nature materials

ARTICLES

https://doi.org/10.1038/s41563-017-0001-5

SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations

Shinhyun Choi[1,2], Scott H. Tan[1,2], Zefan Li[1,2], Yunjo Kim[1,2], Chanyeol Choi[1,2], Pai-Yu Chen[3], Hanwool Yeon[1,2], Shimeng Yu[3] and Jeehwan Kim[1,2,4]*

Although several types of architecture combining memory cells and transistors have been used to demonstrate artificial synaptic arrays, they usually present limited scalability and high power consumption. Transistor-free analog switching devices may overcome these limitations, yet the typical switching process they rely on—formation of filaments in an amorphous medium—is not easily controlled and hence hampers the spatial and temporal reproducibility of the performance. Here, we demonstrate analog resistive switching devices that possess desired characteristics for neuromorphic computing networks with minimal performance variations using a single-crystalline SiGe layer epitaxially grown on Si as a switching medium. Such epitaxial random access memories utilize threading dislocations in SiGe to confine metal filaments in a defined, one-dimensional channel. This confinement results in drastically enhanced switching uniformity and long retention/high endurance with a high analog on/off ratio. Simulations using the MNIST handwritten recognition data set prove that epitaxial random access memories can operate with an online learning accuracy of 95.1%.

Various types of analog switching device have been demonstrated as synapses for neuromorphic computing[1–7]. Most rely on filamentary switching mechanisms, such as oxide-based resistive random access memory (RRAM) and conductive-bridging RAM (CBRAM). Oxide-based RRAM operation is based on alignment of anion vacancies inherent in amorphous-phase binary oxides to form conductive filaments[6,8–11]. While these devices exhibit reasonably good retention and endurance, they suffer from a small on/off ratio and unavoidable temporal (cycle-to-cycle) and spatial (device-to-device) variation due to uncontrollable filament dynamics in an amorphous solid[5,6,8–11]. Resistive switching using single-crystalline-based ternary oxide films has been attempted, where dislocations become active filaments due to the self-doping effect of crystalline defects in $SrTiO_3$ (ref. [12]). However, the amorphous binary oxide has still been a mainstream because the switching performance is not superior to that of amorphous binary oxides. On the other hand, CBRAMs operate on the basis of metal conductive bridging through an amorphous solid electrolyte[4,13–17]. Owing to the high mobility of metal cations, the switching on/off current ratio of CBRAMs is substantially higher than that of the oxide-based RRAMs[4,18–20]. However, uncontrollable ion transport through defects in an amorphous films results in three-dimensional stochastic filament formation resulting in switching variation[13,14,18,21]. These make large-scale analog neural computing impractical without transistors at each resistive switching device. Thus, securing a strategy to confine the filament is an essential step[22].

Here we demonstrate single-crystalline SiGe epitaxial random access memory (epiRAM) with minimal spatial/temporal variations with long retention/great endurance, and a high analog current on/off ratio with tunable linearity in conductance update, thus justifying the suitability of epiRAM for transistor-free neuromorphic computing arrays. This is achieved through one-dimensional confinement of conductive Ag filaments into dislocations in SiGe and enhanced ion transport in the confined paths via defect-selective etch to open up the dislocation pipes. In SiGe epiRAM, the threading dislocation density can be maximized by increasing the Ge content in SiGe or controlling the degree of relaxation[23], and we discovered that 60-nm-thick $Si_{0.9}Ge_{0.1}$ epiRAM contains enough dislocations to switch in tens of nanometre scale devices. When this nanometre device is sampled after switching, Ag filaments confined in the dislocation are visualized via cross-sectional transmission electron microscopy (TEM). In addition, the epitaxy of p–i–p back-to-back diodes in SiGe epiRAM permits self-selection behaviour that can suppress the sneak path during large-scale array operation, and precise doping modulation during epitaxy allows one to modulate the set voltage and read current by varying the Schottky barrier height at the Ag/Si interface. Our simulation based on all of those characteristics of epiRAM shows 95.1% accurate supervised learning with the Modified National Institute of Standards and Technology (MNIST) handwritten recognition data set, which is comparable to a software training baseline of 97%. Thus, our finding is an important step towards developing large-scale and fully functioning neuromorphic hardware.

