Thousands of conductance levels in memristors monolithically integrated on CMOS

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,*]

[1]TetraMem Inc., 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.
[*] Email: jjoshuay@usc.edu

Neural networks based on memristive devices [1-3] have shown potential in substantially improving 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 very 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 (Fig. 1a). Some post-tuning in memristor conductance to adapt local situations may follow afterward or during applications. *Therefore, a critical requirement on memristors for neural network applications is a high-precision programming ability to guarantee uniform and accurate performance across a massive number of memristive networks.* [22-28] That translates into the requirement of many distinguishable conductance levels on each memristive device, not just lab-made devices but more importantly, devices fabricated in foundries. 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 over 2048 conductance levels, *the largest number among all types of memories ever reported*, achieved with memristors in *fully integrated chips with 256 ×256 memristor arrays monolithically integrated on CMOS circuits in a standard foundry*. We have unearthed the underlying physics that previously limited the number of achievable conductance levels in memristors and developed electrical operation protocols to circumvent such limitations. These results reveal insights into the fundamental understanding of the microscopic picture of memristive switching and provide approaches to enable high-precision memristors for various applications.

Memristive switching devices are known for their relatively large dynamical range of conductance, which can potentially lead to a large number of discrete conductance levels.

Different approaches have been developed to accurately program the devices. [31] However, the highest conductance number reported to date has been no more than two hundred.[22, 32] There are no forbidden conductance states within the dynamical range of the device since a memristor is typically 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 achievable within a specific conductance range. Interestingly, we found that such fluctuation can be substantially suppressed, as shown in Figs. 1e and 1f, by applying appropriate electrical stimuli (termed as 'denoising' processes). Importantly, such denoising process does not require any extra circuitry beyond the normal read and program circuits. We incorporated the denoising process into device tuning algorithms and successfully programmed a commercial-semiconductor-manufacturer-made memristor (Figs. 1b-d) into 2048 conductance levels (Fig. 1g), corresponding to 11-bit resolution. Conductive atomic force microscopy (C-AFM) was employed to visualize the evolution of conduction channels during programming and denoising processes. We discovered that a normal switching operation (SET or RSET) 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 principle calculations suggest that these incomplete channels are unstable phase boundaries with dopant levels in a range that is sensitive to trapped charges, contributing to the large fluctuations of each conductance level. We revealed, experimentally and theoretically, that an appropriate voltage in the denoising process either annihilates (weakens) or completes (enhances) these incomplete channels, resulting in a great reduction in fluctuation and a significant increase in memristor precision. The observed phenomena generally exist in memristive switching process with localized conduction channels, and the insights can be applied to most memristive material systems for scientific understanding and technological applications.

Memristors used in this study were fabricated on an 8-inch wafer by a commercial semiconductor manufacturer (Fig. 1b). The fabrication details are given in the *Method* section. Cross-section views of a memristor are shown in Fig. 1c, and the critical resistive switching layers are zoomed-in in Fig. 1d. The electron energy loss spectroscopy (EELS) elemental image is shown in Fig. S1. The device consisting 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 so thin (<1nm) that they appear 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 (Fig. S2). The result shows that an as-programmed state typically has a large fluctuation, which significantly limits the applications of memristors but unfortunately exists in memristive materials generally. [33-36] The data also reveals that a SET operation tends to induce a larger fluctuation in an as-programmed state than a RESET operation. The main contribution of such reading fluctuation is random telegraph noise (RTN) which features

step-like transitions between two or more current levels at random time points under a constant reading voltage. Such RTNs generally exist in memristors. Even fluctuations that are seemingly not step-like may in fact be made of RTN noise, [37] which can be revealed only when the measurement sampling rate is higher than the RTN frequency, as shown in Fig. S3. It has been demonstrated previously by simulations that memristor RTNs may be caused by charges occasionally trapping into certain defects and blocking conduction channels via coulomb screening. [34, 38] However, experiments that directly link trapped charges, conduction channel(s), and RTNs are missing, let alone how to remove RTNs. Although a critical issue for memristors in general, it has been unclear how to reduce RTNs in memristors. They are critical not only for understanding the physical origin of memristor RTNs but also for revealing the entire microscopic 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. One example is given in Fig. 1e, where 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 detailed in Fig. S4, a single memristor was tuned into 2,048 conductance states between 50 and 4,144 μS, with a 2 μS interval between every two neighboring states. All states were read by a voltage sweeping from 0 to 0.2V, as shown in Fig. 1g. The zoomed-in view of the current-voltage curves is given as the lower inset to Fig. 1g, showing well-distinguishable states and the superb linearity of each state. Three nearest neighboring states after denoising are shown in Fig. 1f, where a constant 0.2V voltage reads each state for 1,000 seconds. The current fluctuation of every state is within 0.4 μA, corresponding to 2 μS in conductance. No significant overlap was observed in the neighboring states. The zoom-in view of the measurement result at high conductance states is shown in Fig. S5. Memristors from multiple chips of an 8-inch wafer were measured, demonstrating a great programming uniformity across the entire wafer, as shown in Fig. S6. We further adopted 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 both in Fig. 1g as an upper inset and in Fig. S7. For these demonstrations using the on-chip circuitry, the programming precision was limited by the precision of the on-chip Analog/Digital conversion peripheral circuitry, which was 6-bit (64 levels) in this design. The testing setup and the schematic of the driving circuits are shown in Fig. S8. The extra system cost caused by the denoising process is estimated in SI-9, showing that due to a relatively smaller voltage needed for denoising than for a typical SET/RESET programming, the extra energy consumption is only a small fraction of the energy for programming. Further studies show that the denoising operation can also reduce RTNs in other material stacks, e.g., a TaO_x-based memristor, as shown in Fig. S10. Since reading noise has been observed in various resistive switching materials, the above results show that the denoising step is an important, or even essential, process for the training of memristive neural networks as unstable readings lead to incorrect outputs from the neural networks and cannot be compensated by adaptive in-situ training.

