Peer Review File

Manuscript Title: Thousands of conductance levels in memristors monolithically integrated on CMOS

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

The manuscript by M. Rao et al. reports on memristive devices with 2048 stable and distinguishable conductance levels. The devices are integrated on CMOS circuits in a standard foundry. The authors have observed and suggested a way to stabilize the conductance states and minimize the noise in the resistance states by applying a subthreshold voltage pulses that lead to homogeneous resistance values. The authors offer as well an explanation about the origin of this stabilization, based on experiments and calculations.
The manuscript is well structured and written and the topic is important and suitable for the broad readership of Nature. The results are highly interesting and can lead to significant improvement on device/application level. The number of conductance levels is to my knowledge the highest reported and integrating 256x256 memristors array on CMOS in a foundry is impressive.
I have however, concerns of different severity regarding the proposed mechanism that need to be carefully addressed prior to a decision on the manuscript.

Comments on the text
The authors have proposed that there are multiple filaments formed during the switching, where one is causing the initial SET and several others are formed incomplete. After applying the denoising procedure these additional filaments either grow or are reduced (partially or completely dissolved). To experimentally support this, cAFM technique was used (e.g. Fig. 2 d-g).

1. cAFM is a surface technique showing a conductivity map of the surface. It is not necessary that the filament(s) continue in depth, keeping the same size and shape. The filament(s) could have an “umbrella” or “tree” shape with its large part located at the interface scanned by cAFM, continuing much thinner in depth. Applying voltages of small magnitudes will than increase the diameter of the top conductive area (Fig. 2f-g) or decrease it (Fig. 2d-e) but this must not necessarily change the filament in depth. Alternatively, the filament(s) could have a cone shape with its large base to the bottom. (See for reference works by A. Kenyon or U. Celano). Applying a denoising voltage to cone shaped filament one will modify the tip of the filament, making it larger or thinner but must not necessarily change it in depth.

2. There are several studies showing that multiple filaments co-exists but no one claimed these filaments are very close to each other. Cross-section TEM images of filaments in VCM/OxRAM devices, have shown more filaments of similar size as in fig.2d-g (about 10 nm) but away from each other (see for example Chem. Mater. 2017, 29, 7, 3164). The shape and structure of the filament(s)

in depth is at that stage not verified.

3. About the discussion on the origin of RTN. I fully agree that the results speak that this can be caused by electronic effects, but in the same time I do not see sufficiently strong arguments that this is not an ionic/atomic effect (or combined). At the filament tip apex or tiny region(s) atoms/ions can jump out or in, depending on the chemical potential gradient and excess surface energy. Such jumps will be related to change in the electronic conductivity and as well will influence the RTN. The interpretation on the electronic nature of RTN may be correct, but is not unequivocally verified by experiment.

The authors have claimed that “Combining the statistical analysis of RTN currents with the atomistic Simulations, we identify that interstitial oxygen defects are responsible for the observed RTNs (defect structure and electronic structure shown in the insets to Fig. 3b). An interstitial oxygen defect hosts a localized electronic state, acting as a trap for electrons. The trap state is an antibonding orbital”

4. This statement and discussion about the role of oxygen interstitials is a very critical point that I cannot agree. Oxygen interstitials are negatively charged species (O^2-) at interstitial lattice position. To keep electroneutrilty their negative charge is compensated by positive charges e.g. electron holes. But definitely not by electrons. The authors should carefully justify this assumption and consult with literature on defect chemistry of HfO2. Most the researchers are of opinion that oxygen interstitials are not participating (or to very little extent) in the ionic conduction due to their large size and high activation energy for transport. Most important is that their chemistry is not allowing to accept more electrons.

5. As the authors are writing in the text “interstitial oxygen and its adjacent oxygen atom” it should be made clear in the text that oxygen (irrespective on regular or interstitial position) is in ionic form (anion). Atomic oxygen (neutral) is practically not present in the crystal lattice (the oxygen partial pressure is about 10^-25 for room temperature).

Technical comments
a. What is the size of the cAFM tip?

b. (C)AFM (even used in vacuum) is sensitive towards surface adsorbates (see for details works by K.Szot). Have the authors used some pre-treatment to remove the adsorbates?

c. Is Ti really in metallic state? In such thickness I would expect at least partial oxidation.

d. The stacks for standard electrical measurements and for cAFM studies are reversed (top and bottom). From chemical point of view, this makes difference and the devices are not equivalent. If oxide is sputtered onto Ti and Ta the oxidation of these metals will be much stronger compared to the case where Ti and Ta are sputtered onto the oxide. Would be a transfer of the experience of one device type to another correct?

Referee #2 (Remarks to the Author):

This paper discusses the underlying physics that previously limited the number of achievable conductance levels in memristors and developed electrical operation protocols to circumvent such limitations. It reports over 2048 conductance levels achieved with memristors in fully integrated chips with 256 x 256 memristor
arrays monolithically integrated on CMOS circuits. The overall paper is well written and easy to follow. However, the reviewer has concerns on a few issues that need to be addressed before the significance of the work can be fully appreciated.

1. Even though the authors mentioned that the 2,048 levels were achieved on the entire array, the paper lacks statistical results on all the memristors. Fig. 1 seems to report the results of only 1 memristor. The uniformity and scalability of the approach remains unclear.

2. The paper does not mention anything about data retention. How long can the data programmed in 2,048 levels be reliably kept? Some additional data is needed here.

3. Comparing with write-verify, denoise method suppresses fluctuation caused by RTN and makes it possible for single memristor to obtain more conductance levels. Unlike precise calculations in scientific computing, neural networks can tolerate small fluctuations. From what is shown in Fig. 1e, assuming a read voltage of 0.2 V, the conductance perturbation is about 2% of the target value (about 148uS), which may be tolerable for the inference accelerator of the neural network. In edge computing application (many are one-step write), speed and energy are also very important metrics. The denoise method has a slight advantage in conductance precision, but is it still better than the write-verify method after comprehensive speed, energy and other factors? Seems to need further clarification. This comparison is especially important given the most recent advances in selective write-verify, see Yan et al, "SWIM: Selective Write-Verify for Computing-in-Memory Neural Accelerators," DAC'22.

4 Fig. 2b and c, the residence time of RTN in c is obviously longer than that in b, may need explanation.

5. Correction: Page 7, “The residence time at the higher … the solid curves in Fig. 2b.” should be Fig. 3b.

Referee #3 (Remarks to the Author):

This work by Rao et al. reports a record high number of conductance levels in memristive devices based on understanding of the conductance fluctuation mechanism and introduction of a critical denoising operation during device programming. Improvement of the weight precision in memristive devices is key to the performance and efficiency of memristor based in-memory computing systems, and hence the results in this work represent a significant step forward, especially for neural network models and applications that demand high-precision weight parameters. Theoretical studies were performed to understand the root cause of conductance fluctuations that previously limit weight precision, and subsequent conductive AFM characterization was employed to understand the role of the proposed denoising process. The manuscript is well structured and written, and the proposed protocol might be important for the research field.

This reviewer, however, has some technical questions on the work, as detailed below.

1. The authors have utilized conductive AFM to characterize the evolution of conduction channels before and after the denoising process in order to clarify the denoising mechanism. This will need to land the AFM tip onto the same spot of the sample surface for performing the denoising operation, accompanied by scanning spatial conductivity maps in the two states. My question is how the location of the tip was calibrated so as to ensure the same position? Will possible shift in tip position affect the mapping results and observed evolution of channels?

2. Figure 2 shows the incomplete channels are either annihilated or completed after denoising process, whereas the base current levels didn't change after denoising. Although the main conductance of the device should be dominated by the main completed channels, certain reduction in the base current level may be expected after annihilation of the incomplete channels. More importantly, completing the partial filament to form fully connected filament should increase the conductance, because it becomes part of the main channel and the overall filament grows larger. This seems different from the results in Figure 2b-c, and further clarification on this point is needed.

3. Is it of general applicability to attribute conductance fluctuation in memristors solely to RTN? This work shows that set operation tends to induce larger fluctuation than reset operation. This is different from many previous reports, which reveals stronger fluctuations in high resistance states, and therefore may suggest existence of different physical pictures. A few examples can be found in Refs. 30, 31, etc. The configuration of the conduction channel could be changed for the chosen conductance state, either in different cycles or different devices, and that is related to 'atomic effect' instead of purely 'electronic effect'. Further clarification on this point is needed.

4. How to define the number of states is essential for this work and needs to be justified more clearly. This work used a fixed value of 2 uS as the interval between every two neighboring states, and therefore achieved 2048 states between 50~4144 uS (about 4094 us range). It is also mentioned that the amplitude of variation is proportional to the actual current level, namely higher conductance state has larger variations. Is the 2 uS interval sufficient for those states with high conductance? Fig. 1f and the inset of Fig. 1g present zoomed-in images for rather low conductance states, but showing it for high-conductance states as well will be more convincing.

5. The tuning process for the high precision programming as depicted in Fig. S3 is quite complex and requires a large number of iterative operations. This is OK for small arrays but will easily lead to extremely long programming time when loading weights into memristor neural networks for chips with large memristor capacity. I would recommend the authors to characterize the programming time for each cell and array, and ideally project it to NNs with practical capacities. Moreover, although the denoising process doesn't involve additional circuitry, the closed-loop programming still leads to large overhead in periphery circuitry, which requires careful evaluations.

6. This work has experimentally and theoretically clarified the correlation between RTN and conductance fluctuation/programming precision in memristive devices, and in turn proposed a denoising process that allows high-precision programming. Although this is important for the application of memristive neural networks, the study and understandings on the RTN mechanism itself have been covered by previous works in Refs. 29, 30, 39, 40 & 41, etc. The authors are recommended to point out the new understandings revealed in this work more clearly.

Author Rebuttals to Initial Comments:

Text in blue: Questions / comments from reviewers

Text in black: Response from the authors

Text in red: Changes made to the manuscript

Text in green: Text in the original submission

*The marked change location in manuscript correspond to the version of the manuscript file with changes highlighted (instead of the clean version).

Reviewer #1

The manuscript by M. Rao et al. reports on memristive devices with 2048 stable and distinguishable conductance levels. The devices are integrated on CMOS circuits in a standard foundry. The authors have observed and suggested a way to stabilize the conductance states and minimize the noise in the resistance states by applying a subthreshold voltage pulses that lead to homogeneous resistance values. The authors offer as well an explanation about the origin of this stabilization, based on experiments and calculations.
The manuscript is well structured and written and the topic is important and suitable for the broad readership of Nature. The results are highly interesting and can lead to significant improvement on device/application level. The number of conductance levels is to my knowledge the highest reported and integrating 256x256 memristors array on CMOS in a foundry is impressive.
I have however, concerns of different severity regarding the proposed mechanism that need to be carefully addressed prior to a decision on the manuscript.

We thank the reviewer for the positive comments.

The authors have proposed that there are multiple filaments formed during the switching, where one is causing the initial SET and several others are formed incomplete. After applying the denoising procedure these additional filaments either grow or are reduced (partially or completely dissolved). To experimentally support this, cAFM technique was used (e.g. Fig. 2 d-g).

1. cAFM is a surface technique showing a conductivity map of the surface. It is not necessary that the filament(s) continue in depth, keeping the same size and shape. The filament(s) could have an “umbrella” or “tree” shape with its large part located at the interface scanned by cAFM, continuing much thinner in depth. Applying voltages of small magnitudes will than increase the diameter of the top conductive area (Fig. 2f-g) or decrease it (Fig. 2d-e) but this must not necessarily change the filament in depth. Alternatively, the filament(s) could have a cone shape with its large base to the bottom. (See for reference works by A. Kenyon or U. Celano). Applying a denoising voltage to

cone shaped filament one will modify the tip of the filament, making it larger or thinner but must not necessarily change it in depth.

We thank the reviewer for bringing up this good point. We agree that in addition to the possibility of multiple filaments we mentioned in the manuscript, the electrical conduction may also go through a 'single' filament with either a cone (smaller top end) or tree (smaller bottom end) shape. Among these two shapes, we think the cone shape is more likely in the specific device we used for cAFM because the cAFM data has revealed that the area size of the top end of the filament can significantly affect the total conductance, which is not expected for a tree shape (i.e., removing a branch of the crown of a tree may not change its conductance significantly). A cone shape is also consistent with the material stack of the cAFM device, where active Ta electrode is at the bottom and inert electrode Pt (AFM tip) is on the top. Our cross-sectional TEM data for such material stack in *Scientific reports* 6.1 (2016): 1-8 has revealed a cone shape filament with its narrow end pointing up. This result is also consistent with cAFM results in *Nano letters* 14.5 (2014): 2401-2406 (mentioned by the reviewer), where the filament is also larger at active metal side.

Nevertheless, the shape or number of the filament(s) is not the focus of the manuscript and does not change the denoise mechanisms or operation protocols. A 'single' filament may be viewed as multiple filaments in parallel for the denoising purpose. Strictly speaking, cAFM is more than a surface technique as it reflects the electrical conduction mapping underneath the surface. Annihilating a filament is to electrically 'break' any part of it in the 3D critical conduction path, not necessarily being the surface or bottom.

To include all the possibilities and make the manuscript more rigorous, we added the following sentence in the revision:

(Paragraph 1 on page 6) "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."

