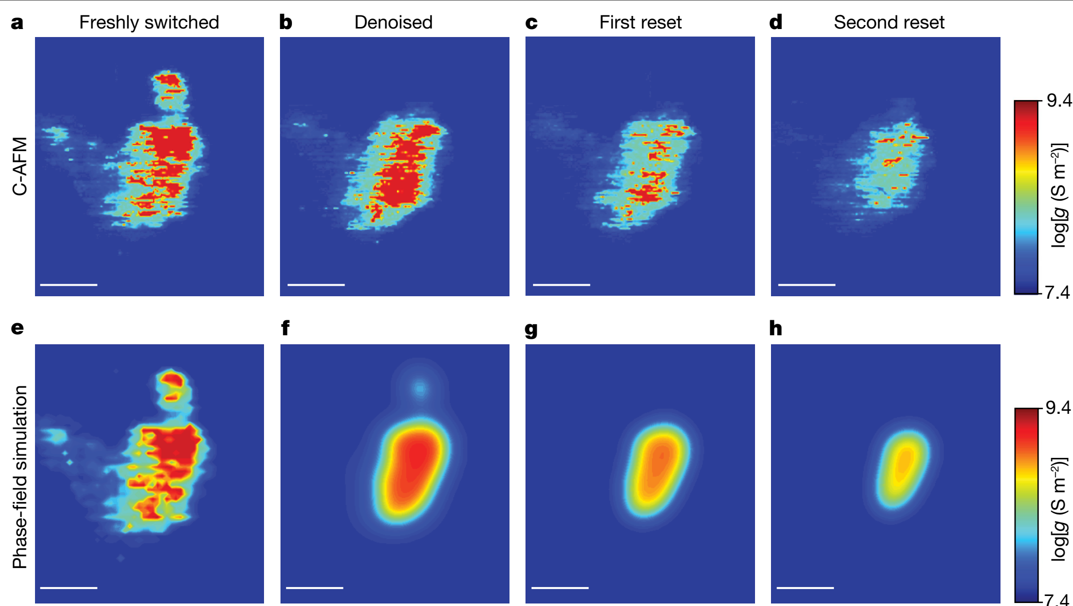

Fig. 4 | Mechanism of denoising using subthreshold voltage, identified using C-AFM measurements and phase-field theory simulations. a–d, After switching (**a**), the conduction channel is first denoised by a 0.2 V voltage (**b**) and then reset twice with a 0.5 V voltage (**c**,**d**), as measured by C-AFM. **e–h**, Phase-field simulations of the conduction channels when the device is freshly switched (**e**), then denoised (**f**), and reset twice (**g**,**h**). The dynamics of the conductive and insulating phase fields are simulated on the basis of the phase transition energy pathway from the first-principles calculation. We propose that the conductive and insulating phases are the orthorhombic phase with a high number of oxygen vacancies and the monoclinic phase without oxygen vacancies, respectively. The denoising process is captured by the phase-field relaxation, in which the island of the incomplete channel disappears and the phase boundary sharpens.

with incomplete channels rather than being induced by the transition process (by atomic motion) between incomplete and complete channels. Previously, a theoretical framework was established for the electronic RTN mechanism[33,34,44–46], in which the electrical conduction of the incomplete conduction channels was frequently blocked by Coulomb repulsion when nearby defects trapped electrons and became negatively charged. RTN caused by the atomic motion induced by external voltage stimuli is random, and irregular in amplitude even when the device is driven by regular voltage pulses[47].

To identify the type of defect that traps or detraps charges, we measured memristor RTN at different voltages and performed theoretical analyses as described in Supplementary Information Section 14. First-principles calculations indicate that the defects might be oxygen interstitials that have large relaxation energies and thus long trapping or detrapping times, consistent with the measurement shown in Supplementary Information Section 14 and Supplementary Fig. 15. It was also previously reported[44] that charge trapping or detrapping at oxygen interstitials may be responsible for RTN in oxide memristors. The strongly non-equilibrium condition during device programming probably drives oxygen ions from conduction channels into their surrounding regions[48] (Supplementary Fig. 16), leading to oxygen interstitial defects and potentially providing a type of trapping or detrapping source. By further analysing the relationship between characteristic duration of RTN and the reading voltage amplitude, we propose that RTN is predominantly induced by an electronic effect rather than an atomic effect in our device (Supplementary Information Section 17).