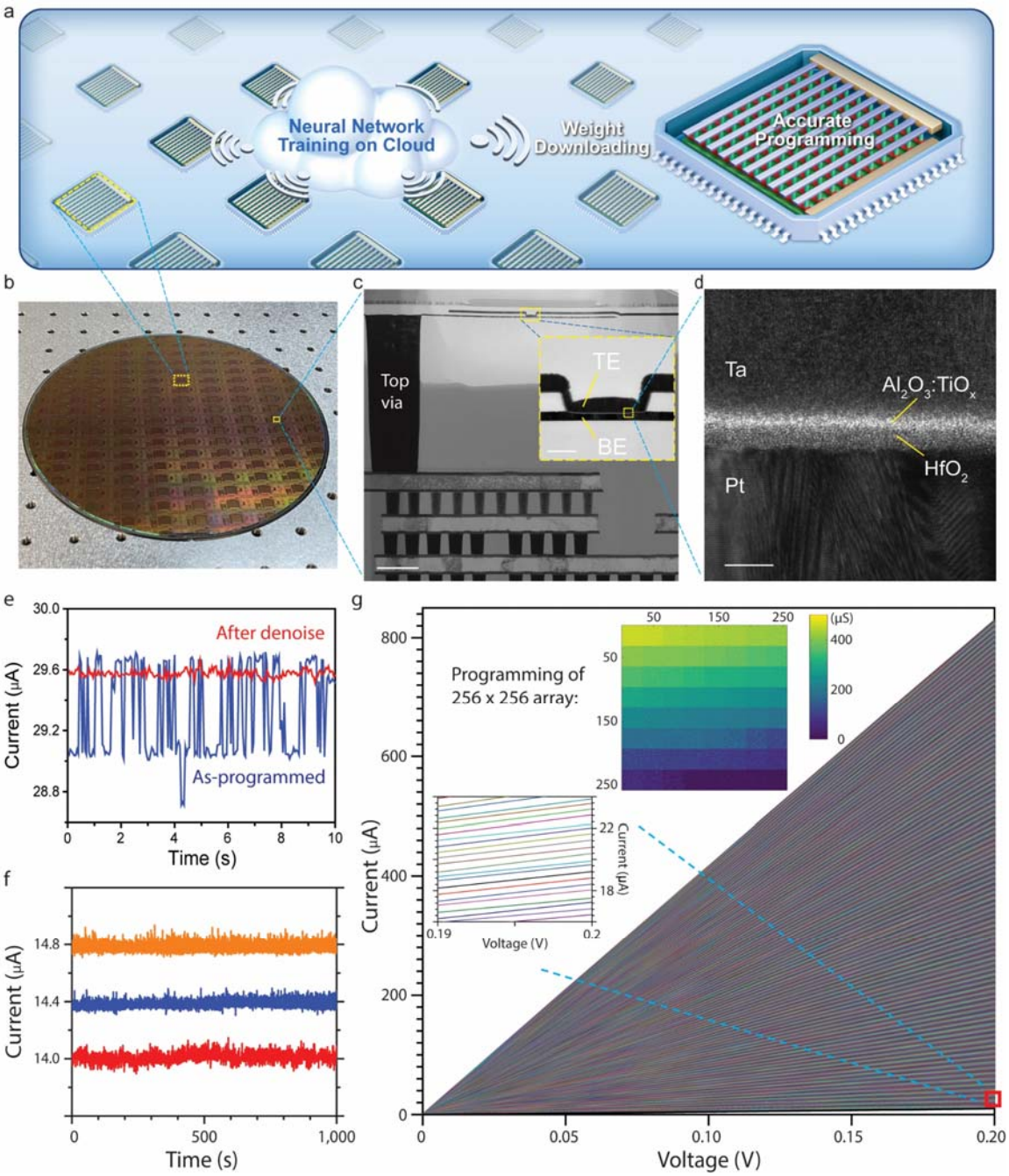


Fig. 1 High precision memristor for neuromorphic computing. a) The most likely scheme of the large-scale application of memristive neural networks for edge computing. Neural network training is performed in the cloud. The obtained weights are downloaded and accurately programmed into a massive number of memristor arrays distributed at the edge, which imposes high-precision requirements on memristive devices. b) The photo of an 8-inch wafer with memristors fabricated by a commercial semiconductor manufacturer. c) HR-TEM image of the cross-section view of a memristor. Pt and Ta serve as bottom and top electrodes, respectively. Scale bar (inset): 1 μm (100 nm). d) The zoomed-in image of

the memristor material stack. Scale bar: 5 nm. e) The as-programmed (blue) and after-denoising (red) currents of a memristor are read by a constant 0.2V voltage. The denoising process eliminated the large amplitude random telegraph noise (RTN) observed in the as-programmed state (see method). f) Zoomed-in view of three nearest neighboring states after denoising. The current of each state was read by a constant 0.2V voltage. No large-amplitude RTN was observed, and all the states can be clearly distinguished. g) An individual memristor on the chip was tuned into 2048 resistance levels by a high-resolution off-chip driving circuitry, and each resistance level was read by a DC voltage sweeping from 0 to 0.2V. The target resistance was set from 50 μS to 4,144 μS with 2 μS interval between neighboring levels. All readings at 0.2V are less than 1 μS from the target conductance. The lower inset shows a zoomed-in view of the resistance levels. The upper inset shows experimental results of an entire 256×256 array programmed by its 6-bit on-chip circuitry into sixty-four 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 million cycles, demonstrating the high endurance and robustness of the devices.

Deciphering the underlying reason for the above discoveries is essential for offering a reliable solution to the critical technology problem 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 measurement to precisely locate the active conduction channel(s) and scan all the surrounding regions. The details of the measurement can be found in *Method* and Fig. S11. A customized device was fabricated for the C-AFM measurements. The 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 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, i.e., 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 and Fig. 2c. The conductance scanning results before and after denoising corresponding to the reading results of Fig. 2b (2c) are shown in Fig. 2d (2f) and Fig. 2e (2g), respectively. Comparing the conductance maps in Fig. 2d and Fig. 2e, it is observed that the main part of the conduction channel (the 'complete' channel) remains nearly the same while the positive denoising voltage annihilates an island-like channel (the 'incomplete' channel). In contrast, the negative denoising voltage (Fig. 2f and Fig. 2g) 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 part provided by incomplete channels. These incomplete channels were formed together with complete channels but are smaller in size. Such incomplete channels were also observed in SrTiO3-based resistive switching devices. [43] A memristor can be denoised by eliminating incomplete channels (either removing or completing them). Incomplete channels are more sensitive to voltage stimuli when compared to complete channels, which

makes it possible to tune the former without affecting the latter by using appropriate electrical stimuli. Further studies suggest that such a mechanism is general and can be performed in other material stacks (Fig. S12) as well. 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, which, however, does not change the denoising mechanisms or operation protocols.

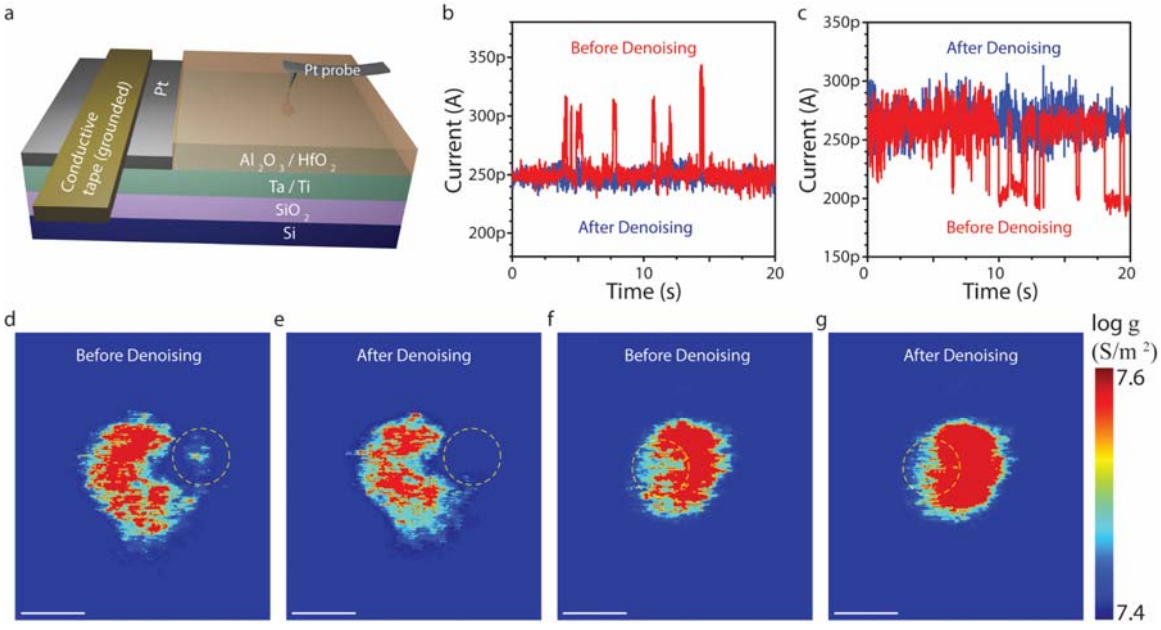


Fig. 2 Direct observation of the evolution of conduction channels in the denoising process through conductive atomic force microscope (C-AFM). a) A schematic of the customized memristor structure and C-AFM testing setup. C-AFM probe played the role of the top electrode in the customized device. Since Ta is easily oxidized in air and not practical to be used as the probe material, a Pt probe was adopted, which served the same role as that of the bottom Pt electrode of a standard memristor we used. To maintain the material stack of a standard memristor, the customized memristor has a reversed structure. b) The current readings by 0.1V voltage before (red) and after (blue) a denoising process by a sub-threshold RESET voltage. c) The current readings by 0.1V voltage before (red) and after (blue) a denoising process by a sub-threshold 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). All scale bar: 10 nm.