We also included *Nano letters* 14.5 (2014): 2401-2406 into the reference list:

(Paragraph 2 on page 5)"…Visualizing the evolution of conduction channels during electrical operations is informative for this purpose.[37-40] …"

2. There are several studies showing that multiple filaments co-exists, but no one claimed these filaments are very close to each other. Cross-section TEM images of filaments in VCM/OxRAM devices, have shown more filaments of similar size as in fig.2d-g (about 10 nm) but away from each other (see for example Chem. Mater. 2017, 29, 7, 3164). The shape and structure of the filament(s) in depth is at that stage not verified.

As mentioned above, we agree with the reviewer the seemingly isolated island(s) in cAFM may or may not belong to the same big filament. If they are isolated multiple

filaments in our cAFM devices, the reason they are closer might be related to the way they were formed, namely they were created by cAFM tip within a very small region, as shown in Fig. 2 d-g.

In addition to adding the following sentence:

(Paragraph 1 on page 6) "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."

We now also included this important reference in the revision:

(paragraph 2 on page 5) "…These incomplete channels were formed together with complete channels but are smaller in size. Such incomplete filaments were also observed in SrTiO3-based resistive switching devices. [41] A memristor can be denoised by eliminating incomplete channels…"

3. About the discussion on the origin of RTN. I fully agree that the results speak that this can be caused by electronic effects, but in the same time I do not see sufficiently strong arguments that this is not an ionic/atomic effect (or combined). At the filament tip apex or tiny region(s) atoms/ions can jump out or in, depending on the chemical potential gradient and excess surface energy. Such jumps will be related to change in the electronic conductivity and as well will influence the RTN. The interpretation on the electronic nature of RTN may be correct, but is not unequivocally verified by experiment.

We thank the reviewer for raising the important question about the possibility of the atomic effect on RTN. Indeed, the atomic effect can also cause fluctuation and RTN as shown in *Scientific reports,* vol. 9, no. 1, pp. 1-9, 2019. Atoms jumping in and out of the filament region can influence the channel's conductivity and lead to RTN noise. However, we noticed that the ionic RTN reported there is driven by purposedly designed periodic voltage (0.7V pulses) stimuli, in contrast to a constant small voltage (e.g. 0.1-0.2 V) used for reading purpose in our measurement condition. Even driven by the regular periodic voltage stimuli, the ionic RTNs were random and irregular in amplitude, in constrast ot our obervations. While the chemical potential gradient and excess surface energy normally drive the ions/atoms towards one direction, the *back-and-forth fluctuation* changes are more likely associated with thermal fluctuations. In addition, atomic effect cannot explain the following experimental observation in our measurement: as shown in Fig. 3b, τ_1 is negatively related to the reading voltage while τ_0 is almost independent of the reading voltage.

This is well explained if we assume the RTN is of electronic origin. But if we assume the RTN has an atomic origin, the lifetime of each state should be determined by the transition state theory:

$$\frac{1}{\tau_1} = \nu_1^{'} \exp\left(-\frac{E_s - E_1 + qV_{s1}}{k_B T}\right), \qquad \frac{1}{\tau_0} = \nu_0^{'} \exp(-\frac{E_s - E_0 + qV_{s0}}{k_B T})$$

where E_1, E_2, E_s are the 1 state, 0 state, and saddle point energy; $\nu_1^{'}$ and $\nu_0^{'}$ are the effective oscillation frequency of 1 and 0 states; V_{s1} and V_{s0} are the electric voltage between the saddle point and 1 and 0 states, respectively. Therefore, we expect that when the voltage increases, both τ_1 and τ_0 should change because of the change of V_{s0}, V_{s1} (electric effect), and T (temperature effect of electric current). It is hard to explain why τ_0 is almost constant and exhibits a sharp contrast with τ_1. Therefore, we think the data indicates that the RTN in our experiment is unlikely to be an atomic effect.

Related changes have been made in Supplementary information SI-16 and mentioned in main text:

(Paragraph 1 on page 10) "…so τ_0 (τ_e) is approximately independent of the applied voltage, which is inconsistent with atomic motion induced RTNs (see SI-16)"

The authors have claimed that "Combining the statistical analysis of RTN currents with the atomistic Simulations, we identify that interstitial oxygen defects are responsible for the observed RTNs (defect structure and electronic structure shown in the insets to Fig. 3b). An interstitial oxygen defect hosts a localized electronic state, acting as a trap for electrons. The trap state is an antibonding orbital"

4. This statement and discussion about the role of oxygen interstitials is a very critical point that I cannot agree. Oxygen interstitials are negatively charged species (O^2-) at interstitial lattice position. To keep electroneutrilty their negative charge is compensated by positive charges e.g. electron holes. But definitely not by electrons. The authors should carefully justify this assumption and consult with literature on defect chemistry of HfO2. Most the researchers are of opinion that oxygen interstitials are not participating (or to very little extent) in the ionic conduction due to their large size and high activation energy for transport. Most important is that their chemistry is not allowing to accept more electrons.

We thank the reviewer for mentioning this unclarity. We agree with the reviewer's comments that O_i^{2-} can not accept more electrons. As the reviewer mentioned, the oxygen interstitial O_i^{2-} traps two holes to compensate its negative charge and keep electroneutrality: $O_i^{2-} + 2h^+ = O_i^0$. When we previously wrote that the oxygen interstitial traps electron in the manuscript, *we actually meant that the charge-neutral oxygen interstitial O_i^0 (that has already trapped two holes) traps one electron*, or namely, *fills one hole*:

$$O_i^0 + e^- = (O_i^{2-} + 2h^+) + e^- \rightarrow (O_i^{2-} + h^+) \rightarrow O_i^{1-}$$

We previously referred to this as an interstitial oxygen defect that traps one electron, while we agree with the reviewer that this description is confusing. We made changes to our description to clarify this point with proper references included:

(Paragraph 3 on page 7) "… A close-shell interstitial oxygen ion O_i^{2-} attracts two holes h^+ to compensate its charge and form an interstitial oxygen defect in neutral states, O_i^0 [PHYSICAL REVIEW B 81, 161201(R) (2010)]. O_i^0 hosts a localized empty σ^* state in the upper band gap [Microelectronic Engineering 80 (2005) 408–411], acting as a trap for electrons. The trap state is an antibonding orbital of the interstitial oxygen and the adjacent oxygen. After capturing an electron that annihilates one h^+ in the O_i^0, the defect becomes negatively charged O_i^{1-} [Microelectronic Engineering 80 (2005) 408–411], and the trap state is occupied by a spin-up electron (Fig. S14) …"

As suggested by the reviewer, we double confirm the literature on the defect chemistry of HfO_2 [Microelectronic Engineering 80 (2005) 408–411; PHYSICAL REVIEW B 81, 161201(R) (2010)]. We see that the charge-neutral state O_i^0 of oxygen interstitial defect has an empty trap state in the band gap close to the conduction band minimum, showing that it is possible to trap an electron from the conduction band and form O_i^{1-}, which is also a widely studied charge state [Microelectronic Engineering 80 (2005) 408–411; PHYSICAL REVIEW B 81, 161201(R) (2010); Microelectronic Engineering 84 (2007) 2370–2373]. Furthermore, we check with a previous HfO_2 memristor experiment in ref. ["Random Telegraph Noise Nano-spectroscopy in High-κ Dielectrics Using Scanning Probe Microscopy Techniques" in *Noise in Nanoscale Semiconductor Devices*: Springer, 2020, pp. 417-440] (we also include citation to this in paragraph 3, page 7), where the neutral-state oxygen interstitial defect O_i^0 is also identified as electron trap defects. As we also find the electron capture and emission time simulated by the O_i^0 defect parameters consistent with our experiments, we think the evidence from the literature and our data is sufficient to claim that O_i^0 acts as an electron trap in the RTN mechanism.

We also agree with the reviewer's point that the oxygen interstitials are not participating in ionic conduction, and our first-principles calculation also confirms that they have large sizes and high activation energy. While our point on oxygen interstitials is that they can have charge state switching in a long timescale as a source of RTN and affect the conductivity of adjacent incomplete conductive filaments (made of oxygen vacancies), though the oxygen interstitials themselves do not significantly contribute to the electron conductivity. In the denoising process, there is no migration/creation/annihilation of oxygen interstitial involved. Rather, the oxygen vacancies are the mobile species that can complete or annihilate incomplete filament (either a single filament or a part of a large filament with branches), which makes the conductance of the previous incomplete filament region not sensitive to the impact of electron trapping/detrapping process anymore. Note that the difference between the complete and incomplete filaments is that the latter has a segment with a much lower oxygen vacancy concentration, which make it much more sensitive to the impact of trapping/detrapping process, as shown in Fig. 3 c, d, e, f.

We have now pointed it out in the main text:

(Paragraph 1 on page 12) "…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. O_i^0). Although the specific phases involved may be different…"

5. As the authors are writing in the text "interstitial oxygen and its adjacent oxygen atom" it should be made clear in the text that oxygen (irrespective on regular or interstitial position) is in ionic form (anion). Atomic oxygen (neutral) is practically not present in the crystal lattice (the oxygen partial pressure is about 10^-25 for room temperature).

We thank the reviewer for pointing out this terminology issue.

We correct the "oxygen atom" into "oxygen ion" and double-check the whole manuscript to clarify that the elements are in ionic form. Other changes include:

Fig. 3 caption: oxygen atoms -> oxygen ions; Hf atoms -> Hf ions

Technical comments
a. What is the size of the cAFM tip?

As mentioned in the method part, the conductive tips used here was SCM-PIT-V2 with a tip radius around 25 nm. During the measurement, the contacting between the tip and sample surface would result in the deformation in both of them, leading to that the contact radius is dependent on the force applied on the tip. In previous works, it was shown however that even standard tips with a total tip radius in the range of 20 nm have an effective contact radius in the sub-nanometre range. [*Appl. Surf. Sci., 2007, 253 , 3615—3626; J. Appl. Phys., 2015, 117 , 214305*]. Here, we also estimated the contact radius by using the standard calibration sample from the vendor. A relationship between the estimated contact radius and applied force (which is proportional to the value of the set point) is given in the followed figure. When the set point was set to 10 nm, which was used for the filament scanning, the contact radius is about 1.2 nm, which is small enough to resolve the details of the filament. When the set point was set to 80 nm, which was used for the in situ denoising, the contact radius is about 64.4 nm, which is large enough to cover the entire filament area (with a diameter around 20 nm here).

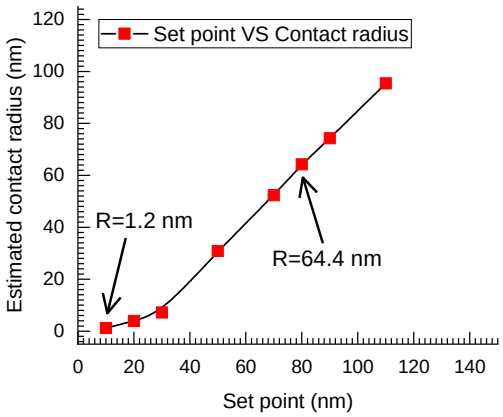

Fig. R1-1. The relationship between the estimated contact radius and set point.

Fig.R1-1 has been included in supplementary information as Fig. S11 and referred to in the main text and *Method*:

(Paragraph 2 on page 5) "…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…"

(Method: C-AFM measurement section) "…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."

b. (C)AFM (even used in vacuum) is sensitive towards surface adsorbates (see for details works by K.Szot). Have the authors used some pre-treatment to remove the adsorbates?

Indeed, the surface absorbates are very critical to the CAFM measurement. To remove the surface absorbates, we did the followed pre-treatment. First, the sample was cleaned by ultrasonic wave for 10 minutes in the acetone, isopropyl alcohol, and deionized water sequentially. Then, using the III-V etcher (Oxford), we cleaned the sample surface with O_2 plasma (100W of power, 30s of treatment time, 50sccm of pressure). The cleanness (statistics the roughness via measured height) of the surface before and after treatment are shown in Fig. R1-2. After treatment the roughness (the full width at half maximum of height distribution) of sample reduced from 0.427 nm to 0.232 nm, indicating the high cleanness of our sample surface.

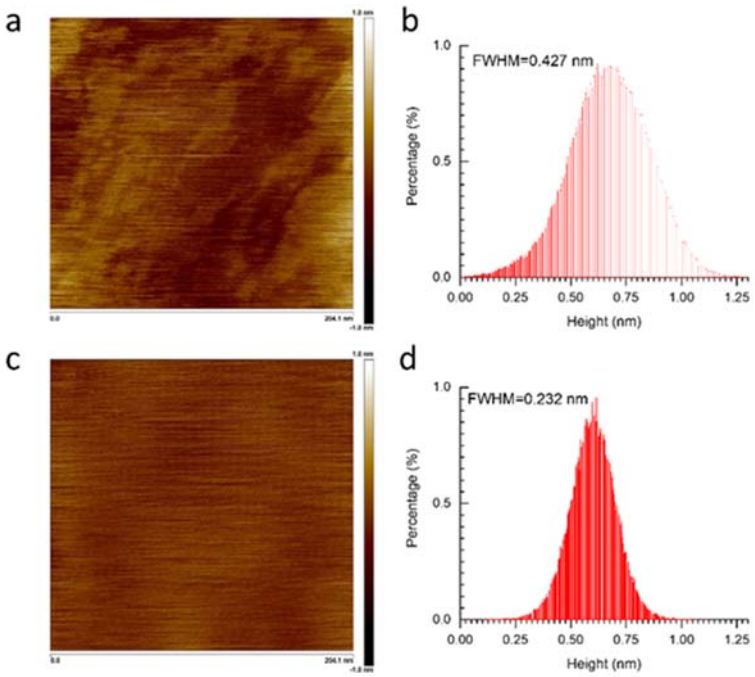

Fig. R1-2. Cleanness of the sample surface. (a) high-resolution height measurement by AFM for the sample surface before treatment. (b) the statistics of the height in (a), the full width at half maximum (FWHM) of the statistics profile was given as the parameter to indicate the roughness. (c) high-resolution height measurement by AFM for the sample surface after treatment. (d) the statistics of the height in (c), the full width at half maximum (FWHM) of the statistics profile was given as the parameter to indicate the roughness.

c. Is Ti really in metallic state? In such thickness I would expect at least partial oxidation.