It is known that dislocation pipes provide preferential diffusion paths in crystalline solids[24]. Thus, injection of immiscible cations into single-crystalline films with dislocations can restrict filaments into dislocation pipes under electrical bias. The crucial step before cation injection is widening the dislocations to facilitate ion transport using defect-selective etching (see Fig. 1a for the schematics of the concept). We have performed heteroepitaxial growth of 60-nm-thick

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona, USA. [4]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Shinhyun Choi and Scott H. Tan contributed equally to this work. *e-mail: jeehwan@mit.edu

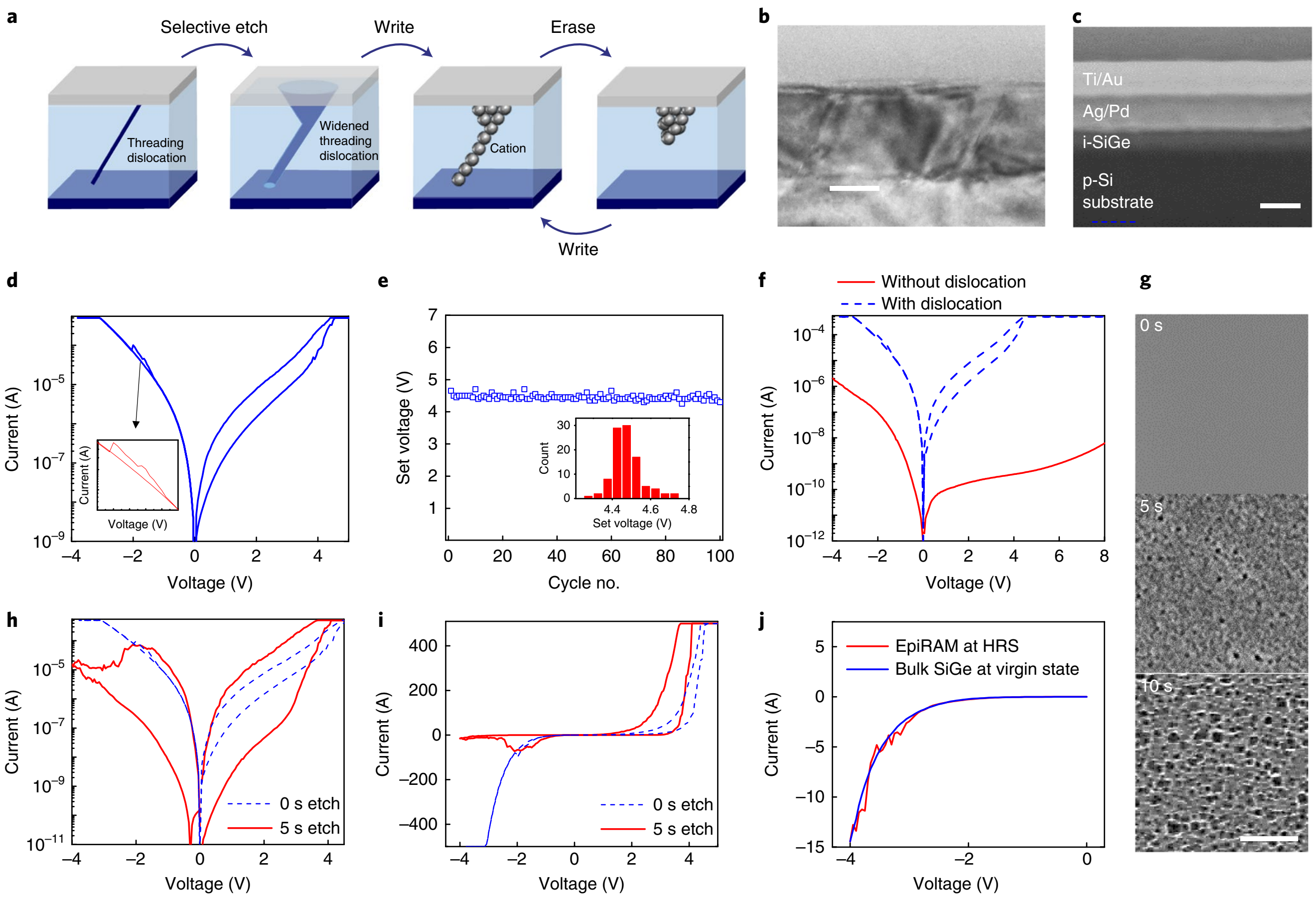

Fig. 1 | Impact of dislocation on the characteristics of the SiGe epiRAM. a, A conceptual schematic of the epiRAM during switching. **b**, Cross-sectional TEM image of 60 nm SiGe grown on a Si substrate. Scale bar, 25 nm **c**, Cross-sectional SEM image of an epiRAM device. Scale bar, 100 nm **d**, Measured d.c. *I*–*V* characteristics of epiRAM with unwidened dislocations. Inset: zoomed-in image to show difficult reset process. **e**, The set voltage variation of the epiRAM with unwidened dislocations over 100 quasi-static *I*–*V* sweeps. Inset: histogram for set voltage distribution. **f**, Measured d.c. *I*–*V* characteristics of a Ag/dislocation-free i-Si/p-Si substrate device, where no switching behaviour is observed even when applying a very high voltage. **g**, Plan-view SEM images of epiRAM after etching for 0 s, 5 s and 10 s. Scale bar, 200 nm. **h**, Semilogarithmic d.c. *I*–*V* characteristics of 0 s and 5 s etched epiRAM. **i**, Linear-scale d.c. *I*–*V* characteristics of 0 s and 5 s etched epiRAM. **j**, *I*–*V* characteristics of epiRAM at high resistance state (HRS) and SiGe at the virgin state.