To understand the mechanism of denoising, we studied the microscopic origin of RTNs in memristors. A critical question is whether RTN is induced by an 'atomic effect' or 'electronic effect'. As shown in Fig. S13, incomplete channels are consistently observed in a C-AFM scanning whenever RTN is observed. Once incomplete channels are eliminated, RTN disappears. Such result indicates that RTN is a phenomenon in company with incomplete channels rather than being induced by the transition process (via atomic motion) between incomplete and complete channels. Previously, an insightful theoretical framework on the electronic RTN mechanism is established in ref. [33, 34, 44-46], where the electrical conduction of the incomplete conduction channels is frequently blocked by Coulomb repulsion when nearby defects trap electrons and become negatively charged. RTNs based on atomic motion induced by external voltage stimuli are random and irregular in amplitude even driven by regular voltage pulses. [47]

To identify the type of defect that traps/detraps charges, we measured memristor RTN at different voltages and performed theoretical analysis as shown in SI-14. First principle calculations suggest that the defects might be oxygen interstitials which feature large relaxation energies and thus long trapping/detrapping times, consistent with measurement results shown in SI-14 and Fig. S15. It was also reported in [44] that charge trapping/detrapping at oxygen interstitials may be responsible for RTN in oxide memristors. The strongly non-equilibrium condition during device programming likely drives oxygen ions from conduction channels into their surrounding regions (see in ref. [48] and Fig. S16), leading to oxygen interstitial defects and providing a type of trapping/detrapping source among other possibilities. By further analyzing the relationship between RTN characteristic time and the reading voltage amplitude, we propose that 'electronic effect' rather than 'atomic effect' induced RTN dominates in our device, as shown in SI-17.

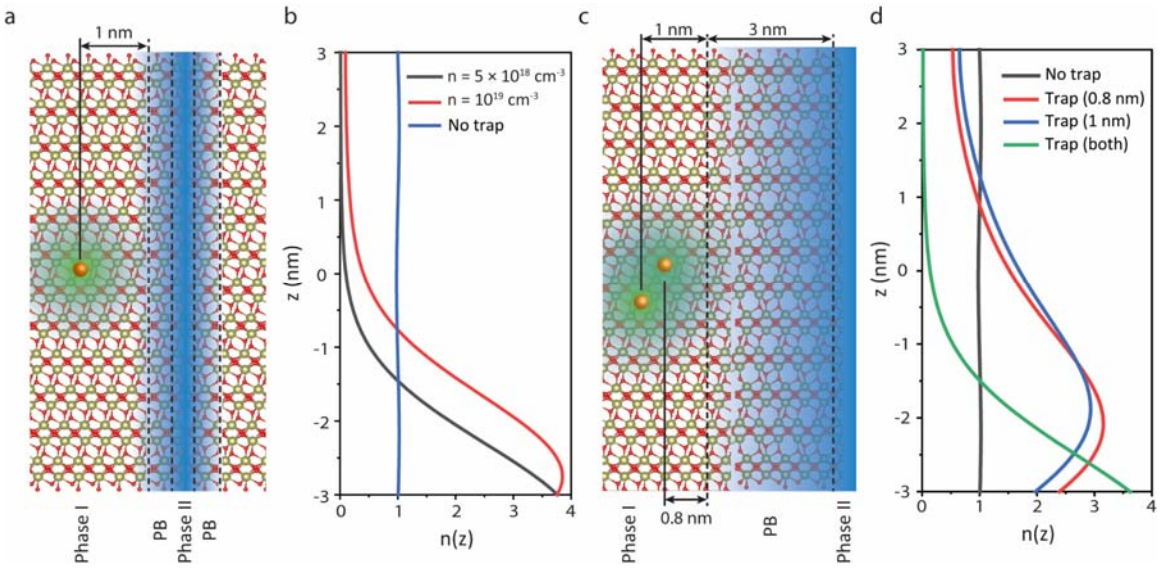


Fig. 3 Trapped-charge-induced conductance change in incomplete conduction channels. (a) The schematic where the RTN-responsible defect (orange) is 1 nm away from

an island-like conduction channel (blue). The channel is formed by a conductive phase region (phase II) and the phase boundary region (PB). (b) Transport electron wave function corresponding to (a). z denotes the position of the channel along the electron transport direction (from -3nm to 3nm), and n(z) shows the normalized integration of the transport electron wave function 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} or 10^{19} cm^{-3} with one electron trapped at the defect, respectively, and the blue line is $n(z)$ with no electron trapped. (c) The schematic where two defects (orange) are away from a channel that is attached to the main conduction channel. The PB region is 3 nm in width in this case. (d) Transport electron wave function corresponding to (c). The red/blue lines represent $n(z)$ when one electron is trapped in the defect 0.8/1 nm away from the channel, respectively, and the green/black lines are $n(z)$ when both/none defects trap electrons. The carrier density in the channel for the simulation is 5×10^{18} cm^{-3}.

The incomplete channel blocking process was modeled as shown in Fig. 3. According to C-AFM experiments, the device region can be classified into three phases: the non-conductive phase (phase Ⅰ), the conductive phase (phase Ⅱ), and the region between them, which features an intermediate conductance (phase boundary, PB). During the programming or denoising operations, these PB regions form or disappear, accompanying the observation of RTN and its annihilation, indicating that some RTN-inducing incomplete channels are located in these PB regions. Fig. 3a shows the schematic of the case where a defect is trapping/detrapping an electron 1 nm away from an island-like incomplete channel whose width is 1 nm. The transport electron wave functions $\psi(x, y, z)$ with / without a trapped charge are plotted in Fig. 3b by the probability density at each cross-section of the channel $n(z) = \int |\psi(x, y, z)|^2 dxdy$ (z is the axis along the channel). This reflects what proportion of the injected electron propagates through the channel. To mimic the case where there are different percentages of phase Ⅱ, two charge carrier densities (averaged over Phase Ⅰ and Phase Ⅱ regions) were used for the simulations. The results suggest 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). Fig. 3c corresponds to another common case as observed in C-AFM, where the incomplete channel is attached to the main channel with multiple charge traps around. Fig. 3d shows that the trapped charge close to the incomplete channel tends to have a bigger impact on conductance than the one far away. It is also observed that the impact of multiple charge traps can enhance each other and lead to a multiplied change of conductance as the thick PB region is completely blocked in this case. Compared to previous models using classical carrier drift-diffusion equations, we employ quantum transport formalism to simulate the influence of charged defects on channel conductivity, confirming that the Coulomb blockade mechanism applies to nanoscale channels in our

case. It can be further 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]