Thanks for pointing this out, we revealed this point using the electron energy loss spectroscopy (EELS) of the memristor and observed that Ti is indeed partially oxidized.

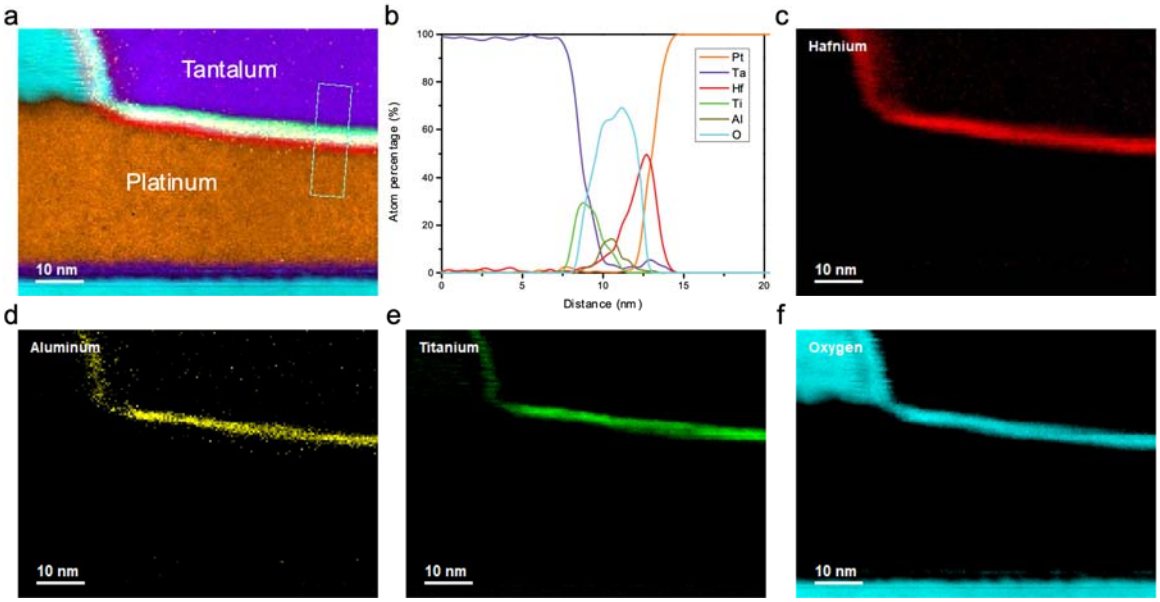

Fig. R1-3 The electron energy loss spectroscopy (EELS) elemental image of the memristor shown in Fig. 1. a) The stacked profile of all elements. b) Atomic percentage in the rectangle marked in a). c) to f) The elemental distribution of c) Hafnium, d) Aluminum, e) Titanium, f) Oxygen.

We changed the label "Ti" to "TiO_x" Fig. 1d.

We are now including Fig. R1-3 as supplementary information Fig. S1:

(Paragraph 2 on page 2)"…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…"

d. The stacks for standard electrical measurements and for cAFM studies are reversed (top and bottom). From chemical point of view, this makes difference and the devices are not equivalent. If oxide is sputtered onto Ti and Ta the oxidation of these metals will be much stronger compared to the case where Ti and Ta are sputtered onto the oxide. Would be a transfer of the experience of one device type to another correct?

We totally agree with the reviewer on that the device for CAFM has a reversed structure compared to the device for standard electrical measurements and their characteristics may not be exactly the same. The reason we did this reversing is that among the three electrode metals (Pt, Ti, Ta), only Pt is suitable as the probe material. We also tried using Ta as the CAFM probe to form a CAFM sample with Pt as bottom electrode and HfO_2 (3 nm) as resistive switching layer. The result turns out that the device cannot be electroformed. The reason is likely that Ta probe gets oxidized in air, resulting in a thick TaO_x layer covering Ta metal. Therefore, we believe the CAFM sample we designed is

the best device that can be used to mimic the standard device for mechanism study under such constraints.

Nevertheless, we have observed similar CAFM and denoise behavior with different types of oxides, which suggest the independence of oxides and general applicability of the conclusions. As one example, we performed CAFM experiment as shown in Fig. S12 which does not have Ti and shows very similar denoising behavior as the device with Ti, suggesting that the denoising mechanism we proposed is valid in different oxde material systems and stacks. The slight difference between the standard device and CAFM device may not affect our general conclusions.

Referee 2:

This paper discusses the underlying physics that previously limited the number of achievable conductance levels in memristors and developed electrical operation protocols to circumvent such limitations. It reports over 2048 conductance levels achieved with memristors in fully integrated chips with 256 x 256 memristor arrays monolithically integrated on CMOS circuits. The overall paper is well written and easy to follow. However, the reviewer has concerns on a few issues that need to be addressed before the significance of the work can be fully appreciated.

We thank the reviewer for the positive comments.

1. Even though the authors mentioned that the 2,048 levels were achieved on the entire array, the paper lacks statistical results on all the memristors. Fig. 1 seems to report the results of only 1 memristor. The uniformity and scalability of the approach remains unclear.

Thanks for the questions of uniformity and scalability. 2,048 levels can be achieved with a randomly chosen device on the chip. As mentioned in the manuscript, the on-chip driving circuitry in our chips has 6-bit ADC and DACs. Measuring all the devices in the 256x256 arrays is only possible by using the on-chip 6-bit (64 levels) driving circuits, with which we have shown highly uniform programming of the entire array into sixty-four 32x32 blocks, and each block is programmed into one of the 64 conductance levels (as shown in Fig. 1g, which is already the state of the art for such large arrays). In order to demonstrate the capability of 2048 levels, we need to use an instrument with over 11 bit precision, such as Keysight B 1500 as we adopted in our measurements, which however, only allows one to program the device one by one and is very time consuming. Nevertheless, the verification of uniformity and programmability can also be achieved by programming randomly chosen devices in arrays or dies to pre-determined target values with a pre-determined distribution tolerance. In such experiment, we went a step further to *show the uniformity not only across an array but also cross the entire 8-inch wafer*. To verify the uniformity and scalability of our high precision programming approach, we

measured 67 dies (50% of all available dies) evenly distributed on an entire 8-inch wafer (numbered in red, as shown in Fig. R2-1). For each of the 67 dies we programmed a randomly chosen memristor into 6 pre-determined target conductance levels (100 μS, 105 μS, 110 μS, 490 μS, 495 μS, 500 μS) with 2 μS variation tolerance (critera for 2048 levels) and analyzed the programming uniformity, as shown in Fig. R2-2. *Such contour map analysis is widely used in standard commercial foundries to evaluate wafer uniformity*. The excellent uniformity of devices across the entire 8-inch wafer can be seen by how narrow the conductance scale bars (2 μS) are in panels a-f and also by how steep the cumulative probability curves (67 data points on each curve corresponding to 67 dies) are in panels g and h. Such level of uniformity and scalability has never been demonstrated before.

Fig. R2-1 The photo of wafer used for programming uniformity measurement. Red numbers marked the dies that were measured.

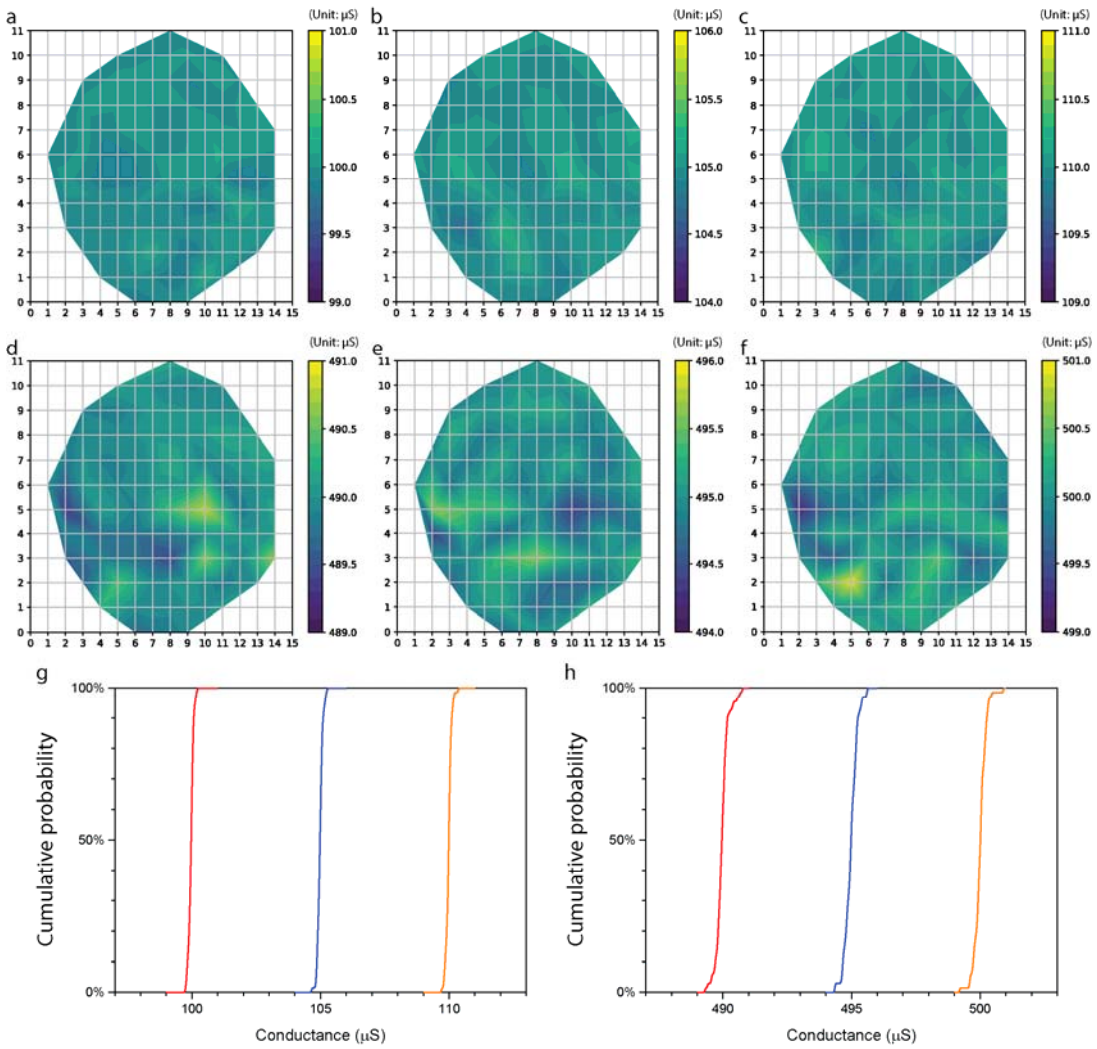

Fig. R2-2 Programming result of randomly chosen memristors on 67 different dies of an 8 inch wafer. a) to f) The contour map of conductance as a result of memristors on the same wafer but different dies being programed into 6 different conductance state. In the 15 by 11 matrix, each block that has color other than white correspond to a die used in the measurements. g) Cumulative probability curve of conductance as a result of memristors being programmed into 100 μS, 105 μS, 110 μS. h) Cumulative probability curve of conductance as a result of memristors being programmed into 490 μS, 495 μS, 500 μS.

Fig. R2-1 and Fig. R2-2 are now included in supplementary information as Fig. S6 and mentioned in the main text:

(Paragraph 2 on page 3) "…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"

2. The paper does not mention anything about data retention. How long can the data programmed in 2,048 levels be reliably kept? Some additional data is needed here.

We thank the reviewer for bringing up this important question. To demonstrate the data retention, we sampled three devices in three different conductance states and used elevated temperature (125℃) to accelerate the measurement process as a great retention at room temperature has been shown with a similar device stack before (*Scientific reports* 6.1 (2016): 1-8). The conductance of three devices were measured with 0.2V at room temperature at certain time points (e.g. 10, 50 and 100 hours) after they were kept at 125 ℃. The largest difference measured for the three devices across 100hrs can be seen from Fig. R2-3, where the conductance change remained smaller than 2 μS (defined in our 2,048 level criteria) for all the devices.