The incomplete channel blocking process was modelled as shown in Fig. 3. On the basis of C-AFM experiments, we classified the device region as three phases: the non-conductive phase (phase I), the conductive phase (phase II) and the region between them, which has an intermediate conductance (phase boundary). During the programming or denoising operations, these phase-boundary regions form or disappear, accompanying the observation of RTN and its removal, indicating that some RTN-inducing incomplete channels are located in these phase boundary regions. Figure 3a shows a defect trapping or detrapping an electron 1 nm away from an island-like incomplete channel that has a width of 1 nm. The transport electron wavefunctions $\psi(x, y, z)$ with or without a trapped charge are visualized in Fig. 3b by the probability density at each cross-section of the channel $n(z) = \int |\psi(x, y, z)|^2 \, \mathrm{d}x\mathrm{d}y$ (where z is the axis along the channel). The wave functions show what proportion of the injected electron propagates through the channel. To mimic the different percentages of phase II, two charge carrier densities (averaged over phase I and phase II) were used for the simulations. The results indicate that the incomplete channel is fully blocked at a lower charge carrier density (lightly doped with oxygen vacancies, corresponding to less phase II) and partially blocked at a higher charge carrier density (heavily doped, corresponding to more phase II). Figure 3c corresponds to another commonly observed C-AFM result, in which the incomplete channel is attached to the main channel with multiple charge traps around it. Figure 3d shows that a trapped charge close to the incomplete channel tends to have a larger impact on conductance than one far away. Furthermore, the effect of multiple charge traps can enhance each other and lead to a multiplied change of conductance because the thick phase boundary region is completely blocked. Compared with previous models using classic carrier drift-diffusion equations, we use quantum transport formalism to simulate the influence of charged defects on channel conductivity, confirming that the Coulomb blockade mechanism applies to nanoscale channels. Furthermore, we inferred that two or more (N) charge-trapping defects can lead to complex RTN patterns with a maximum of 2^N levels, which is consistent with previous reports[45,46].

Because the RTN originates from the incomplete conduction channels, the denoising process is associated with the disappearance of both the island and the blurry boundary of the main channel. A subthreshold voltage that is much smaller than the set or reset voltages can decrease the RTN because of the phase-field relaxation, as shown in Fig. 4. For this specific material system, the relatively conductive and insulating

phases (phase II and phase I, respectively; Fig. 3) are the orthorhombic and monoclinic phases of HfO_2, because the orthorhombic phase is stabilized by a high number of oxygen vacancies[49]. The denoising voltage provides a driving force for the phase relaxation through both temperature effects and the current-induced forces, enabling the system to relax towards an equilibrium state. The free energy F and equation of motion of the system are as follows:

$$\Delta F = \int \left[\Delta f_0(\eta) + \frac{1}{2} K (\nabla \eta)^2 \right] \mathrm{d}V$$

$$\frac{1}{\alpha} \frac{\partial \eta(r)}{\partial t} = - \frac{\delta \Delta F[\eta]}{\partial \eta(r)} = - \frac{\partial \Delta f_0}{\partial \eta} + K \nabla^2 \eta$$