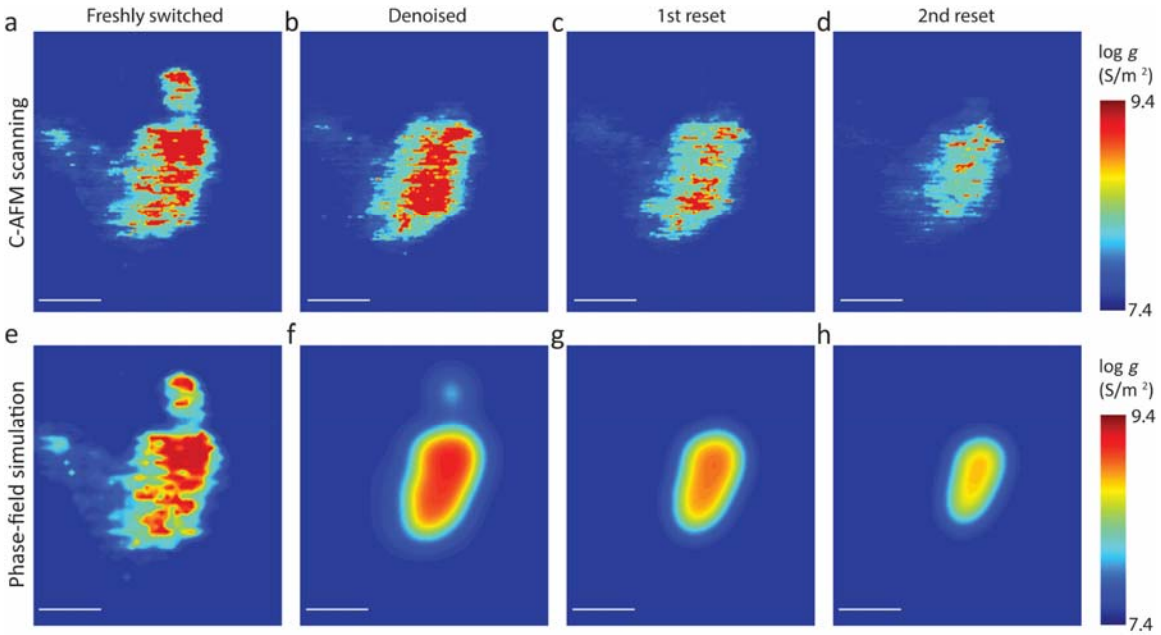


Fig 4. Mechanism of denoising using sub-threshold voltage, identified through C-AFM measurement and phase-field theory simulation. From left to right, the conduction channel is first denoised by a 0.2 V voltage and then RESET with a 0.5 V voltage twice. The dynamics of conductive and insulating phase fields are simulated based on the phase transition energy pathway from the first principle calculation. We hypothesize that the conductive and insulating phases are the orthorhombic phase with a high concentration of oxygen vacancy and the monoclinic phase without oxygen vacancy, respectively. The denoising process is captured by the phase-field relaxation, where the island of the incomplete channel disappears, and the phase boundary sharpens.

As 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. The reason why a 'sub-threshold' voltage that is much smaller than the SET or RSET voltages, can decrease the RTN is explained by 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 in Fig. 3) are the orthorhombic (*o*) and monoclinic (*m*) phases of HfO_2, as the *o* phase is stabilized through a high concentration of oxygen vacancy.[49] The denoising voltage provides a driven force for the phase relaxation by both the temperature effects and the current-induced force, enabling the

system to relax towards an equilibrium state. The free energy F and equation of motion of the system are as follows:

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

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

Where η is the order parameter (here use the monoclinic angle) describing the transition from *m* to *o* 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 principle calculations. Using the phase-field simulation, we derive a similar behavior 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, where the high-energy boundary region is reduced. During the RESET process, the conduction channel shrinks in size, and its conductivity also decreases, as oxygen vacancy is drifted away from the switching-active region by the strong voltage. Intuitively, the incomplete conduction channels, namely 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, they are prone to change (either being completed or annihilated), which can be activated by a sub-threshold voltage. On the other hand, different from the complete main conduction channel, electron transport of incomplete channels can be readily blocked by trapped charges as shown in Fig. 3, making them the main source of RTN noises. The situation is more severe for a conductance state obtained by a SET switching process as conduction channel creation and growth are a positive feedback process, which happens faster and faster and leaves no time for maturation of the newly formed conduction channels before the end of each switching pulse. In the denoising process, there is no need of migration, annihilation or creation of trap sites (e.g., interstitial oxygen defects). Although the specific phases involved may be different for different oxide systems, the approach used here and the conclusions drawn are generally applicable.

In summary, we have achieved 2,048 conductance levels in a memristor, over an order of magnitude higher than previous demonstrations and the highest among all known memories. Importantly, these were obtained in memristors of a fully integrated chip fabricated in a commercial foundry. We have revealed 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 operated on the entire 256 × 256 crossbar using the on-chip driving circuitry designed for regular reading and programming without any extra hardware overhead. These results not only provide critical insights into the microscopic picture of the memristive switching process but also represent a leap forward

in commercializing memristor technology as hardware accelerators of machine learning and artificial intelligence for edge applications. In addition, such analog memristors may also enable electronic circuits capable of growing for the recently proposed mortal computation in the future [30].

Data availability

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

Code availability

The algorithm for memristor high precision programming has been presented in the supplementary information. The code for physical modeling/simulations are available at https://github.com/htang113/HfO2-memristor-denoise/tree/main.

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 Inc. J.J.Y also serves as co-founder and paid consultant of TetraMem Inc. R. M., Q. X., and J. J. Y. were also partially supported by the Air Force Office of Scientific Research (AFOSR) through the MURI program under contract no. FA9550-19-1-0213, the USA Air Force Research Laboratory (AFRL) (Prime Contract Nos.: FA8650-21-C-5405 and FA8750-22-1-0501), and by the National Science Foundation under contract no. 2023752. The authors thank Alan 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. made electrical measurements. H.T., M.R., and J.L. designed and carried out the simulation. J.W., M.R., H.L., H.C., and H.W. designed and carried out the CAFM 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 The authors declare no competing financial interests.

Additional information

Supplementary information: The online version contains supplementary material available at: xxx

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

Reference

[1] L. Chua, "Memristor-the missing circuit element," *IEEE Transactions on circuit theory,* vol. 18, no. 5, pp. 507-519, 1971.

[2] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, "Electrochemical metallization memories—fundamentals, applications, prospects," *Nanotechnology,* vol. 22, no. 25, p. 254003, 2011.

[3] Y. Yang and R. Huang, "Probing memristive switching in nanoionic devices," *Nature Electronics,* vol. 1, no. 5, pp. 274-287, 2018.

[4] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," *Advances in neural information processing systems,* vol. 29, 2016.

[5] W. Wan *et al.*, "A compute-in-memory chip based on resistive random-access memory," *Nature,* vol. 608, no. 7923, pp. 504-512, 2022.

[6] S. Kumar, X. Wang, J. P. Strachan, Y. Yang, and W. D. Lu, "Dynamical memristors for higher-complexity neuromorphic computing," *Nature Reviews Materials,* pp. 1-17, 2022.

[7] C.-X. Xue *et al.*, "A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices," *Nature Electronics,* vol. 4, no. 1, pp. 81-90, 2021.

[8] M. Lanza *et al.*, "Memristive technologies for data storage, computation, encryption, and radio-frequency communication," *Science,* vol. 376, no. 6597, p. eabj9979, 2022.

[9] P. Yao *et al.*, "Fully hardware-implemented memristor convolutional neural network," *Nature,* vol. 577, no. 7792, pp. 641-646, 2020.

[10] W. Zhang *et al.*, "Neuro-inspired computing chips," *Nature electronics,* vol. 3, no. 7, pp. 371-382, 2020.

[11] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature electronics,* vol. 1, no. 6, pp. 333-343, 2018.

[12] M. A. Zidan, J. P. Strachan, and W. D. Lu, "The future of electronics based on memristive systems," *Nature electronics,* vol. 1, no. 1, pp. 22-29, 2018.

[13] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proceedings of the IEEE,* vol. 106, no. 2, pp. 260-285, 2018.

[14] S. Jung *et al.*, "A crossbar array of magnetoresistive memory devices for in-memory computing," *Nature,* vol. 601, no. 7892, pp. 211-216, 2022.