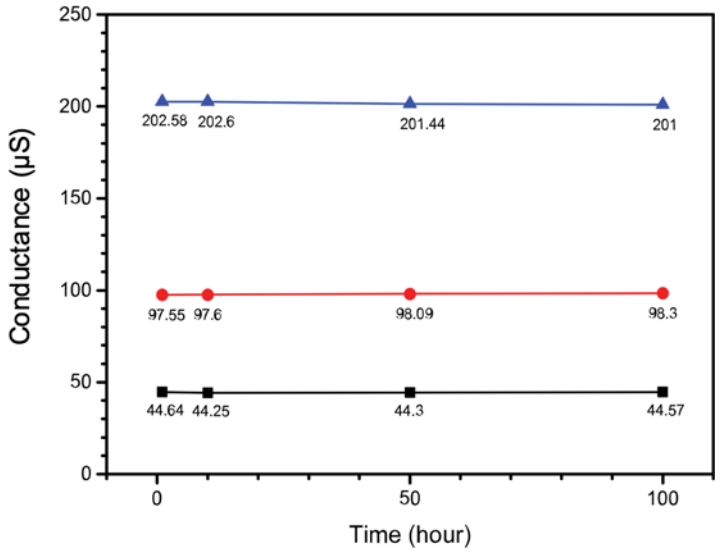

Fig.R2-3 Device conductance remains for 100 h at 125 ℃.

To measure the retention statistics, we programmed 2400 devices to three representative conductance levels: 40 μS (zone 1), 200 μS (zone 2), 550 μS (zone 3) and kept reading them over time at 85 ℃. The drift of conductance was read at 30, 50, 70, 90 hours, as shown in Fig. R2-4. The histograms of change of conductance from initial reading are centered at zero for all measurements, suggesting that there is no trend of drifting during the measurement period.

In addition, as shown in Fig. R2-4, only a very small number of devices exhibit a condtutance change over 2 uS (2uS is the criterion for 2048 levels) over the course the retention test. Given that the fluctuations observed in the array measurements are a combined result of array parasitic, on-chip driving circuit, communication path from chip

to controlling FPGA, and so on, the devices exhibit a great retention at multiple condutacne levels.

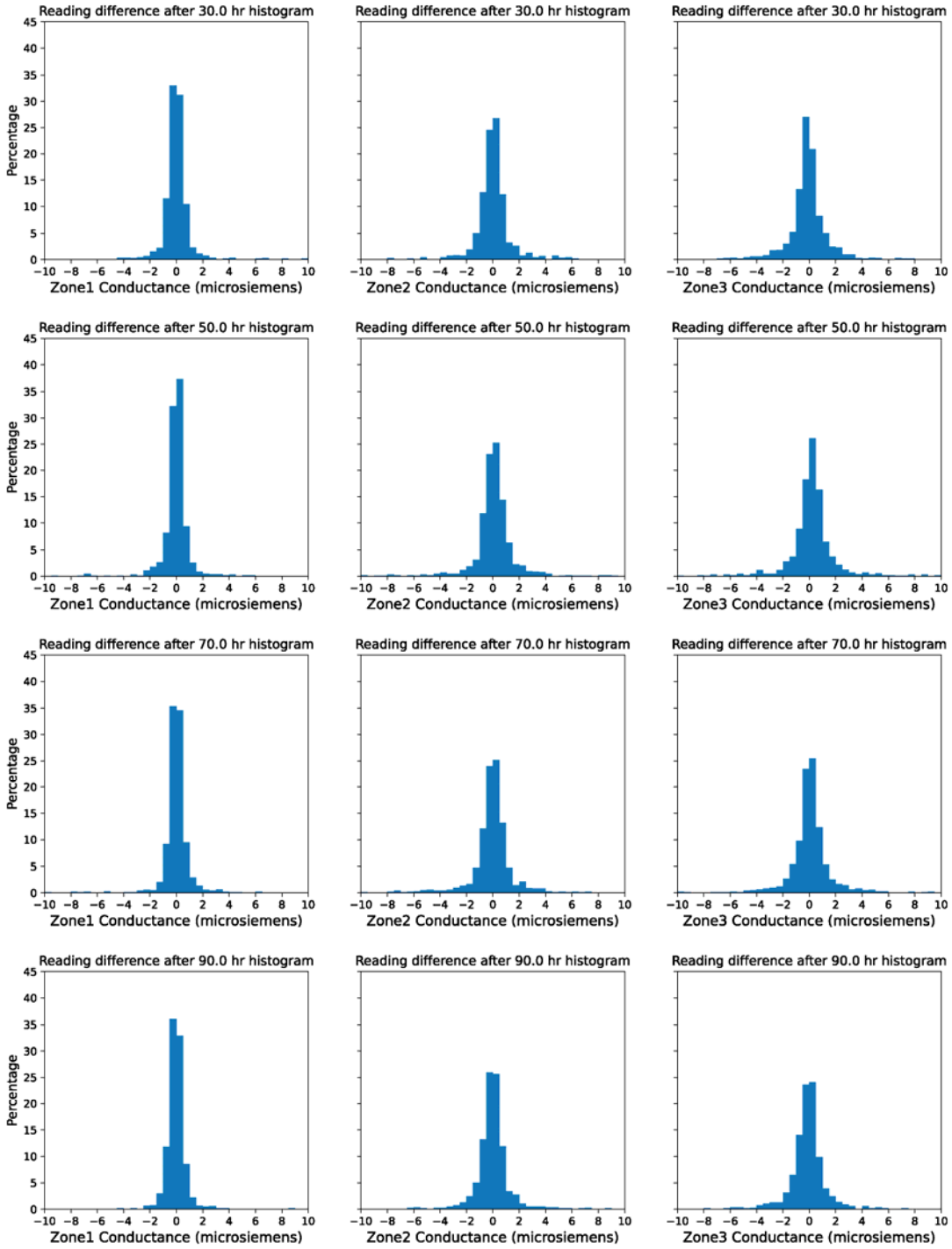

Fig. R2-4 The conductance drift in three programmed regions measured at 85 ℃ at 4 time points after the initial programming. Devices in zone 1 were programmed to ~40 μS,

devices in zone 2 were programmed to ~200 μS. Devices in zone 3 were programmed to ~550 μS.

We also measured a single device retention at even high temperatures (250 ℃, 300 ℃, 350 ℃) as shown in Fig. R2-5, where the extrapolation to lower temperatures shows a long retention. However, we should note that binary failure mode (large condutance jumps) was used in the measurements for this plot, which is not a direct measurement of multilevel retention (as Figs. R2-3 and R2-4) but a reference for long-term retention prediction.

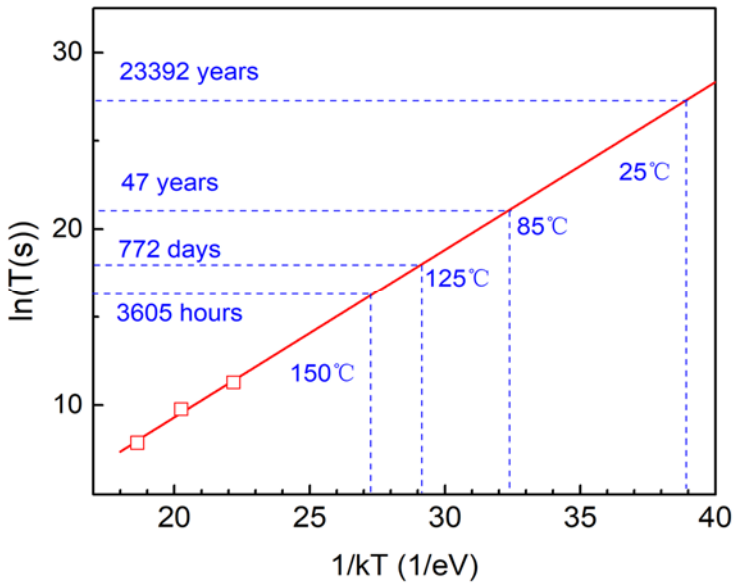

Fig. R2-5 Retention test at high temperature with Arrhenius fitting. The device retention was measured by keep reading the device at its low conductance state until the conductance suddenly jumped at 250 ℃, 300 ℃, 350 ℃ in a vacuum chamber. At each temperature, the measurement was performed 5 times and the data for plotting and fitting is the median value of the 5 measurements.

3.Comparing with write-verify, denoise method suppresses fluctuation caused by RTN and makes it possible for single memristor to obtain more conductance levels. Unlike precise calculations in scientific computing, neural networks can tolerate small fluctuations. From what is shown in Fig. 1e, assuming a read voltage of 0.2 V, the conductance perturbation is about 2% of the target value (about 148uS), which may be tolerable for the inference accelerator of the neural network. In edge computing application (many are one-step write), speed and energy are also very important metrics. The denoise method has a slight advantage in conductance precision, but is it still better than the write-verify method after comprehensive speed, energy and other factors? Seems to need further clarification. This comparison is especially important given the most

recent advances in selective write-verify, see Yan et al, "SWIM: Selective Write-Verify for Computing-in-Memory Neural Accelerators,"DAC'22.

We appreciate the insightful comments of the reviewer. A perturbation of 2% of the range means that device cannot have more than 50 distinguishable states, or less than 6-bit precision, which is insufficient in most neural network applications.

We appreciate many previous work on improving the programming speed and accuracy for neural network applications, like "SWIM: Selective Write-Verify for Computing-in-Memory Neural Accelerators,"DAC'22" as the reviewer mentioned. However, the RTN brings a new challenge that was not fully studied. *As cycle-to-cycle variations, the RTN induced flucutations could be more detrimental as it cannot be compensated by adaptive in-situ learning algorithms and breaks the stability and repeatability of inference computing.*

It is true that adding the denoising process will consume extra energy and elongate the programming process, but not by a lot though. To verify that, we calculated the energy and time consumptions. Based on the algorithm of high precision programming in Fig. S4, the total process can be divided into 3 phases: tuning (gradually switch the device to a target value), fluctuation scan (scan to verify whether the fluctuation of the device is acceptable or not), and stabilization (denoising operation using subthreshold voltage pulses).

The energy consumption is calculated as follows:

// INPUT: the target matrix W_{target} , the error history matrix W_{err} and the set and reset voltage history matrices V_{set} and V_{reset} of the whole array during 200 tuning cycles. All scalars are known parameters.

For $i \leftarrow 1 \ldots 200$ do

Calculate conductance matrix G_i from W_{target} and $W_{err,i}$

Calculate mean $P_{read,i}, P_{set,i}\ and\ P_{reset,i}$ from known $V_{read}, V_{set,i}$ and $V_{reset,i}$ and then averaging over all devices.

End for

$E_{tune} \leftarrow sum\left(P_{read,i} + P_{set,i} + P_{reset,i}\right) * t_{pulse}$

$E_{scan} \leftarrow P_{read,200} * n_{scan} * t_{pulse}$

$P_{stab} \leftarrow mean(V_{stab}^2 * G_{200})$

$E_{stab} \leftarrow P_{stab} * n_{stab} * t_{pulse}$

Energy of different tuning cycles are estimated from the mean tuning energy $\frac{E_{tune}}{200} * n_{tune}$.

The time estimation is counting the number of pulses of each phase as the pulse width is fixed (10 us here).

The energy and time consumption of the high precision programming algorithm is summarized in Fig. R2-6.

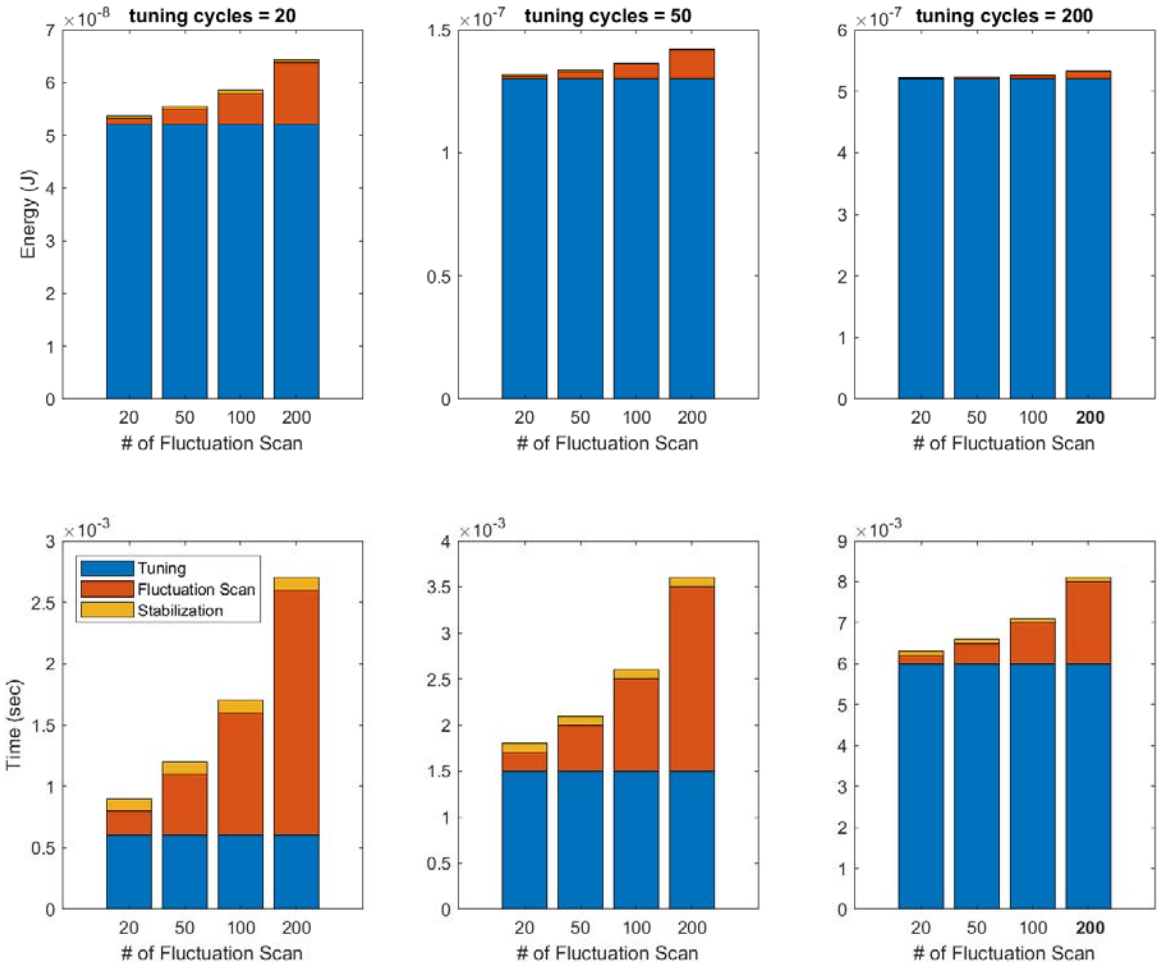

Fig. R2-6. The energy and time consumption per cell of different phases of the high precision programming algorithm. Different tuning cycles from 20, 50 to 200 and different fluctuation scan times from 20, 50, 100 to 200 are calculated based on the experimental data with 200 tuning cycles and 200 fluctuation scans.