intrinsic $Si_{0.9}Ge_{0.1}$ onto p-type Si(001) substrates at 750 °C, at which SiGe films can be partially relaxed[25,26]. In ultrathin partially relaxed heteroepitaxial films, threading dislocation density is typically very high due to incomplete relaxation[23]. The threading dislocation density in our SiGe films is counted to be in the range of $10^{11}/cm^2$, which is dense enough to provide dislocations in well-scaled devices[25,26] (see Supplementary Fig. 1 for the scanning electron microscopy (SEM) image showing the dislocation density). The vertically aligned threading dislocations are imaged by cross-sectional TEM (Fig. 1b), where the strain field from threading dislocations is observed as diagonal extensions through the epitaxial layer. The threading arms extended to the SiGe surface were visualized under SEM after a selective defect decoration (see Fig. 1g). Figure 1c shows a cross-sectional SEM image of an epiRAM device. Silver (Ag) was selected as the active metal due to its limited solid solubility in SiGe and its inability to form compounds with Si and Ge (http://www.factsage.cn/fact/documentation/binary/binary_figs.htm)[27,28]. Other metals that can form compounds with SiGe are not selected because it can form a very strong filament that is difficult to reset (http://www.factsage.cn/fact/documentation/binary/binary_figs.htm)[29]. As displayed in Fig. 1d,e, SiGe epiRAM with dislocations shows exceptionally uniform resistive switching with only 1.7% temporal set voltage variation (σ/μ) during 100 switching cycles without widening dislocations (see Methods for the method to define the set voltage). Notably, the current in the high-resistance state and that in the low-resistance state (LRS) also maintain temporal uniformity (see Supplementary Fig. 2). When amorphous Si is used instead of crystalline SiGe, undefined conductive paths and miscibility of Ag in the amorphous phase result in large temporal set voltage variation (28%) as shown in Supplementary Fig. 3. This clearly contrasts the set voltage uniformity achieved when employing SiGe epitaxial films. Switching behaviour is not observed when a dislocation-free homoepitaxial intrinsic Si film is used as a switching medium within the same device stack, as shown in Fig. 1f. This further suggests that threading dislocations in heteroepitaxial SiGe accommodate a Ag filament through the epilayer.