[15] V. K. Sangwan and M. C. Hersam, "Neuromorphic nanoelectronic materials," *Nature nanotechnology,* vol. 15, no. 7, pp. 517-528, 2020.

[16] G. W. Burr, "A role for analogue memory in AI hardware," *Nature Machine Intelligence,* vol. 1, no. 1, pp. 10-11, 2019.

[17] S. Chen *et al.*, "Wafer-scale integration of two-dimensional materials in high-density memristive crossbar arrays for artificial neural networks," *Nature Electronics,* vol. 3, no. 10, pp. 638-645, 2020.
[18] E. J. Fuller *et al.*, "Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing," *Science,* vol. 364, no. 6440, pp. 570-574, 2019.
[19] C. Choi *et al.*, "Reconfigurable heterogeneous integration using stackable chips with embedded artificial intelligence," *Nature Electronics,* pp. 1-8, 2022.
[20] D.-H. Lim, S. Wu, R. Zhao, J.-H. Lee, H. Jeong, and L. Shi, "Spontaneous sparse learning for PCM-based memristor neural networks," *Nature communications,* vol. 12, no. 1, pp. 1-14, 2021.
[21] X. Xu *et al.*, "Scaling for edge inference of deep neural networks," *Nature Electronics,* vol. 1, no. 4, pp. 216-222, 2018.
[22] Y. Sun *et al.*, "A Ti/AlO x/TaO x/Pt analog synapse for memristive neural network," *IEEE Electron Device Letters,* vol. 39, no. 9, pp. 1298-1301, 2018.
[23] S. Stathopoulos *et al.*, "Multibit memory operation of metal-oxide bi-layer memristors," *Scientific reports,* vol. 7, no. 1, pp. 1-7, 2017.
[24] H. Kim, M. Mahmoodi, H. Nili, and D. B. Strukov, "4K-memristor analog-grade passive crossbar circuit," *Nature communications,* vol. 12, no. 1, pp. 1-11, 2021.
[25] M. A. Zidan *et al.*, "A general memristor-based partial differential equation solver," *Nature Electronics,* vol. 1, no. 7, pp. 411-420, 2018.
[26] C. Li *et al.*, "Analogue signal and image processing with large memristor crossbars," *Nature electronics,* vol. 1, no. 1, pp. 52-59, 2018.
[27] C. Mackin *et al.*, "Optimised weight programming for analogue memory-based deep neural networks," *Nature Communications,* vol. 13, no. 1, pp. 1-12, 2022.
[28] S. Choi *et al.*, "SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations," *Nature materials,* vol. 17, no. 4, pp. 335-340, 2018.
[29] P. Yao *et al.*, "Face classification using electronic synapses," *Nature communications,* vol. 8, no. 1, pp. 1-8, 2017.
[30] G. Hinton, "The Forward-Forward Algorithm: Some Preliminary Investigations," *arXiv preprint arXiv:2212.13345,* 2022.
[31] Z. Yan, X. S. Hu, and Y. Shi, "SWIM: Selective Write-Verify for Computing-in-Memory Neural Accelerators," *arXiv preprint arXiv:2202.08395,* 2022.
[32] B. Chen *et al.*, "A memristor-based hybrid analog-digital computing platform for mobile robotics," *Science Robotics,* vol. 5, no. 47, p. eabb6938, 2020.
[33] S. Choi, Y. Yang, and W. Lu, "Random telegraph noise and resistance switching analysis of oxide based resistive memory," *Nanoscale,* vol. 6, no. 1, pp. 400-404, 2014.
[34] D. Ielmini, F. Nardi, and C. Cagli, "Resistance-dependent amplitude of random telegraph-signal noise in resistive switching memories," *Applied Physics Letters,* vol. 96, no. 5, p. 053503, 2010.
[35] F. M. Puglisi, P. Pavan, A. Padovani, L. Larcher, and G. Bersuker, "Random telegraph signal noise properties of HfOx RRAM in high resistive state," in *2012 Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, 2012: IEEE, pp. 274-277.

[36] J.-K. Lee *et al.*, "Extraction of trap location and energy from random telegraph noise in amorphous TiO x resistance random access memories," *Applied Physics Letters,* vol. 98, no. 14, p. 143502, 2011.
[37] F. M. Puglisi, A. Padovani, L. Larcher, and P. Pavan, "Random telegraph noise: Measurement, data analysis, and interpretation," in *2017 IEEE 24th International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*, 2017: IEEE, pp. 1-9.
[38] F. M. Puglisi, N. Zagni, L. Larcher, and P. Pavan, "Random telegraph noise in resistive random access memories: Compact modeling and advanced circuit design," *IEEE Transactions on Electron Devices,* vol. 65, no. 7, pp. 2964-2972, 2018.
[39] Y. Yang, X. Zhang, L. Qin, Q. Zeng, X. Qiu, and R. Huang, "Probing nanoscale oxygen ion motion in memristive systems," *Nature communications,* vol. 8, no. 1, pp. 1-10, 2017.
[40] F. Puglisi, "Noise in resistive random access memory devices," in *Noise in Nanoscale Semiconductor Devices*: Springer, 2020, pp. 87-133.
[41] F. Hui and M. Lanza, "Scanning probe microscopy for advanced nanoelectronics," *Nature electronics,* vol. 2, no. 6, pp. 221-229, 2019.
[42] U. Celano *et al.*, "Three-dimensional observation of the conductive filament in nanoscaled resistive memory devices," *Nano letters,* vol. 14, no. 5, pp. 2401-2406, 2014.
[43] H. Du *et al.*, "Nanosized conducting filaments formed by atomic-scale defects in redox-based resistive switching memories," *Chemistry of materials,* vol. 29, no. 7, pp. 3164-3173, 2017.
[44] F. M. Puglisi, L. Larcher, A. Padovani, and P. Pavan, "A complete statistical investigation of RTN in HfO 2-based RRAM in high resistive state," *IEEE Transactions on Electron Devices,* vol. 62, no. 8, pp. 2606-2613, 2015.
[45] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Statistical fluctuations in HfO x resistive-switching memory: Part II—Random telegraph noise," *IEEE Transactions on Electron Devices,* vol. 61, no. 8, pp. 2920-2927, 2014.
[46] T. Becker *et al.*, "An electrical model for trap coupling effects on random telegraph noise," *IEEE Electron Device Letters,* vol. 41, no. 10, pp. 1596-1599, 2020.
[47] S. Brivio, J. Frascaroli, E. Covi, and S. Spiga, "Stimulated ionic telegraph noise in filamentary memristive devices," *Scientific reports,* vol. 9, no. 1, pp. 1-9, 2019.
[48] F. Miao *et al.*, "Anatomy of a nanoscale conduction channel reveals the mechanism of a high‐performance memristor," *Advanced materials,* vol. 23, no. 47, pp. 5633-5640, 2011.
[49] Y. Zhou *et al.*, "The effects of oxygen vacancies on ferroelectric phase transition of HfO2-based thin film from first-principle," *Computational Materials Science,* vol. 167, pp. 143-150, 2019.

Method

1. Memristor fabrication:

Standard memristor integrated with CMOS driving circuits:

The CMOS part was fabricated in a standard 180 nm process line in a commercial semiconductor manufacturer with exposed tungsten via at the top. Memristors were processed in the same process line with customized materials / recipes. After tungsten via surface oxide cleaning, Pt bottom electrodes were sputtered and patterned on the vias. Memristor holes were defined by etching through a patterned SiO_2 isolation layer (~ 100nm) and stops at the surface of Pt. Resistive switching layer (HfO_2/Al_2O_3) and top electrode (Ti/Ta) were filled into the etched holes sequentially, where the resistive switching layers were fabricated by atomic layer deposition and the top electrode was fabricated by sputter. Finally, standard aluminum interconnect was made to connect the top electrode to bond pads for electrical testing.