In the experiments reported here, we used 200 tuning cycles and 200 fluctuation scans, corresponding to the right most bar of the right most sub-plot. In more optimized algorithms, we could shorten the cycle numbers needed by combining coarse and fine tuning processes. It can be seen that the *extra energy added by the denoising processing*

(fluctuation scan and stabilization) is actually a relatively small fraction of the energy needed to switch the device to the target value anyway. The time needed for each cell is a few milliseconds, which make the total time needed for a large array (e.g., 1000x1000) a few seconds in semi-parallel operation schemes (row by row).

In addition, there are other reasons that *make the denoising process not just desirable but even necessary for reliable computing*:

1. The device and protocols are mainly for edge computing. The advantage of analog memristive crossbar array mainly lies in fast and energy efficient inference. It is rare to frequently retrain the arrays distributed on the edge. The infrequent training makes the extra time and energy affordable and worthwhile for a high inference performance.
2. The existence of incomplete channels is detrimental for the array. They result in reading accuracy degradation; furthermore, because they are sensitive to sub-threshold voltage, they are also at higher risks in unpredictable re-shaping / re-composition by inference operations. The denoising process, not only reduces the reading error, but also removed those unstable incomplete channels which reduces long-term risks for applications. So we believe it is an essential process to suppress those RTNs for long term reliabilities*, which cannot be achieved by other approaches, including the Write-Verify approach*.

The above analysis has now been included in supplementary information as SI-9. We also included the important reference “SWIM: Selective Write-Verify for Computing-in-Memory Neural Accelerators,"DAC'22” in the main text:

(Paragraph 1 on page 2): “…Different approaches have been developed to accurately program the devices. [30] However, the highest conductance number reported to date has been no more than two hundred…”

SI-9 has now been mentioned in the main text:

(Paragraph 2 on page 3) “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.”

4 Fig. 2b and c, the residence time of RTN in c is obviously longer than that in b, may need explanation.

We thank the reviewer for the question. Indeed, the residence times in Fig. 2b and Fig. 2c show an obvious difference, for which one may hypothesize that the residence time depends on whether the RTN comes from an island-like channel or boundary-like

incomplete channels. Moreover, the residence time actually has a large distribution range, and each specific sample has different residence time without a clear trend.

This is not surprising as the residence time depends on the position of the charge-trapping oxygen interstitial defect, which is different in different samples. This is shown in ref. [*IEEE Transactions on Electron Devices*, vol. 62, no. 8, pp. 2606-2613, Aug. 2015]:

$$\tau_c = c_0 N_C e^{\frac{x_T}{\lambda_c}} e^{\frac{E_c(V_{READ})}{kT}}$$

$$\tau_e = c_0 N_V e^{\frac{t_B - x_T}{\lambda_c}} e^{\frac{E_e}{kT}}$$

We notice that the capture and emission time τ_c and τ_e are related with the trap distance from the injecting electrode, x_T. As the oxygen interstitial charge trap can be anywhere in the device, x_T has a range of variation of $\sim 6\ nm$, which is generally random among different samples. The Coulomb screening length, λ_c, is in the lengthscale of nanometer assuming carrier density of n of 10^{18}~10^{19} cm^{-3} of the conductive channel. Therefore, the position of oxygen interstitial defect can lead to a large range of the factor e^{x_T/λ_c}. If $n = 10^{19}$ cm^{-3}, this estimation gives the residence time with a variance of a factor of 52.9, which can span two orders of magnitudes. That explains why different samples exhibit evidently different residence time with randomness.

5. Correction: Page 7, "The residence time at the higher … the solid curves in Fig. 2b." should be Fig. 3b.

Thank reviewer for spotting this typo. It has been fixed now.

Referee #3 (Remarks to the Author):

This work by Rao et al. reports a record high number of conductance levels in memristive devices based on understanding of the conductance fluctuation mechanism and introduction of a critical denoising operation during device programming. Improvement of the weight precision in memristive devices is key to the performance and efficiency of memristor based in-memory computing systems, and hence the results in this work represent a significant step forward, especially for neural network models and applications that demand high-precision weight parameters. Theoretical studies were performed to understand the root cause of conductance fluctuations that previously limit weight precision, and subsequent conductive AFM characterization was employed to understand the role of the proposed denoising process. The manuscript is well structured and written, and the proposed protocol might be important for the research field.

We thank the reviewer for the positive comments.

This reviewer, however, has some technical questions on the work, as detailed below.

1. The authors have utilized conductive AFM to characterize the evolution of conduction channels before and after the denoising process in order to clarify the denoising mechanism. This will need to land the AFM tip onto the same spot of the sample surface for performing the denoising operation, accompanied by scanning spatial conductivity maps in the two states. My question is how the location of the tip was calibrated so as to ensure the same position? Will possible shift in tip position affect the mapping results and observed evolution of channels?

We thank the review for the good question. As we addressed on the method part and the response to the Reviewer 1. The contact radius for the scanning the shape of the conduction channels and the in situ denoising of the conduction channels are controlled to be different by tuning the set point, which is corresponding to the applied force of the AFM tip. In detail, When the set point was set to 10 nm, which was used for the conduction channel scanning, the contact radius is about 1.2 nm, which is small enough to resolve the details of the channel. When the set point was set to 80 nm, which was used for the in situ denoising, the contact radius is about 64.4 nm, which is large enough to cover the entire channel area (with a diameter around 20 nm here). The relationship between the contact radius and the set point is shown in followed figure. Thus, during the denoising operation, the estimated contact diameter is about 130 nm (64.4 of the radius) which is much larger than the diameter of the channel (filament), which can increase the tolerate the positioning error of the XY scanning in AFM.

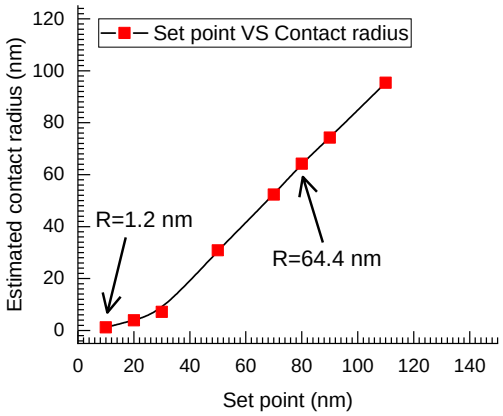

Fig. R3-1. The relationship between the estimated contact radius and set point.

Moreover, during the measurement, we use an adaptive XY control of the equipment, which can give a great accuracy in position tracking by minimizing the errors introduced by creep and thermal drift, which is much better than the conventional closed-loop XY control. We compare the XY drift with normal XY closed-loop control and that with the

adaptive XY closed-loop control by continuously scanning the same area, as shown in followed figure. On the standard calibration sample from vender, the adaptive XY closed-loop control gives a drift about 5 nm, which is much smaller than that with the normal XY closed loop control (25 nm), and the drift (5 nm) is much smaller than the redundancy (64.4-10=54.4 nm) between the contact radius (64.4 nm) and filament diameter (~20 nm). Thus, by enlarging the tip contact radius during the denoising operation and minimizing the XY drift by using the adaptive XY control, we can ensure the entire filament spot was covered by the tip and the possible small drift in tip position would not affect the mapping results, which allows us to observe the evolution of channels.

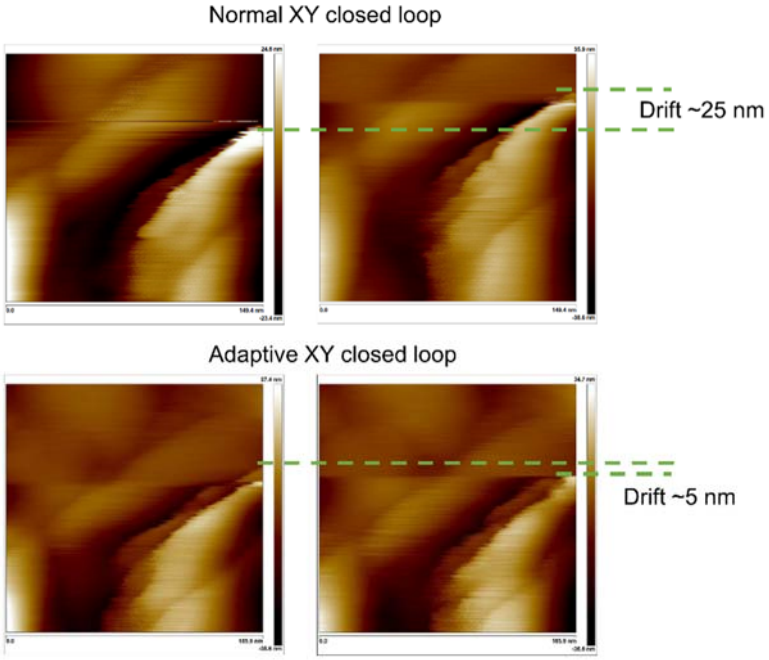

Fig. R3-2. The XY drift from the modes with normal XY closed loop control (top) and adaptive XY closed loop control (bottom) on the standard calibration sample from vender.

The additional probe information has been included in supplementary information as Fig. S11 and mentioned in the main text:

(Paragraph 2 on page 5) "…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…"

2. Figure 2 shows the incomplete channels are either annihilated or completed after denoising process, whereas the base current levels didn't change after denoising.

Although the main conductance of the device should be dominated by the main completed channels, certain reduction in the base current level may be expected after annihilation of the incomplete channels. More importantly, completing the partial filament to form fully connected filament should increase the conductance, because it becomes part of the main channel and the overall filament grows larger. This seems different from the results in Figure 2b-c, and further clarification on this point is needed.

We thank the reviewer for this good question. The results in Fig. 2 indicate that the fluctuating amplitude in two cases is equal to the conductance contribution of the incomplete filaments, that can be completely blocked by the trapped charges nearby.

In the case of Fig. 2b, d, e, the extra current on top of the base current comes from the conductance of the incomplete filament (the island) when it is not blocked by the trapped charge nearby. When trap site(s) is charged, such extra conduction is *fully blocked*, and the total current goes to the base current level. By the denoising operation, such incomplete filament(s) is annihilated, which removes the extra conduction by the filament permanently and sets the current level to based current. We'd like to point out that in this device, if we used sub-threshold SET (instead of RESET) voltage to denoise, the extra conductance of the incomplete filament would be added to the 'base' conductance and the final current reading would be close to the top level of the red curve in Fig. 2b.

It has also been reported in Ref. [40] that filaments with certain doping concentration and size can be completely blocked, which is consistent with our observations.

In the case of Fig. c, f, g, during the denoising we used a sub-threshold SET voltage to make such incomplete filament region grow into a complete one by introducing more dopants in that region. As shown in Fig. R3-3, a certain level of dopant addition (such as those involved in the denoising process) does not noticeably change the filament condutance per se, but dramatically changes the sensitivity of the filament to charge blocking effect (by charges trapped in trap sites nearby, e.g., O interstitals). *Therefore, the main difference of the 'incomplete' and 'complete' filaments are not their condtuance but their sensitivity to the charge blocking effect.* Before the denoising process, the blocking effect to the incomplete region along the boundary is responsible for the current dips. After the denosing process, that region is not blocked anymore and the total current returns to the base current. Because the denosing process doesn't significantly change the condutance of the impcomlete region, the based current does not change before and after denosing process in a notciable way. Those are supported by our simulation results in Fig. R3-3.