While the SiGe epiRAM with untreated dislocations demonstrates extremely uniform switching behaviour, measured digital and analog switching on/off ratios are ~10 and ~3, respectively. The small on/off ratio limits the number of distinctive conductive states for training large-scale neuromorphic arrays[30] (see Supplementary Fig. 4 for analog switching). We suspect this originates from incomplete filament rupture owing to the tight spatial accommodation of Ag into dislocations[31,32] (see the inset of Fig. 1d showing incomplete reset). To facilitate ionic transport through dislocations, we performed defect-selective etching to widen the dislocation pipes. SiGe epilayers on Si substrates were dipped into Schimmel etchant[33], which is conventionally used to selectively etch dislocations in SiGe for defect density characterization. Fig. 1g shows SEM images of the SiGe surface morphology at different etch times; wider threading dislocation pipes are observed after longer etching times. The cross-sectional TEM of epiRAM in the LRS (5-s-etched sample) indicates that the silver filament is well confined in the threading dislocation pipe (see Supplementary Fig. 5 for a TEM image showing the Ag filament in a widened dislocation). Dislocation-selective etch results in the following noticeable changes in the switching behavior of epiRAM: at negative bias, the device effectively resets; at positive bias, the on/off current ratio increases by three orders of magnitude; and the set voltage reduces by 0.7 V (see Fig. 1h,i). To evaluate the effectiveness of rupturing the filaments on reset, we compare the negative-bias *I*–*V* profile of bulk SiGe (with no filament formation history) to that of epiRAM after filament rupturing. As shown in Fig. 1j, the *I*–*V* curve of the epiRAM after the reset overlaps with that of bulk SiGe, indicating effective retraction of the Ag (ref. [30]; see Supplementary Fig. 6). This is further supported by the substantially reduced off-current at the positive bias. Moreover, the increased on-current and lowered set voltage of etched epiRAM indicate facilitated ionic movement into the dislocation pipes after etching. As a result, etched epiRAM with widened dislocations exhibits a higher on/off ratio (10^4) than non-etched epiRAM.

Filament confinement in dislocations results in exceptionally low temporal variation while the uniform distribution of dislocations throughout the SiGe film allows for low spatial variation (measured for 500 devices). These low variations are essential for accurate pattern learning and recognition when implemented into neuromorphic hardware[3,34,35]. We characterized the temporal variation of an epiRAM device by measuring the set voltage over 700 switching cycles. The measured temporal variation (σ/μ) is as low as 1% (see Fig. 2a). This cycle-to-cycle uniformity of epiRAM makes a clear contrast to that of many amorphous-based device architectures even after modification to improve temporal/spatial uniformity by metal doping, field localization by nanoparticles, or confinement of cation transport by nanopore graphene[36–40]. However, additional widening of dislocation channels (more than 5 s of etching) increases switching variation, possibly because of excessive stochastic pathways due to loss of the confinement effect (see Fig. 2b and Supplementary Fig. 7). In addition to exceptional temporal uniformity, epiRAM exhibit excellent spatial uniformity contributed by the well-distributed dislocations across the wafer (see Fig. 1g). The set voltage was mapped for 100 devices from two batches (Fig. 2c). All measured epiRAM devices show comparable resistive switching with spatial variation of only 4.9% and uniform batch-to-batch performance (see Fig. 2d). It should be noted that epiRAM unprecedentedly exhibits a high device yield with excellent temporal and spatial uniformity and such high uniformity is maintained in nanoscale devices (see Supplementary Fig. 8), thus justifying suitability of our epiRAM for implementation in high-density, large-scale neuromorphic arrays. As shown in Supplementary Fig. 8, all measured 25-nm-sized devices exhibit comparable performances to 5-μm-sized devices with high uniformity. This suggests that the threading density in our SiGe films (~10^{11} cm^{-2}) is sufficient enough to operate dense nanoscale devices. More interestingly, the *I*–*V* characteristics of epiRAM with size between 25 nm and 5 μm are comparable, and the Ge composition in SiGe does not severely alter the *I*–*V* characteristics (see Supplementary Figs. 9 and 10). These findings imply that switching occurs only in a localized area and a limited number of dislocations are probably responsible for the majority of ionic movement among all existing dislocations under the electrode. In addition, the *I*–*V* characteristics are maintained when the epiRAM is formed in the form of a crossbar, suggesting the suitability of epiRAM for neuromorphic computing (see Supplementary Fig. 11).