Customized memristor for C-AFM measurement

The customized device was fabricated in a university cleanroom on a Si wafer covered by thermally oxidized SiO_2 (~100nm). 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 the 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 the exposed bottom electrode from oxidation and serve as the ground contact during C-AFM measurement.

2. Electrical measurements:

Single device measurement

Electrical measurement of the standard memristor (foundry made complete memristor with top electrode) was performed on a Keysight B1500A semiconductor device analyzer equipped with a B1530A waveform generator/fast measurement unit (WGFMU). To realize the algorithm as shown in Fig. S4, we built a program using C# to control the electrical operations of B1500A.

Array measurement

The schematic of 1-transistor-1-memristor array with on-chip driving circuits and the testing setup is shown in Supplementary information Fig. S7 and Fig. S8.

C-AFM measurement

The C-AFM was performed by the Bruker Dimension Icon system with a conductive probe (SCM-PIT-V2, 0.01-0.025 Ohm-cm of resistivity) under the contact mode. When performing electrical operations including SET, RESET and read, C-AFM probe was at a fixed position. The conduction channel was first formed with a voltage of 4V. During the

in-situ SET/RESET and reading operations for the chosen conduction channel, the setpoint was set at a relatively large number (~80 nm) to increase the strength of the pressing force to make a large contact area between tip and sample surface. The setpoint 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 nm by 150 nm region surrounding the conductive channel. During this measurement, the set point was set to a small value (~10 nm) for high resolution. The relationship between the contact radius and set point can be found in Fig. S11.

3. First principle calculation

The oxygen interstitial defects' atomic and electronic structure are calculated by the density functional theory (DFT) with the projector augmented wave (PAW) method [1] implemented in Vienna *ab initio* simulation package (VASP) [2]. The generalized gradient approximation (GGA) is employed with the Perdew–Burke–Ernzerhof (PBE) exchange-correlation function [3]. The cut-off energy is set as 400 eV, and the *k*-point mesh is sampled by the Monkhorst-Pack method [4] with a separation of 0.2 rad/Å. The atomic structure of the oxygen interstitial defect is constructed by including one oxygen atom in the $2 \times 2 \times 2$ supercell of the *m*-phase HfO_2 crystal. The initial position of the included oxygen atom is set according to ref. [5], and the atomic configuration is fully relaxed. The force on each atom converges to 0.01 eV/Å, 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, Fig. S14, and Fig. S15 are then extracted from DFT calculation results.

Simulation the impact of trapped charge to conductive channel:

We simulate the Coulomb blockade effect through the quantum transport of conduction electron in a cuboid conduction channel, as shown in Fig. 3. The length of the conductive channel is set as $L = 6$ nm to match the channel length in 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 $V(\vec{r})$ acting on the carriers. Assuming the electric conductance outside the channel is negligible, the quantum transport of electron in the channel can be described by the following equations:

$$\begin{cases} \left[-\frac{\hbar^2}{2m^*}\nabla^2 + V(\vec{r})\right]\psi(x,y,z) = (E - E_c)\psi(x,y,z) \\ V(\vec{r}) = \frac{e^2}{4\pi\epsilon_0\epsilon_r}\sum_i \frac{e^{-|\vec{r}-\vec{r}_i|/\lambda_D}}{|\vec{r}-\vec{r}_i|} \\ \psi|_{x=0} = \psi|_{x=d} = \psi|_{y=0} = \psi|_{y=d} = 0 \end{cases} \quad (571)$$

Where m^* is the effective mass of the conductive band of HfO_2, set as $0.11m_e$ according to [6]. E is the eigen energy of the transport electron, set as 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 $\vec{r}_i$, where ϵ_r is the relative dielectric constant (set as 16 according to [7]) and λ_D is the Debye screening length evaluated as $\lambda_D = \sqrt{\frac{\epsilon_0\epsilon_r k_B T}{ne^2}}$ (the temperature T is set as 300 K).

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

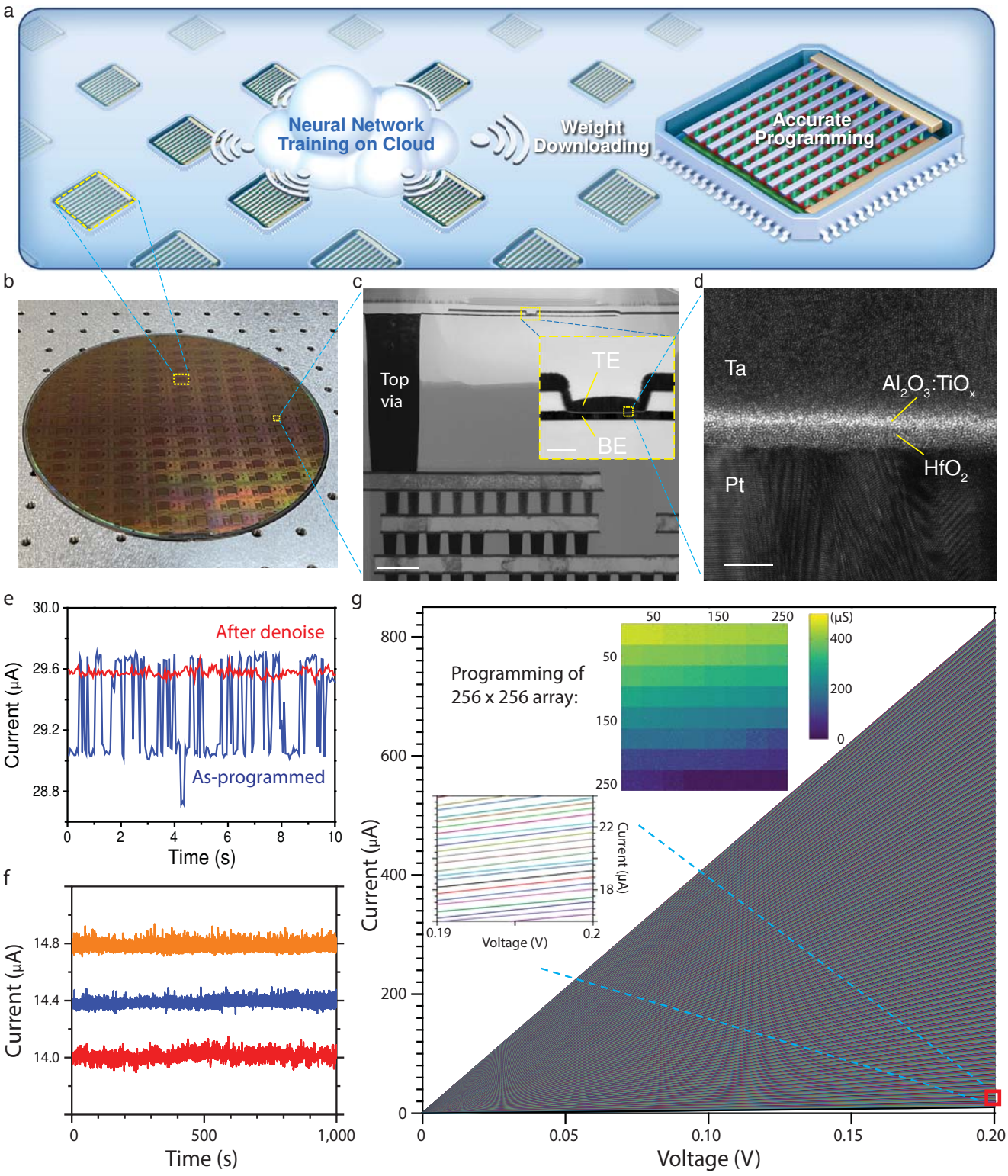
$$\begin{cases} \psi|_{z=0} = e^{ik_{11}z}\sin\frac{\pi x}{d}\sin\frac{\pi y}{d} + \sum_{m,n} T_{mn}e^{-ik_{mn}z}\sin\frac{m\pi x}{d}\sin\frac{n\pi y}{d} \\ \frac{\partial\psi}{\partial z}|_{z=0} = ike^{ikz}\sin\frac{\pi x}{d}\sin\frac{\pi y}{d} - \sum_{m,n} T_{mn}ik_{mn}e^{-ik_{mn}z}\sin\frac{m\pi x}{d}\sin\frac{n\pi y}{d} \\ \psi|_{z=L} = \sum_{m,n} R_{mn}e^{ik_{mn}z}\sin\frac{m\pi x}{d}\sin\frac{n\pi y}{d} \\ \psi|_{z=L} = \sum_{m,n} R_{mn}ik_{mn}e^{ik_{mn}z}\sin\frac{m\pi x}{d}\sin\frac{n\pi y}{d} \\ k_{mn} = \sqrt{\frac{2m^*}{\hbar^2}(E - E_c) - (m^2 + n^2 - 2)\left(\frac{\pi}{d}\right)^2} \end{cases} \quad (580)$$