where η is the order parameter (here, the monoclinic angle) describing the transition from the monoclinic to the orthorhombic phase, Δf_0 is the free energy density for a system with a certain order parameter and K is the gradient energy parameter. The energy density Δf_0 is derived from the first-principles calculations. Using the phase-field simulation, we derive a similar behaviour as observed by the C-AFM: after denoising, the island disappears and the boundary of the main channel sharpens. The disappearance and sharpening of the boundary are driven by the energy barrier between the two phases, in which the high-energy boundary region is reduced. During the reset process, the conduction channel shrinks in size and its conductivity also decreases because the strong voltage drives the oxygen vacancy away from the switching-active region. The incomplete conduction channels—that is, the islands and boundary regions in a freshly switched state—are frozen in a highly non-equilibrium state because they are always formed at the end of the set or reset voltage pulse and do not have a chance (sufficient time) to reach the same stable state as the more mature complete channel region formed earlier. Therefore, these incomplete conduction channels are prone to change; the completion or removal can be induced by a subthreshold voltage. In contrast to the electron transport in the complete main conduction channel, that of incomplete channels can be readily blocked by trapped charges (Fig. 3), making them the main source of RTN. The situation is more severe for a conductance state obtained by a set switching process because the creation and growth of a conduction channel comprise a positive feedback process, which happens faster and faster and leaves no time for the maturation of the newly formed conduction channels before the end of each switching pulse. In the denoising process, there is no need for the migration, annihilation or creation of trap sites (for example, interstitial oxygen defects). Although the specific phases involved may be different for different oxide systems, our approach and conclusions are generally applicable.

Summary

We have achieved 2,048 conductance levels in a memristor which is more than an order of magnitude higher than previous demonstrations[22,32]. Notably, these were obtained in memristors of a fully integrated chip fabricated in a commercial factory. We have shown the root cause of conductance fluctuations in memristors through experimental and theoretical studies and devised an electrical operation protocol to denoise the memristors for high-precision operations. The denoising process has been successfully applied to the entire 256 × 256 crossbars using the on-chip driving circuitry designed for regular reading and programming without any extra hardware. These results not only provide crucial insights into the microscopy picture of the memristive switching process but also represent a step forward in commercializing memristor technology as hardware accelerators of machine learning and artificial intelligence for edge applications. Moreover, such analog memristors may also enable electronic circuits capable of growing for the recently proposed mortal computations[30].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05759-5.