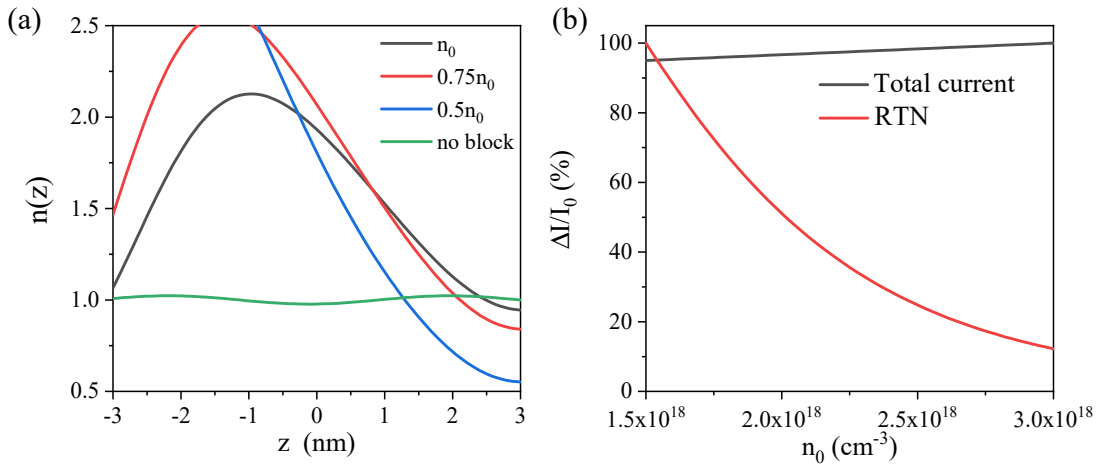

Fig. R3-3 Simulation of (a) transport wave function in the boundary-type incomplete channel and (b) variation percentage of the total current and RTN when changing the carrier density. Model parameters are set as follow: the carrier density n_0 in (a) is 3×10^{18} cm^{-3}, the charged defect distance is d = 1 nm, and the boundary width is set as a = 5 nm, transport electron energy E is 0.2 eV above the CBM, the temperature T = 300 K, and we assume boundary contributes 10% of total current in (b).

Assuming the carrier density drops 40% (from 2.5×10^{18} to 1.5×10^{18}), the blocking/ unblocking conductance is 4 times different (the red curve), while the total current (condutance) only has a 4% variation.

We'd also like to mention that there are cases denoising process may lead to change of base resistance of the overall filament region, which is the reason devices may sometime need to re-enter fine-tuning loop after denoising trials as shown in the tuning algorithm in Fig. S4. This is because the denoising voltage has a small chance to make changes to both incomplete channels and complete channels. However, by carefully controlling the signal parameters, the change of base conductance is usually very small, as shown in Fig. R3-4. In this case, only stabilization pulse is applied to the device and the mean current value was read out after each stabilization.

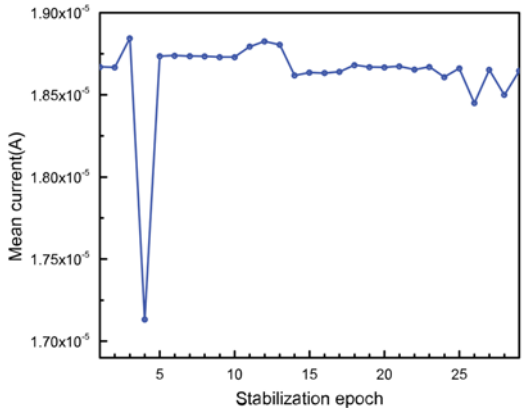

Fig. R3-4 Memristor mean readout current history during stabilization.

3. Is it of general applicability to attribute conductance fluctuation in memristors solely to RTN? This work shows that set operation tends to induce larger fluctuation than reset operation. This is different from many previous reports, which reveals stronger fluctuations in high resistance states, and therefore may suggest existence of different physical pictures. A few examples can be found in Refs. 30, 31, etc. The configuration of the conduction channel could be changed for the chosen conductance state, either in different cycles or different devices, and that is related to 'atomic effect' instead of purely 'electronic effect'. Further clarification on this point is needed.

That's a very good question. RTN may not be the solely source of conductance fluctuation in memrsitors, but is the largest one that prevents us from obtaining many distinguishable levels. As shown in Fig. S3, some fluctuations seemingly not RTNs are actually proven to be RTNs when sufficiently fast sampling rate is used to collect the data. We'd also like to clarify that in our manuscript, 'SET Operation' is used to just refer to the operation voltage polarity that tends to switch the device towards a slightly more conductive state and it does not mean that the device become very conductive (in the so-called SET state) after every SET Operation. Similarly, 'RESET Operation' is used to refer to the operation voltage polarity that tends to switch the device towards a slightly more resistive state. Our observation is that when a device freshly reaches a certain conductance state, its fluctuation level may often depends on how it gets to this conductance state, i.e. by a SET operation or a RESET operation, and the former (SET operation) often leads to a higher fluctuation level than the latter (RESET operation) even for the same conductance level. We agree with the reviewer on that this is likely due to the different atomic configuration of the conduction channel for a nominally same/similar conductance level. The denoising process is really to change atomic (and oxygen vacancy) configuration and concentration so that it becomes less sensitive to the impact of charge trapping nearby. The step-like conductance fluctuation is mainly from the charge trapping RTNs. In the experiment of Fig. S2, where weak SET / RESET operations were done so that the device conductance increase / decrease in a small range (100 ± 10 μS). In this case no iterative tuning was performed. Statistics of as-

SET/RESET state (without denoising) fluctuations were measured and their statistics were plotted in Fig. S2. The RESET operations lead to conductance states with smaller fluctuation levels than those of SET operations. Because the as-SET and as-RESET states are in a similar conductance range, it does not contradict with result from Ref. 30. The result of Ref. 32 (that was Ref. 30 in the original submission) shows the relationship between noise level and conductance level (in Ref. 32, the device has stronger fluctuation at low conductance level) while in our experiment in Fig. S2, we were trying to study the impact of programming method (instead of conductance) on fluctuation.

We also considered the scenario of atomic-induced fluctuation in our case.

First of all, we think the possibilities of atomic-induced fluctuation further suggest the necessities of performing a fluctuation reading. With the help of feedback loops as shown in Fig. S4, such atomic-induced fluctuation can be resolved together with RTN.

Regarding whether RTN itself in the device in this work is atomic effect or electronic effect. We believe the latter is more likely the case, which is also supported by the following observation.

As shown in Fig. 3b, τ_1 is negatively related to the reading voltage while τ_0 is almost independent of the reading voltage.

This is well explained if we assume the RTN is of electronic origin. But if we assume the RTN has an atomic origin, the lifetime of each state should be determined by the transition state theory:

$$\frac{1}{\tau_1} = \nu_1' \exp\left(-\frac{E_s - E_1 + qV_{s1}}{k_B T}\right), \qquad \frac{1}{\tau_0} = \nu_0' \exp(-\frac{E_s - E_0 + qV_{s0}}{k_B T})$$

where E_1, E_2, E_s are the 1 state, 0 state, and saddle point energy; ν_1' and ν_0' are the effective oscillation frequency of 1 and 0 states; V_{s1} and V_{s0} are the electric voltage between the saddle point and 1 and 0 states, respectively. Therefore, we expect that when the voltage increases, both τ_1 and τ_0 should change because of the change of V_{s0}, V_{s1} (electric effect), and T (temperature effect of electric current). It is hard to explain why τ_0 is almost constant and exhibits a sharp contrast with τ_1. Therefore, we think the data indicates that the RTN in our experiment is unlikely to be an atomic effect.

Related changes have been made in Supplementary information SI-16 and mentioned in main text:

(Paragraph 1 on page 10) "…independent of the applied voltage, which is inconsistent with atomic motion induced RTNs (see SI-16)."

Clarifications have been made in the caption of Fig. S2 to make it clearer:

"in which 100 programming operations ended with a SET operation and the other 100 programming operations ended with a RESET operation. Small voltages were applied so that the device is only slightly SET/RESET within a small conductance range within ±10 μS of 100 μS after each programming operation,"

4. How to define the number of states is essential for this work and needs to be justified more clearly. This work used a fixed value of 2 uS as the interval between every two neighboring states, and therefore achieved 2048 states between 50~4144 uS (about 4094 us range). It is also mentioned that the amplitude of variation is proportional to the actual current level, namely higher conductance state has larger variations. Is the 2 uS interval sufficient for those states with high conductance? Fig. 1f and the inset of Fig. 1g present zoomed-in images for rather low conductance states, but showing it for high-conductance states as well will be more convincing.

Thanks for pointing this issue out.

First of all, we proved that 2 μS is sufficient for the highest range according to our criteria, as shown in Fig. R3-5.

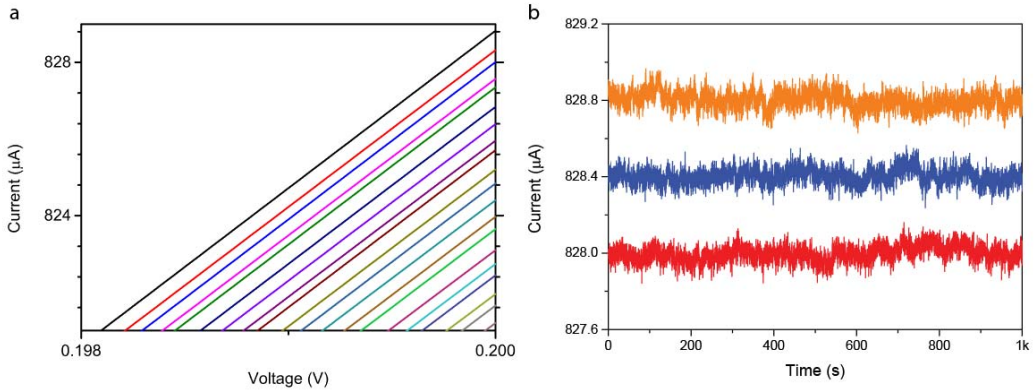

Fig. R3-5 Programming result of memristor at high conductance state. a) Zoom-in image of Fig. 1g near 4144 μS (~828uA). b) Zoomed-in view of three nearest neighboring states near 4144 μS after denoising. The current of each state was read by a constant 0.2V voltage. No large-amplitude RTN was observed, and all the states can be clearly distinguished.

The reason that in Fig. 3, the fluctuation is proportional to the current level, is because the device state is not changed during the three voltage levels (0.15V, 0.2V, 0.25V) and the fluctuation amplitude is determined by the same incomplete filament at those voltages.

When we program the device into a different conductance state, we may remove existing incomplete channels and/or form new incomplete channels which may not have a conductance that is proportional to the total conductance. Meanwhile, if the incomplete

filament is very conductive (due to large size or high oxygen vacancy density), the nearby trapped charge may not be able to completely block the filament, as inferred from results in Fig. 3. As a result, at very high conductance range, we may not see RTN amplitude increase linearly with the overall conductance to very large; in fact, we frequently observe RTN with similar amplitude at high conductance region:

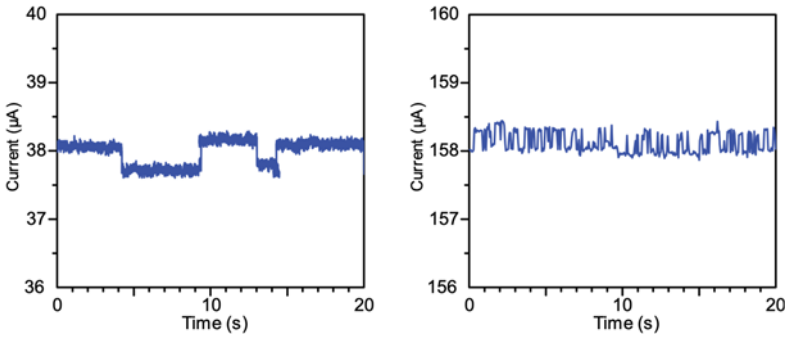

Fig. R3-6 RTN observed at different conductance levels.

Fig. R3-5 is now included in supplementary information Fig. S5 and referred to in the main text:

(Paragraph 2 on page 3): "…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…"

5. The tuning process for the high precision programming as depicted in Fig. S3 is quite complex and requires a large number of iterative operations. This is OK for small arrays but will easily lead to extremely long programming time when loading weights into memristor neural networks for chips with large memristor capacity. I would recommend the authors to characterize the programming time for each cell and array, and ideally project it to NNs with practical capacities. Moreover, although the denoising process doesn't involve additional circuitry, the closed-loop programming still leads to large overhead in periphery circuitry, which requires careful evaluations.

We appreciate this important question of the reviewer. Counterintuitively, adding the stabilization phase to a conventional write-and-verify tuning loop may help to reach the desired precision faster and save some programming time, especially on a larger array. It might even be a necessary step to reduce the cycle-to-cycle variation for inference operations and reach a higher accuracy for long term.

First, a close loop algorithm is necessary to reach a certain precision, for example, 8-bit, which is applicable to most NN applications. This is because many non-idealities of real

devices can reduce the effectiveness of any delicate algorithm without feedback loop and gradual approximation, for example, device-to-device variation. And that is why our fitting modeling in Fig. S4 (Fig. S3 in the original submission) can only be used for coarse tuning. With that being said, either close loop tuning or stabilization process can be controlled at the software level without extra circuitry. As shown in Figure 1e, an as-programmed device may have a noise amplitude of about 2%, which means that device may not reach even 6-bit (64 states). Thousands level of states are only obtainable with the denoising process.

Second, we have calculated and estimated the additional energy and time consumption per cell under several possible combination of tuning parameters.

Based on the algorithm of high precision programming in Fig. S4, the total process can be divided into 3 phases: tuning (gradually switch the device to target values), fluctuation scan (scan to verify whether the fluctuation of the device is acceptable or not), and stabilization (denoising operation using subthreshold voltage pulses).