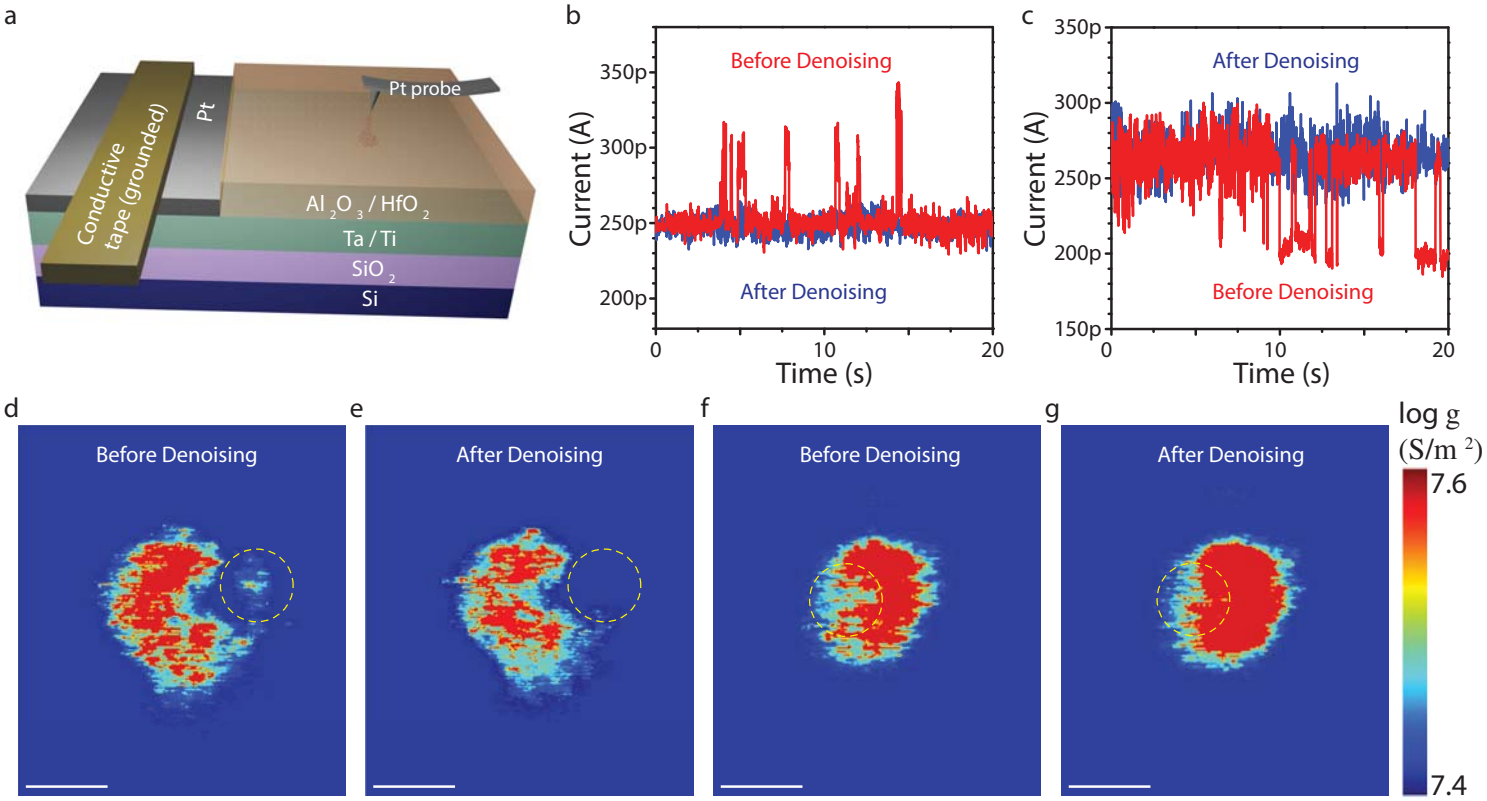
The electron transport is then shown by the probability density function of the electron wave function at each cross section of the channel $n(z) = \int|\psi(x,y,z)|^2 dxdy$, reflecting what proportion of the injected electron propagates through the channel. If $n(L) \simeq 0$, the electron transport 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, as discussed in the results part.

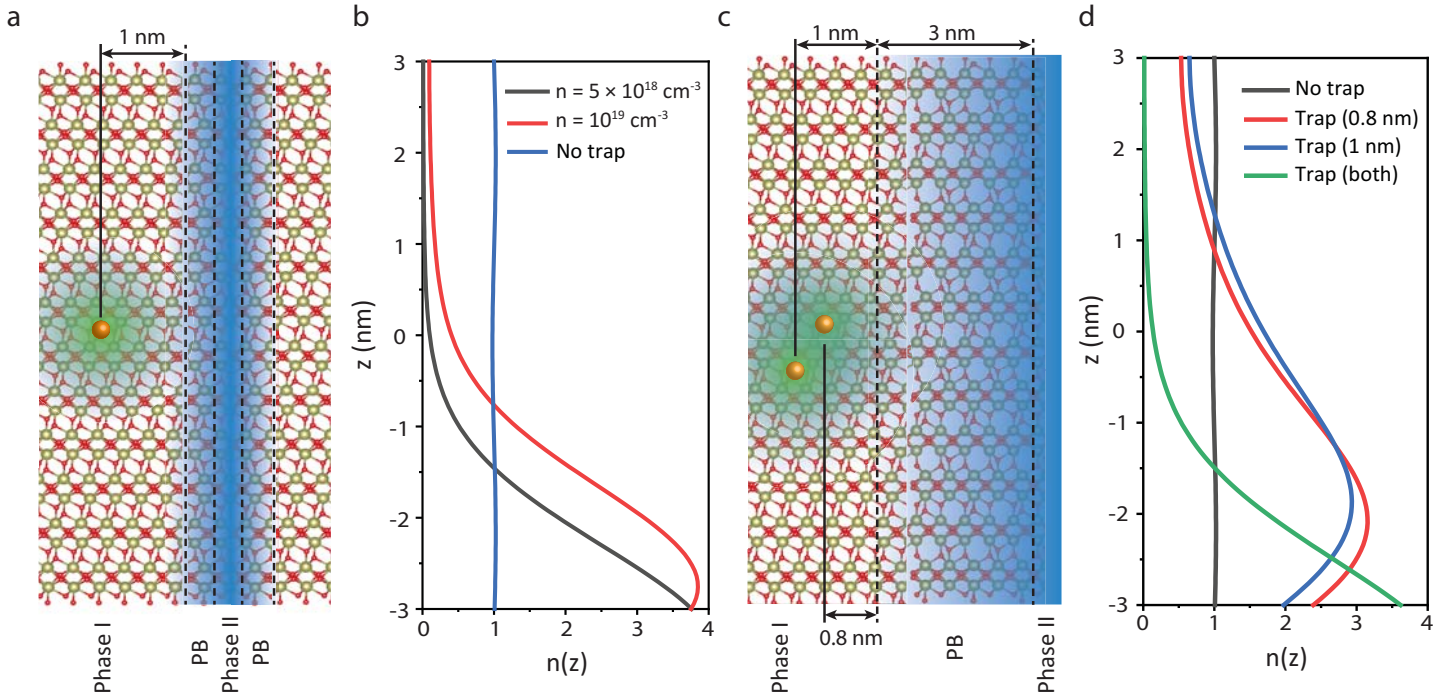
Reference for method:

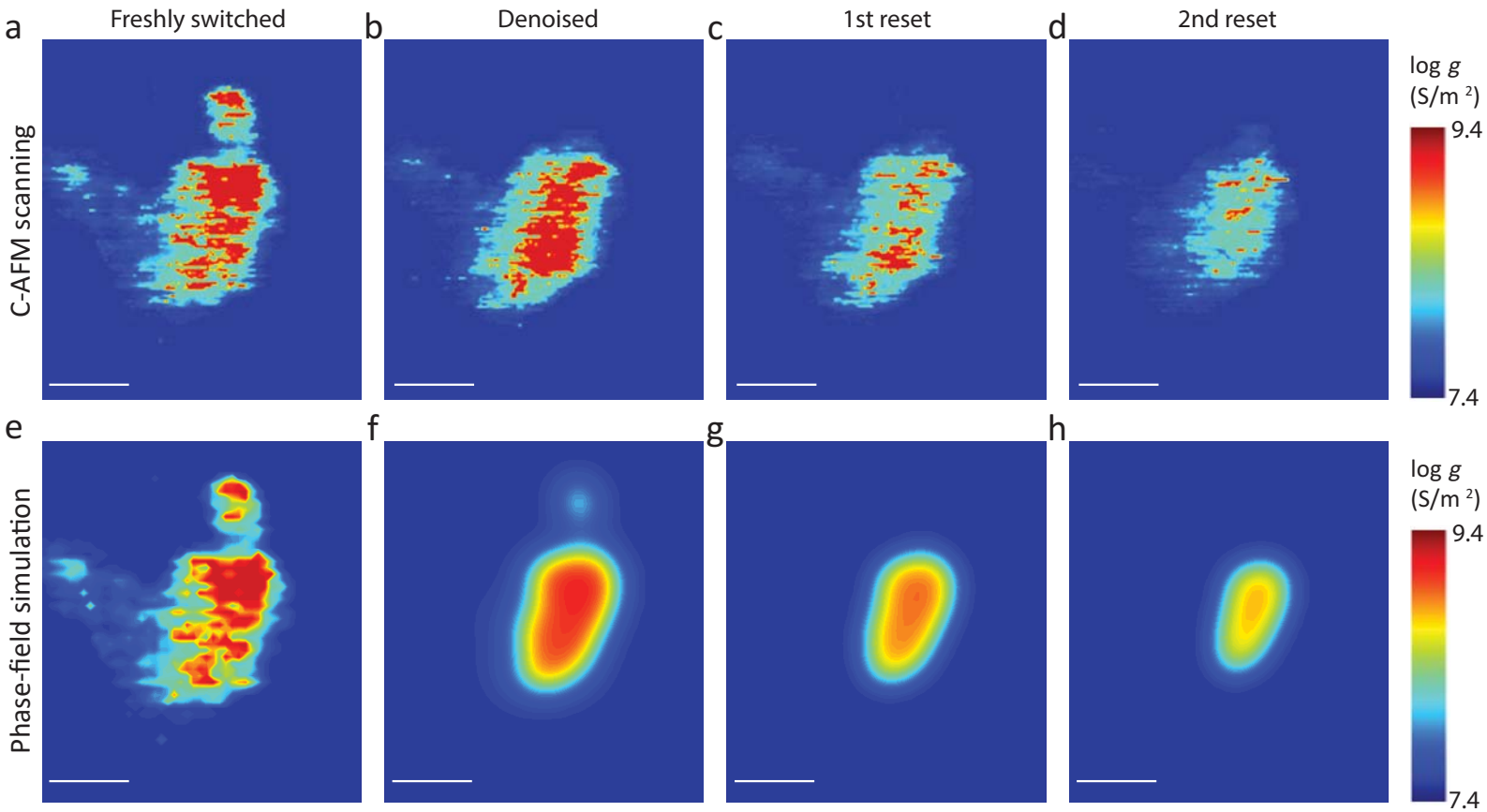
[1] G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," *Physical review b,* vol. 59, no. 3, p. 1758, 1999.

[2] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Physical review B,* vol. 54, no. 16, p. 11169, 1996.
[3] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Physical review letters,* vol. 77, no. 18, p. 3865, 1996.
[4] H. J. Monkhorst and J. D. Pack, "Special points for Brillouin-zone integrations," *Physical review B,* vol. 13, no. 12, p. 5188, 1976.
[5] J. Lyons, A. Janotti, and C. Van de Walle, "The role of oxygen-related defects and hydrogen impurities in HfO2 and ZrO2," *Microelectronic engineering,* vol. 88, no. 7, pp. 1452-1456, 2011.
[6] S. Monaghan, P. Hurley, K. Cherkaoui, M. Negara, and A. Schenk, "Determination of electron effective mass and electron affinity in HfO2 using MOS and MOSFET structures," *Solid-State Electronics,* vol. 53, no. 4, pp. 438-444, 2009.
[7] X. Zhao and D. Vanderbilt, "First-principles study of structural, vibrational, and lattice dielectric properties of hafnium oxide," *Physical Review B,* vol. 65, no. 23, p. 233106, 2002.


a
Neural Network Training on Cloud
Weight Downloading
Accurate Programming
b
c
Top via
TE
BE
d
Ta
Al_2O_3:TiO_x
Pt
HfO_2
e
After denoise
As-programmed
Current (μA)
Time (s)
f
Current (μA)
Time (s)
g
Programming of 256 x 256 array:
(μS)
Current (μA)
Voltage (V)


a
Pt probe
Conductive tape (grounded)
Pt
Al 2O 3 / HfO 2
Ta / Ti
SiO 2
Si
b
Before Denoising
After Denoising
Current (A)
350p
300p
250p
200p
0
10
20
Time (s)
c
After Denoising
Before Denoising
Current (A)
350p
300p
250p
200p
150p
0
10
20
Time (s)
d
Before Denoising
e
After Denoising
f
Before Denoising
g
After Denoising
log g
(S/m 2)
7.6
7.4


a
1 nm
Phase I
PB
Phase II
PB
b
n = 5 × 10^18 cm^-3
n = 10^19 cm^-3
No trap
z (nm)
n(z)
c
1 nm
3 nm
Phase I
0.8 nm
PB
Phase II
d
No trap
Trap (0.8 nm)
Trap (1 nm)
Trap (both)
z (nm)
n(z)


a
Freshly switched
b
Denoised
c
1st reset
d
2nd reset
C-AFM scanning
log g
(S/m 2)
9.4
7.4
e
f
g
h
Phase-field simulation
log g
(S/m 2)
9.4
7.4