1. Chua, L. O. Memristor—the missing circuit element. *IEEE Trans. Circuit Theory* **18**, 507–519 (1971).
2. Valov, I., Waser, R., Jameson, J. R. & Kozicki, M. N. Electrochemical metallization memories—fundamentals, applications, prospects. *Nanotechnology* **22**, 254003 (2011).
3. Yang, Y. & Huang, R. Probing memristive switching in nanoionic devices. *Nat. Electron.* **1**, 274–287 (2018).
4. Wen, W., Wu, C., Wang, Y., Chen, Y. & Li, H. Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems 29* (eds Lee, D. D., et al.), 2082–2090 (Curan Associates, 2016).
5. Wan, W. et al. A compute-in-memory chip based on resistive random-access memory. *Nature* **608**, 504–512 (2022).
6. Kumar, S., Wang, X., Strachan, J. P., Yang, Y. & Lu, W. D. Dynamical memristors for higher-complexity neuromorphic computing.*Nat. Rev. Mater.* **7**, 575–591 (2022).
7. Xue, C.-X. et al. A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices. *Nat. Electron.* **4**, 81–90 (2021).
8. Lanza, M. et al. Memristive technologies for data storage, computation, encryption, and radio-frequency communication. *Science* **376**, eabj9979 (2022).
9. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).
10. Zhang, W. et al. Neuro-inspired computing chips. *Nat. Electron.* **3**, 371–382 (2020).
11. Ielmini, D. & Wong, H.-S. P. In-memory computing with resistive switching devices. *Nat. Electron.* **1**, 333–343 (2018).
12. Zidan, M. A., Strachan, J. P. & Lu, W. D. The future of electronics based on memristive systems. *Nat. Electron.* **1**, 22–29 (2018).
13. Yu, S. Neuro-inspired computing with emerging nonvolatile memorys. *Proc. IEEE* **106**, 260–285 (2018).
14. Jung, S. et al. A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature* **601**, 211–216 (2022).
15. Sangwan, V. K. & Hersam, M. C. Neuromorphic nanoelectronic materials. *Nat. Nanotechnol.* **15**, 517–528 (2020).
16. Burr, G. W. A role for analogue memory in AI hardware. *Nat. Mach. Intell.* **1**, 10–11 (2019).
17. Chen, S. et al. Wafer-scale integration of two-dimensional materials in high-density memristive crossbar arrays for artificial neural networks. *Nat. Electron.* **3**, 638–645 (2020).
18. Fuller, E. J. et al. Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing. *Science* **364**, 570–574 (2019).
19. Choi, C. et al. Reconfigurable heterogeneous integration using stackable chips with embedded artificial intelligence.*Nat. Electron.* **5**, 386–393 (2022).
20. Lim, D.-H. et al. Spontaneous sparse learning for PCM-based memristor neural networks. *Nat. Commun.* **12**, 319 (2021).
21. Xu, X. et al. Scaling for edge inference of deep neural networks. *Nat. Electron.* **1**, 216–222 (2018).
22. Sun, Y. et al. A Ti/AlO_x/TaO_x/Pt analog synapse for memristive neural network. *IEEE Electron Device Lett.* **39**, 1298–1301 (2018).
23. Stathopoulos, S. et al. Multibit memory operation of metal-oxide bi-layer memristors. *Sci. Rep.* **7**, 17532 (2017).
24. Kim, H., Mahmoodi, M. R., Nili, H. & Strukov, D. B. 4K-memristor analog-grade passive crossbar circuit. *Nat. Commun.* **12**, 5198 (2021).
25. Zidan, M. A. et al. A general memristor-based partial differential equation solver. *Nat. Electron.* **1**, 411–420 (2018).
26. Li, C. et al. Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* **1**, 52–59 (2018).
27. Mackin, C. et al. Optimised weight programming for analogue memory-based deep neural networks. *Nat. Commun.* **13**, 3765 (2022).
28. Choi, S. et al. SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations. *Nat. Mater.* **17**, 335–340 (2018).
29. Yao, P. et al. Face classification using electronic synapses. *Nat. Commun.* **8**, 15199 (2017).
30. Hinton, G. The forward–forward algorithm: some preliminary investigations. Preprint at https://arxiv.org/abs/2212.13345 (2022).
31. Yan, Z., Hu, X. S. & Shi, Y. SWIM: Selective write-verify for computing-in-memory neural accelerators. Preprint at https://arxiv.org/abs/2202.08395 (2022).
32. Chen, B. et al. A memristor-based hybrid analog-digital computing platform for mobile robotics. *Sci. Robot.* **5**, eabb6938 (2020).
33. Choi, S., Yang, Y. & Lu, W. Random telegraph noise and resistance switching analysis of oxide based resistive memory. *Nanoscale* **6**, 400–404 (2014).
34. Ielmini, D., Nardi, F. & Cagli, C. Resistance-dependent amplitude of random telegraph-signal noise in resistive switching memories. *Appl. Phys. Lett.* **96**, 053503 (2010).
35. Puglisi, F. M., Pavan, P., Padovani, A., Larcher, L. & Bersuker, G. Random telegraph signal noise properties of HfOx RRAM in high resistive state. In *2012 Proc. European Solid-State Device Research Conference (ESSDERC)*, 274–277 (IEEE, 2012).
36. Lee, J.-K. et al. Extraction of trap location and energy from random telegraph noise in amorphous TiO_x resistance random access memories. *Appl. Phys. Lett.* **98**, 143502 (2011).