The energy consumption is calculated as follows:

// INPUT: the target matrix W_{target} , the error history matrix W_{err} and the set and reset voltage history matrices V_{set} and V_{reset} of the whole array during 200 tuning cycles. All scalars are known parameters.

For $i \leftarrow 1 \ldots 200$ do

Calculate conductance matrix G_i from W_{target} and $W_{err,i}$

Calculate mean $P_{read,i}, P_{set,i}\ and\ P_{reset,i}$ from known $V_{read}, V_{set,i}$ and $V_{reset,i}$ and then averaging over all devices.

End for

$E_{tune} \leftarrow sum\left(P_{read,i} + P_{set,i} + P_{reset,i}\right) * t_{pulse}$

$E_{scan} \leftarrow P_{read,200} * n_{scan} * t_{pulse}$

$P_{stab} \leftarrow mean(V_{stab}^2 * G_{200})$

$E_{stab} \leftarrow P_{stab} * n_{stab} * t_{pulse}$

Energy of different tuning cycles are estimated from the mean tuning energy $\frac{E_{tune}}{200} * n_{tune}$.

The time estimation is counting the number of pulses of each phase as the pulse width is fixed (10 us here).

The energy and time consumption of the high precision programming algorithm is summarized in Fig. R3-7.

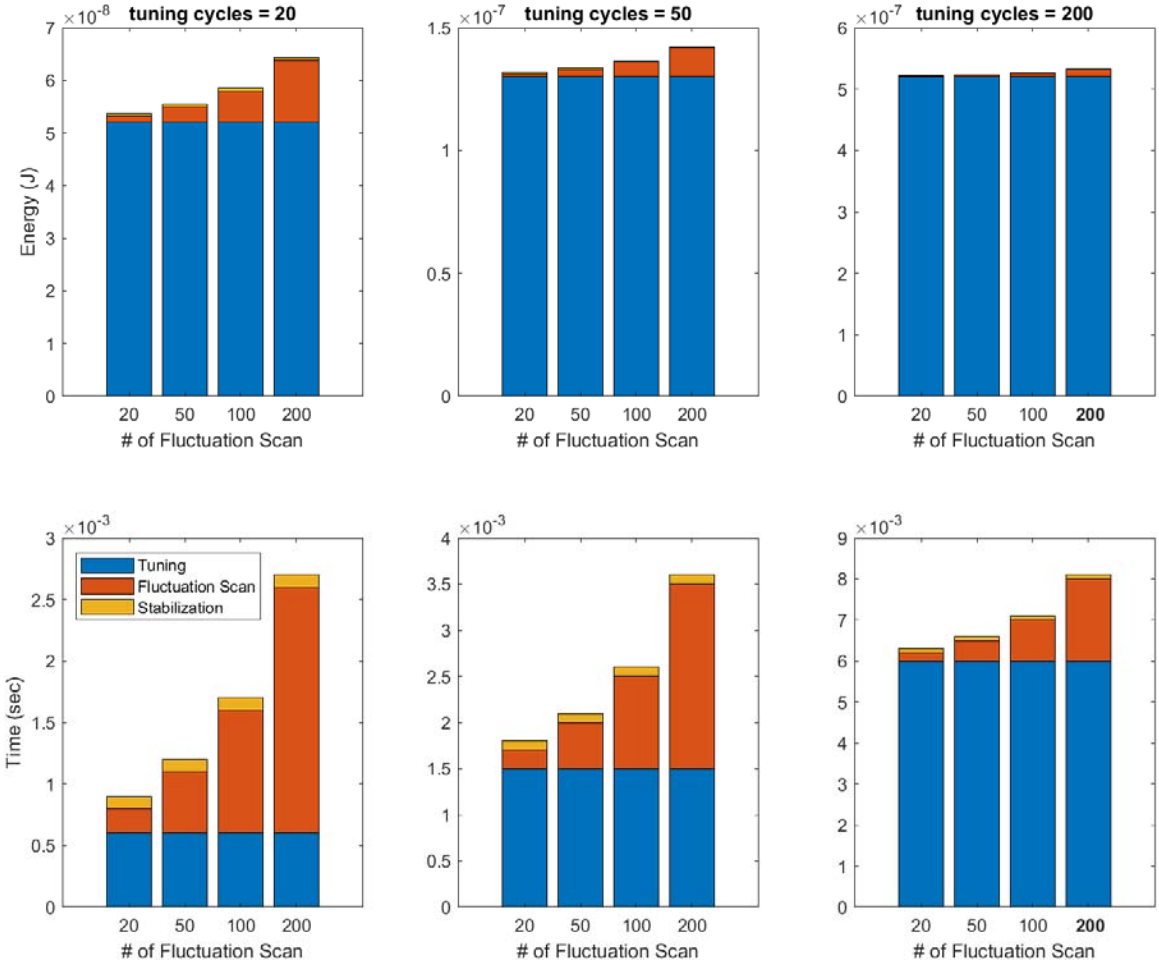

Fig. R3-7. The energy and time consumption per cell of different phases of the high precision programming algorithm. Different tuning cycles from 20, 50 to 200 and different fluctuation scan times from 20, 50, 100 to 200 are calculated based on the experimental data with 200 tuning cycles and 200 fluctuation scans.

In the experiments reported in here, we used 200 tuning cycles and 200 fluctuation scans, corresponding to the right most bar of the right most sub-plot. Tuning cycles could be improved by the optimization of tuning algorithm and improving the chip uniformity. It can be seen that the extra energy added by the denoising processing (fluctuation scan and stabilization) is actually a relatively small fraction of the energy needed to switch the device to the target value anyway. The time needed for each cell is a few milliseconds, which make the total time needed for a large array (e.g., 1000x1000) a few seconds in semi-parallel operation schemes (row by row).

In addition, there are other reasons that make the denoising process necessary for reliable computing:

1. The device and protocols are mainly for edge computing. The advantage of analog memristive crossbar array mainly lies in fast and energy efficient inference. It is rare to frequently train the arrays distributed on the edge. The infrequent training makes the extra time and energy affordable and worthwhile for a high inference performance.
2. The existence of incomplete channels is detrimental for the array. They result in reading accuracy degradation; furthermore, because they are sensitive to sub-threshold voltage, they are also at higher risks in unpredictable re-shaping / re-composition by inference operations. *The denoising process, not only reduces the reading error, but also removed those unstable incomplete channels which reduces long-term risks for applications*. So we believe it is an essential process to suppress those RTNs for long term reliabilities, *which cannot be achieved by other approaches, including the Write-Verify approach*.

The analysis in this part has been included into the supplementary information as SI-9 and mentioned in the main text:

(Paragraph 2 on page 3) "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."

6. This work has experimentally and theoretically clarified the correlation between RTN and conductance fluctuation/programming precision in memristive devices, and in turn proposed a denoising process that allows high-precision programming. Although this is important for the application of memristive neural networks, the study and understandings on the RTN mechanism itself have been covered by previous works in Refs. 29, 30, 39, 40 & 41, etc. The authors are recommended to point out the new understandings revealed in this work more clearly.

We totally agree with the reviewer on that there have been fantastic previous studies on the understandings of the RTN mechanism in Refs. 29, 30, 39, 40 & 41, where the RTN mechanism of Coulomb blockade from oxygen interstitials was established. Based on such previous understandings, we think our work provides further experimental/theoretical support and insights to the RTN theory in the following three aspects:

1. We directly observed the **incomplete channel** and confirmed its relationship with RTN in our devices. Although the Coulomb blockade mechanism is the same, identifying the special role of incomplete channels reveals the details of the microscopic origin of the RTN. Most previous experiments support the filaments' Coulomb blockade mechanism by reading the electric signals, while our work has

the image of the incomplete channel and its correspondence with the appearance and disappearance of RTN. This point is emphasized in paragraph 2 on page 5.

2. We examine the RTN mechanism in a nano-scale channel where **quantum mechanics** govern the electron's transport. Most reported methods focus on the classical case, using the carrier drift-diffusion equation to describe the charge transport. In our case, the scale of conductive channels is down to the nanometer scale, and the quantum effect can play a role in carrier transport and filament blockade. For example, *the blockade of a conductive channel can be more difficult because of quantum tunneling*. Therefore, we performed quantum mechanics calculation on transport wavefunction, closer to our actual scenario in this ~ 5nm conductive channel. We include an emphasis on this point on page 10.
3. The **denoise mechanism** we proposed, namely, how the RTN can be reduced by denoise voltage, has not been studied to the best of our knowledge, as emphasized on page 11.

To clearly acknowledge the mechanism established in previous work, we include clarification on page 7 of the manuscript indicating:

Previously, an insightful theoretical framework on the electronic RTN mechanism is established in reference [*Applied Physics Letters,* vol. 96, no. 5, p. 053503, 2010; *IEEE Transactions on Electron Devices,* vol. 61, no. 8, pp. 2920-2927, 2014].

We also cite ["Random Telegraph Noise Nano-spectroscopy in High-κ Dielectrics Using Scanning Probe Microscopy Techniques" in *Noise in Nanoscale Semiconductor Devices*: Springer, 2020, pp. 417-440], where the role of oxygen interstitial defects in RRAM was discussed.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have addressed most of the comments and improved the manuscript. However, my main concern on the explanation of the mechanism is still not resolved. Involving oxygen interstitials in the model is not justified and I cannot accept it at that stage, based on the calculations provided. Nevertheless, I support a publication of the manuscript, provided the part on the mechanism is either removed, or fundamentally revised/reconsidered.

Comments:
1. Based on the knowledge on defect chemistry (point defects) of HfO2 it is very improbable interstitial oxygen ions to be involved. Oxygen interstitials in general are formed in two situations – a) as a result for anti-Frenkel (intrinsic) disorder and b) incorporated at high oxygen pressures where oxygen cannot be incorporated on an empty vacancy position and goes to interstitial position.

a. In the case of anti-Frenkel defects, the charge of the oxygen interstitials is compensated by an oxygen vacancy (not by an electron hole!). This type of disorder is expected for “normal” oxygen partial pressures.

b. In case of oxygen deficiency (which here seems to be the case), the main point defects are oxygen vacancies, with a charge compensated by electrons. Under these conditions, presence of oxygen interstitials is highly improbable (the concentration of vacancies and interstitials is related by the Frenkel equilibrium constant). To my knowledge, there is no single experimental evidence on oxide ion conductors with oxygen interstitial species at reducing conditions (low oxygen partial pressures or oxygen deficient lattices).

c. At high oxygen partial pressures (typically considered above 1 atm) the concentration of oxygen vacancies is negligible and incorporated oxygen is forced to go to interstitial position. The charge compensation occurs indeed by holes, but they are not necessarily localized (they are rather non-localized). In such a case the assumption should be made that HfO2 is HfO(2+x). However, oxygen over-stoichiometry is not typical for HfO2 but rather HfO(2-x) is always assumed.

So, I do not see any argument based on the defect chemical equilibria and point defect models that support the suggested mechanism as reasonable.

2. The authors speculate that oxygen interstitials attract holes and form neutral interstitial oxygen, which in turn attracts an electron (electron trap). This assumption violates the electroneutrality conditions. One oxygen ion at interstitial position is 2-, two electron holes are 2+. Adding additional electron makes 3- charges vs 2+ charges (violated electroneutrality). Why a neutral specie (assuming theoretically neutral oxygen interstitial exists under these conditions) should be an electron trap? Where the additional electron is coming from, and what should be its compensating charge? Why oxygen vacancies (positively charged) should not be preferred as traps?

3. I have checked the provided literature (Random Telegraph Noise Nano-spectroscopy in High-κ Dielectrics Using Scanning Probe Microscopy Techniques" in Noise in Nanoscale Semiconductor Devices: Springer,2020, pp. 417-440) and found no support for the claims. The authors should verify the citation. In the cited chapter as traps oxygen vacancies are considered!

4. I have found in the reference list of the book (ref 88-91 in the book) theoretical calculations on oxygen interstitials including mentioning neutral interstitials. However, I would not say they support the mechanism suggested in this manuscript.

5. Additional comment: I strongly recommend using Kroeger-Vink notation (relative charges). In the classical notation (absolute charges) both lattice oxygen ions and interstitials are O2-. Mixing relative and absolute charge notations causes significant confusion for the readers.

EXTRA COMMENTS FROM REVIEWER 1 (in the light of your response of 19 December)

I have carefully read the responses and argumentation provided by the authors as well the suggested changes in the manuscript. In my view the manuscript can be accepted implementing the proposed modifications. As also pointed by the authors the suggested mechanism is not vital for the perception of the main results and achievements in the paper. With the changes the authors have made clear that the proposed mechanism is a hypothesis, supported by theoretical calculations and will move the part to the SI. The system is exposed to strongly non-equilibrium conditions and such a constellation cannot be excluded. From my side this is absolutely sufficient.

So I join the other reviewers, recommending acceptance of the manuscript.

Referee #2 (Remarks to the Author):

My previous concerns have been adequately addressed by the authors.

Referee #3 (Remarks to the Author):

The authors made detailed responses to the questions, which addressed my concerns. I have no further comments on the revised manuscript and can recommend publication.

Author Rebuttals to First Revision:

Referee #1 (Remarks to the Author):

The authors have addressed most of the comments and improved the manuscript. However, my main concern on the explanation of the mechanism is still not resolved. Involving oxygen interstitials in the model is not justified and I cannot accept it at that stage, based on the calculations provided. Nevertheless, I support a publication of the manuscript, provided the part on the mechanism is either removed, or fundamentally revised/reconsidered.