37. Puglisi, F. M., Padovani, A., Larcher, L. & Pavan, P. Random telegraph noise: measurement, data analysis, and interpretation. In *2017 IEEE 24th International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*, 1–9 (IEEE, 2017).
38. Puglisi, F. M., Zagni, N., Larcher, L. & Pavan, P. Random telegraph noise in resistive random access memories: compact modeling and advanced circuit design. *IEEE Trans. Electron Devices* **65**, 2964–2972 (2018).
39. Yang, Y. et al. Probing nanoscale oxygen ion motion in memristive systems. *Nat. Commun.* **8**, 15173 (2017).
40. Puglisi, F. M. *Noise in Nanoscale Semiconductor Devices* (ed. Grassor, T.), 87–133 (Springer, 2020).
41. Hui, F. & Lanza, M. Scanning probe microscopy for advanced nanoelectronics. *Nat. Electron.* **2**, 221–229 (2019).
42. Celano, U. et al. Three-dimensional observation of the conductive filament in nanoscaled resistive memory devices. *Nano Lett.* **14**, 2401–2406 (2014).
43. Du, H. et al. Nanosized conducting filaments formed by atomic-scale defects in redox-based resistive switching memories. *Chem. Mater.* **29**, 3164–3173 (2017).
44. Puglisi, F. M., Larcher, L., Padovani, A. & Pavan, P. A complete statistical investigation of RTN in HfO_2-based RRAM in high resistive state. *IEEE Trans. Electron Devices* **62**, 2606–2613 (2015).
45. Ambrogio, S. et al. Statistical fluctuations in HfO_x resistive-switching memory: part II—random telegraph noise. *IEEE Trans. Electron Devices* **61**, 2920–2927 (2014).
46. Becker, T. et al. An electrical model for trap coupling effects on random telegraph noise. *IEEE Electron Device Lett.* **41**, 1596–1599 (2020).
47. Brivio, S., Frascaroli, J., Covi, E. & Spiga, S. Stimulated ionic telegraph noise in filamentary memristive devices. *Sci. Rep.* **9**, 6310 (2019).
48. Miao, F. et al. Anatomy of a nanoscale conduction channel reveals the mechanism of a high-performance memristor. *Adv. Mater.* **23**, 5633–5640 (2011).
49. Zhou, Y. et al. The effects of oxygen vacancies on ferroelectric phase transition of HfO_2-based thin film from first-principle. *Comput. Mater. Sci.* **167**, 143–150 (2019).

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Methods

Memristor fabrication

Standard memristors integrated with CMOS driving circuits. The CMOS part was fabricated in a standard 180-nm process line in a commercial semiconductor manufacturer with an exposed tungsten via at the top. Memristors were processed in the same process line with customized materials and protocols. After surface oxide cleaning of the tungsten via, the Pt bottom electrodes were sputtered and patterned on the vias. Holes for memristors were created by etching through a patterned SiO_2 isolation layer (~100 nm) and terminating at the surface of Pt. The resistive switching layer (HfO_2/Al_2O_3) and top electrode (Ti/Ta) were filled into the etched holes sequentially, in which the resistive switching layers were fabricated by atomic layer deposition and the top electrode was fabricated by sputter. Finally, a standard aluminium interconnect was used to connect the top electrode to bond pads for electrical testing.

A customized memristor for C-AFM measurement. The customized device was fabricated in a university cleanroom on an Si wafer covered with thermally oxidized SiO_2 (~100 nm). The bottom electrode (Ta/Ti) and resistive switching layers (Al_2O_3/HfO_2) were deposited by an AJA sputtering system. The four layers were fabricated continuously in a high-vacuum chamber to avoid oxidation of Ta and Ti. The chip was then patterned and etched to expose part of the bottom electrode. After surface oxide cleaning, Pt was deposited onto the exposed bottom electrode to prevent oxidation and serve as the ground contact during C-AFM measurement.

Electrical measurements

Single-device measurement. Electrical measurements of the standard memristor (factory-made complete memristor with top electrode) were performed on a Keysight B1500A semiconductor device analyser equipped with a B1530A waveform generator and fast measurement unit. To realize the algorithm as shown in Supplementary Fig. 4, we built a program using C# to control the electrical operations of B1500A.

Array measurements. The schematic of the one-transistor–one-memristor array with on-chip driving circuits and the testing set-up is shown in Supplementary Figs. 7 and 8.

C-AFM measurements. C-AFM was performed using a Bruker Dimension Icon system with a conductive probe (SCM-PIT-V2, 0.01–0.025 ohm-cm of resistivity) in a contact mode. When performing electrical operations, including set, reset and read, the C-AFM probe was at a fixed position. The conduction channel was first formed with a voltage of 4 V. During the in situ set, reset and reading operations for the chosen conduction channel, the set point was set to a relatively large number (around 80 nm) to increase the strength of the pressing force to make a large contact area between the tip and sample surface. The set point is a measure of the force applied by the tip to the sample. In contact mode, it is a certain deflection of the cantilever. This deflection is maintained by the feedback electronics, so that the force between the tip and sample is kept constant. When performing the conduction channel morphology mapping, the probe scanned a 150 × 150 nm region surrounding the conductive channel. During this measurement, the set point was set to a small value (around 10 nm) for high resolution. The relationship between the contact radius and set point are shown in Supplementary Fig. 11.