Response: We appreciate the reviewer's constructive comments and suggestions, which have been very helpful for us to improve our manuscript in the first round as well as the 2nd round. Meanwhile, we also want to respectively point out three points:

1) The major conclusions in the mechanism study are still valid regardless of the type of the trapping/detrapping center. These conclusions include: 1. In the filament blocking mechanism simulation (in Fig. 3c to f), the trapping center is considered as a point charge, meaning any charge trapping center (that is not too big to significantly change the local geometry near the simulated filament) that can switch between its neutral and trapped (capturing an electron) states should lead to the same filament-blocking effect. 2. Our denoising mechanism is independent of the source of charge trapping sites, because during the denoising process the trapping sites remain unchanged. Therefore, whether the trapping sites are oxygen interstitials or something else barely make any difference to our main mechanism part (i.e., denoising mechanism) described in the manuscript.
2) We agree with the reviewer that we do not have conclusive evidence to state that the charge trapping sites are certainly oxygen interstitials. We have tuned the language in the revision to reflect this point.
3) We also fully agree with the reviewer regarding the defect chemistry mentioned in reviewer's comments. However, we would also like to argue that oxygen interstitials in HfO_2, although rare (as the reviewer pointed out correctly), are likely in our system under the unusual condition created by the electroforming condition, as explained in the newly added Fig. R1 below. On the other hand, the long characteristic time of RTN as measured well matched the large relaxation energy of an oxygen interstitial, serving as indirect evidence that oxygen interstitial may act as a trapping/detrapping center. (Other than oxygen interstitial, we did not find any other possible defect in our system with such a long characteristic time.)

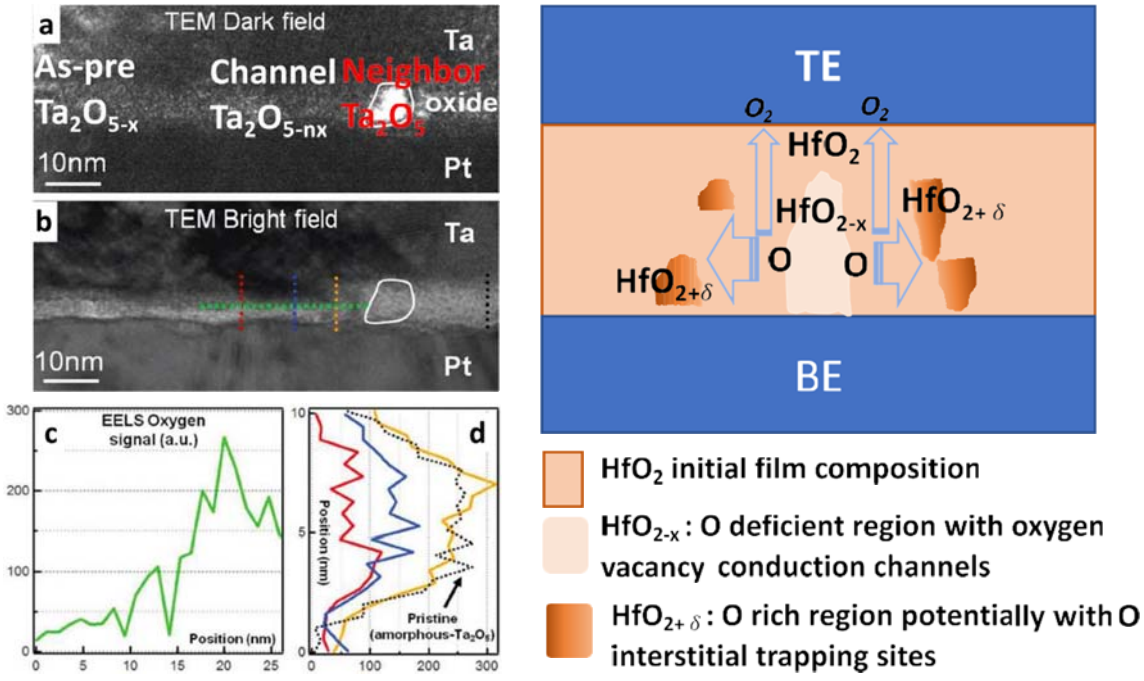

Fig. R1 The formation of O-rich region accompanying O-deficient region during electroforming process. Left figure: our previous experimental observation that for as-prepared Ta_2O_{5-x} film, after electroforming, O concentration decreased in the electroformed conduction channel region (Ta_2O_{5-nx} , n>1) but increased in the neighboring region, where a single crystal Ta_2O_5 is unambiguously identified (circled white region in panel a) with a higher O concentration than the as-prepared film). These results clearly show that O atoms migrated towards the neighborhood of conduction channel. At certain point (~7 nm) the concentration of oxygen even slightly exceeds stoichiometric Ta_2O_5, leading to the existence of $Ta_2O_{5+\delta}$ at that point. (Result adapted from *Advanced materials* 23.47 (2011): 5633-5640.) **Right figure**: Schematic of a possible scenario in electroformed HfO_2 system. Under a high electrical bias during electroforming, some stoichiometric HfO_2 region is electroformed, which creates oxygen vacancies to form electrical conduction channels. Some of the oxygen atoms removed from this region escaped the film as oxygen gas with a high pressure and others migrate to the neighboring region, establishing a O rich region that makes O interstitial defects possible, which can serve as charge trapping sites to make their neighboring conduction channels fluctuate in conduction.

The above figure has been included in the Supplementary Information now.

It was also reported in [*Physical review letters* 89.22 (2002): 225901.] that oxygen interstitials can diffuse in HfO_2 without large energy barrier and oxygen can serve as traps and create intrinsic electric fields.

The corresponding changes as detailed in the end of this letter are made in the main text to reflect the above points.

In the following, we would like to respond to the reviewer's specific comments point-by-point.

Comments:

1. Based on the knowledge on defect chemistry (point defects) of HfO2 it is very improbable interstitial oxygen ions to be involved. Oxygen interstitials in general are formed in two situations – a) as a result for anti-Frenkel (intrinsic) disorder and b) incorporated at high oxygen pressures where oxygen cannot be incorporated on an empty vacancy position and goes to interstitial position.

We agree with the reviewer on the defect chemistry and believe in the case of electroforming in our device, case b) mentioned above applies, as explained in Fig. R1.

a. In the case of anti-Frenkel defects, the charge of the oxygen interstitials is compensated by an oxygen vacancy (not by an electron hole!). This type of disorder is expected for "normal" oxygen partial pressures.

b. In case of oxygen deficiency (which here seems to be the case), the main point defects are oxygen vacancies, with a charge compensated by electrons. Under these conditions, presence of oxygen interstitials is highly improbable (the concentration of vacancies and interstitials is related by the Frenkel equilibrium constant). To my knowledge, there is no single experimental evidence on oxide ion conductors with oxygen interstitial species at reducing conditions (low oxygen partial pressures or oxygen deficient lattices).

We agree with all the points here except that the oxygen deficient region is only in the conduction channel region and the neighboring region surrounding the channel region is not reducing conductions (oxygen rich instead).

c. At high oxygen partial pressures (typically considered above 1 atm) the concentration of oxygen vacancies is negligible and incorporated oxygen is forced to go to interstitial position. The charge compensation occurs indeed by holes, but they are not necessarily localized (they are rather non-localized). In such a case the assumption should be made that HfO2 is HfO(2+x). However, oxygen over-stoichiometry is not typical for HfO2 but rather HfO(2-x) is always assumed.

We agree with all the points here and again would like to emphasize that we do believe high oxygen partial pressures in the neighborhood of the conduction channel under the very drastic electroforming process is possible as it can happen within ns and NOT ALL the generated O atoms have time to form oxygen gases and escape from the film. The fact that damages induced to the devices by gas-bubble eruptions have been commonly observed during electroforming also indicates an oxygen pressure higher than 1 atom has formed locally inside the film during electroforming process, which made O surplus possible in HfO_2 under the electroforming condition.

So, I do not see any argument based on the defect chemical equilibria and point defect models that support the suggested mechanism as reasonable.

We do not have conclusive evidence to show there must be O interstitial defects, but they are likely in our system under electroforming conditions based on the above experimental data and analysis. Even if charge trapping sites are something else other than O interstitials, our denoising mechanism remains the same.

2. The authors speculate that oxygen interstitials attract holes and form neutral interstitial oxygen, which in turn attracts an electron (electron trap). This assumption violates the electroneutrality conditions. One oxygen ion at interstitial position is 2-, two electron holes are 2+. Adding additional electron makes 3- charges vs 2+ charges (violated electroneutrality). Why a neutral specie (assuming theoretically neutral oxygen interstitial exists under these conditions) should be an electron trap? Where the additional electron is coming from, and what should be its compensating charge? Why oxygen vacancies (positively charged) should not be preferred as traps?

It is possible and common for neutral defects to trap electrons and become charged defects, as bypass conduction electrons have a transition probability to fill the localized state (as in Fig. 3a) of a neutral defect, which is not necessarily driven by electrostatic attraction. For example: the widely studied NV center [*Nat Commun* **7**, 12660 (2016)] and SiV center [Phys. Rev. Lett. **120**, 117401] in diamond, their neutral states can trap an electron to form a negatively charged defect state, and the negatively charged state is stable and widely used in quantum technology [Annu. Rev. Phys. Chem. 2014. 65:83–105]. Charges are also trapped in neutral dielectrics (SiNx) in Flash memory devices. If charges can freely move in a spontaneous process, electroneutrality can be satisfied all the time. Otherwise, electroneutrality may be temporarily broken during the charge trapping/detrapping process (as our cases).

The electrons can come from the conduction band of the crystal, as the relevant charge trap defects are close to the conduction channel, conductive electrons can either diffuse or tunnel to the defects.

As stated in the manuscript, oxygen vacancies as defects can also trap charges but they have a much small atomic relaxation energy. This results in a much shorter characteristic time of trapping/detrapping, on the order of nanoseconds instead of milliseconds, which has a negligible impact on RTNs and cannot be resolved by normal testing.

3. I have checked the provided literature (Random Telegraph Noise Nano-spectroscopy in High-κ Dielectrics Using Scanning Probe Microscopy Techniques" in Noise in Nanoscale Semiconductor Devices: Springer,2020, pp. 417-440) and found no support for the claims. The authors should verify the citation. In the cited chapter as traps oxygen vacancies are considered!

We apologize for this confusion in reference. In this place, we intended to cite pages

107-117 of the same book instead of pages 417-440. We corrected the citation and will double check all citations of the manuscript. They also hypothesized that the RTN noise in their HfO_2 memristor can originate from oxygen interstitial defect and provided evidence based on statistical analysis of the electric signals (pages 111-112). The correct citation has been updated to the manuscript:

[40] Puglisi, F. M. "Noise in resistive random access memory devices." *Noise in Nanoscale Semiconductor Devices*. Springer, Cham, 2020. 87-133.

4. I have found in the reference list of the book (ref 88-91 in the book) theoretical calculations on oxygen interstitials including mentioning neutral interstitials. However, I would not say they support the mechanism suggested in this manuscript.

Indeed, we do not have direct experimental evidence to conclude that oxygen interstitial defects for sure lead to the RTN in HfO_2 memristors, though such defects are widely studied in computational work. We agree with the reviewer that involving oxygen interstitial defect is a hypothesis (only with indirect evidence) instead of a solid conclusion.

5. Additional comment: I strongly recommend using Kroeger-Vink notation (relative charges). In the classical notation (absolute charges) both lattice oxygen ions and interstitials are O2-. Mixing relative and absolute charge notations causes significant confusion for the readers.
We agree that the Kroeger-Vink notation is clearer in representing the charge of defect. We changed to Kroeger-Vink notation in the updated manuscript.

Referee #2 (Remarks to the Author):

My previous concerns have been adequately addressed by the authors.
Response: we appreciate the positive comments of the reviewer.

Referee #3 (Remarks to the Author):

The authors made detailed responses to the questions, which addressed my concerns. I have no further comments on the revised manuscript and can recommend publication.

Response: we appreciate the positive comments of the reviewer.

Changes made to the revised manuscript:

We removed all texts claiming that interstitial oxygen *is* the defect that traps/detraps electrons and cause RTN. Instead, we claimed that interstitial oxygen *is possibly* the RTN-responsible defect because: 1. Its calculated characteristic time is consistent with experiment result; 2. The strongly non-equilibrium condition during memristor forming may lead to interstitial oxygen defect generate near the filament. Because the incomplete filament blocking simulation (in the original Fig. 3c-f) treated the charge-trapped defect as a point charge, the results are not changed, and the figures are kept as Fig. 3a – d. The schematics in Fig. 3a and c (as well as their captions) are slightly modified to show that the defect is a charge trapping center (which may or may not be an interstitial oxygen defect).

"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. "

We moved the detailed analysis why interstitial oxygen defects are the likely the RTN-responsible defect into supplementary information (SI-14, Fig. S15 and Fig. S16).

We also did minor changes, including:

1. Fixed the reference issue (reference [40]), as the referee mentioned;
2. We changed the original notation to Kroeger-Vink notation in the updated manuscript;
3. We noticed that analog memristor is a promising candidate for future mortal computation, as G. Hinton brought up recently [*arXiv preprint arXiv*:2212.13345 (2022)]. So, we have added brief comment on that in the introduction/conclusion part of the manuscript.