First-principles calculations

The atomic and electronic structures of the oxygen interstitial defects are calculated using the density functional theory with the projector augmented wave method[50] implemented in the Vienna ab initio simulation package[51]. The generalized gradient approximation is used together with the Perdew–Burke–Ernzerhof exchange-correlation function[52]. The cut-off energy is set to 400 eV and the k-point mesh is sampled using the Monkhorst–Pack method[53] with a separation of 0.2 rad Å^{-1}. The atomic structure of the oxygen interstitial defect is constructed by including one oxygen atom in the $2\times2\times2$ supercell of the monoclinic-phase HfO_2 crystal. The initial position of the included oxygen atom is set as described previously[54] and the atomic configuration is fully relaxed. The force on each atom converges to 0.01 eV Å^{-1}, and the electronic energy converges to 10^{-6} eV. The atomic structure, charge distribution of the trap state and electronic band structure in Fig. 3 and Supplementary Figs. 14 and 15 are then extracted from the calculations of the density functional theory.

Simulation of the effect of a trapped charge on the conductive channel. We simulate the Coulomb blockade effect through the quantum transport of a conduction electron in a cuboid conduction channel (Fig. 3). The length of the conductive channel is set to L = 6 nm to match the channel length in the device. The motion of carriers in the conductive channel is calculated through the effective mass approximation and the Coulomb blockade effect of the RTN-responsible defect is simulated by a screened Coulomb potential $V(\mathbf{r})$ acting on the carriers. Assuming the electric conductance outside the channel is negligible, the quantum transport of an electron in the channel can be described by the following equations:

$$\begin{cases} \left[-\dfrac{\hbar^2}{2m^*}\nabla^2 + V(\mathbf{r})\right]\psi(x,y,z) = (E-E_c)\psi(x,y,z) \\ V(\mathbf{r}) = \dfrac{e^2}{4\pi\epsilon_0\epsilon_r}\sum_i \dfrac{e^{-|\mathbf{r}-\mathbf{r}_i|/\lambda_D}}{|\mathbf{r}-\mathbf{r}_i|} \\ \psi|_{x=0} = \psi|_{x=d} = \psi|_{y=0} = \psi|_{y=d} = 0 \end{cases}$$

where m^* is the effective mass of the conductive band of HfO_2, set to $0.11m_e$ (ref. [55]). E is the eigen energy of the transport electron, set to 0.2 eV above the conduction band minimum E_c estimated by the magnitude of bias voltage of about 0.2 V. The Coulomb potential is the summation of the RTN-responsible defect located at $\mathbf{r}_i$, where ϵ_r is the relative dielectric constant (set to 16 as proposed previously[56]) and λ_D is the Debye screening length calculated as $\lambda_D = \sqrt{\frac{\epsilon_0\epsilon_r k_B T}{ne^2}}$ (the temperature T is set to 300 K).

The transport wavefunction with electrons injected from $x = 0$ with unitary amplitude is then calculated with the following boundary conditions:

$$\begin{cases} \psi|_{z=0} = e^{ik_{11}z}\sin\dfrac{\pi x}{d}\sin\dfrac{\pi y}{d} + \sum\limits_{m,n} R_{mn}e^{-ik_{mn}z}\sin\dfrac{m\pi x}{d}\sin\dfrac{n\pi y}{d} \\ \left.\dfrac{\partial\psi}{\partial z}\right|_{z=0} = ik_{11}e^{ik_{11}z}\sin\dfrac{\pi x}{d}\sin\dfrac{\pi y}{d} - \sum\limits_{m,n} R_{mn}ik_{mn}e^{-ik_{mn}z}\sin\dfrac{m\pi x}{d}\sin\dfrac{n\pi y}{d} \\ \psi|_{z=L} = \sum\limits_{m,n} T_{mn}e^{ik_{mn}z}\sin\dfrac{m\pi x}{d}\sin\dfrac{n\pi y}{d} \\ \left.\dfrac{\partial\psi}{\partial z}\right|_{z=L} = \sum\limits_{m,n} T_{mn}ik_{mn}e^{ik_{mn}z}\sin\dfrac{m\pi x}{d}\sin\dfrac{n\pi y}{d} \\ k_{mn} = \sqrt{\dfrac{2m^*}{\hbar^2}(E-E_c) - (m^2+n^2-2)\left(\dfrac{\pi}{d}\right)^2} \end{cases}$$

The electron transport is then shown by the probability density function of the electron wavefunction at each cross-section of the channel $n(z) = \int|\psi(x,y,z)|^2\mathrm{d}x\mathrm{d}y$, reflecting what proportion of the injected electron propagates through the channel. If $n(L) \simeq 0$, the electron transport is completely blocked; if $n(L) \simeq 1$, the electron goes through the channel with negligible barrier. Three parameters control the Coulomb blockade: the size of channel d, the carrier density n and the distance of the RTN-responsible defect to the channel. These factors lead to the different degrees of Coulomb blockade to the isolated island and main channel.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The algorithm for memristor high-precision programming is included in the Supplementary Information. The code for physical modelling and simulations is available at GitHub (https://github.com/htang113/HfO2-memristor-denoise/tree/main).

50. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
51. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
52. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
53. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188–5192 (1976).
54. Lyons, J. L., Janotti, A. & Van de Walle, C. G. The role of oxygen-related defects and hydrogen impurities in HfO_2 and ZrO_2. *Microelectron. Eng.* **88**, 1452–1456 (2011).
55. Monaghan, S., Hurley, P. K., Cherkaoui, K., Negara, M. A. & Schenk, A. Determination of electron effective mass and electron affinity in HfO_2 using MOS and MOSFET structures. *Solid State Electron.* **53**, 438–444 (2009).
56. Zhao, X. & Vanderbilt, D. First-principles study of structural, vibrational, and lattice dielectric properties of hafnium oxide. *Phys. Rev. B* **65**, 233106 (2002).

Acknowledgements J.J.Y., W.S. and Y.Z. were partially supported by a subcontract (GR1055585 53-4502-0003) from the University of Massachusetts Amherst, with the sponsor being TetraMem. R.M., Q.X. and J.J.Y. were partially supported by the Air Force Office of Scientific Research through the Multidisciplinary University Research Initiative programme under contract no. FA9550-19-1-0213, the US Air Force Research Laboratory (prime contract nos. FA8650-21-C-5405 and FA8750-22-1-0501) and by the National Science Foundation under contract no. 2023752. J.W. and H.W. acknowledge the support by the Army Research Office (grant no. W911NF2120128) and the National Science Foundation (grant no. CMMI-2240407). H.T. and J.L. acknowledge the support by the National Science Foundation (grant no. CMMI-1922206). We thank A. Tan for proofreading the manuscript.

Author contributions J.J.Y. and M.R. conceived the concept. J.J.Y. and Q.X. supervised the entire project. J.J.Y., M.R., Q.X., H.T., J.W. and W.S. designed the experiments and simulations. M.R., M.Z., R.M. and H.J. fabricated the devices. M.R., W.S., Y.Z., B.C., X.J. and Z.W. carried out the electrical measurements. H.T., M.R. and J.L. designed and carried out the simulation. J.W., M.R., H.L., H.-Y.C. and H.W. designed and carried out the C-AFM studies. W.Y., F.K., F.Y., Z.W., M.W., M.H., Q.X., N.G. and J.J.Y. helped with experiments and data analysis. M.R., H.T. and J.J.Y. wrote the paper. All authors discussed the results and implications and commented on the manuscript at all stages.

Competing interests J.J.Y. and Q.X. are co-founders and paid consultants of TetraMem.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05759-5.

Correspondence and requests for materials should be addressed to J. Joshua Yang.

Peer review information *Nature* thanks Yiyu Shi, Ilia Valov and Yuchao Yang for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.