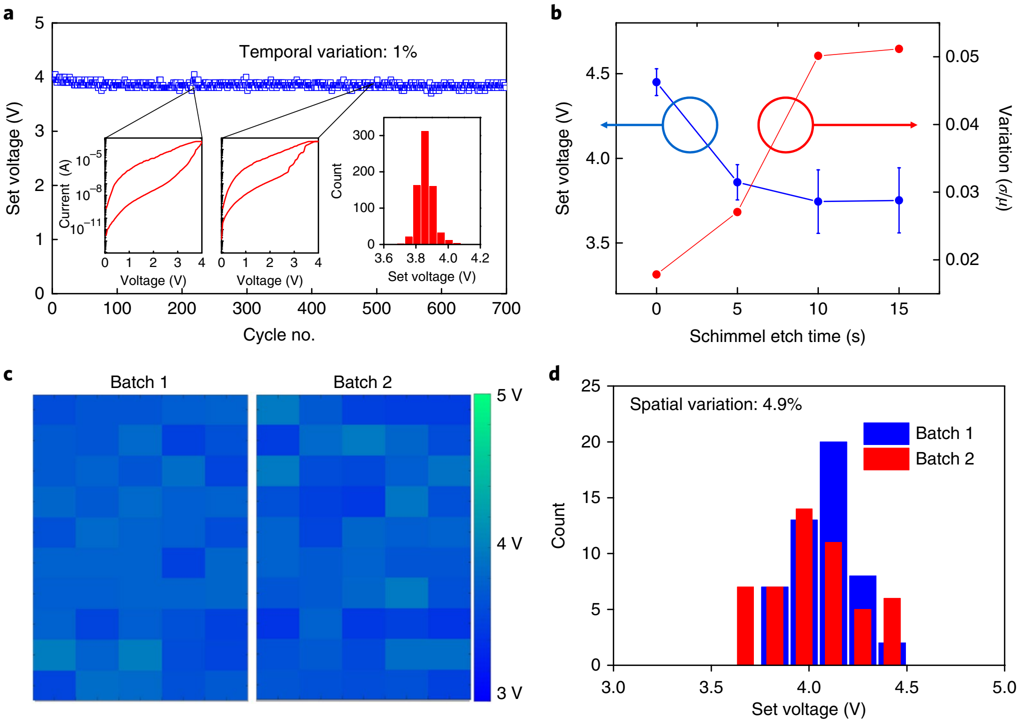

Fig. 2 | The characteristics of the SiGe epiRAM after widening the dislocation pipes with a dislocation-selective etch. a, The temporal set voltage variation of the epiRAM over 700 quasi-static *I*–*V* sweeps. Insets: d.c. *I*–*V* plots at the 250th cycle and the 500th cycle, along with the histogram for the set voltage distribution. **b**, The set voltage and its variation (σ/μ) depending on the defect-selective etch time. The error bars show the standard deviation from 100 d.c. cycles. **c**, Map of set voltages at devices overlapped on optical micrographs of the epiRAM devices (50 devices from batch 1 and 50 devices from batch 2 are shown in the map). **d**, Histogram for the spatial set voltage distribution shown in **c**.

For accurate artificial neural network training, the analog on/off ratio must be sufficiently high to access multiple synaptic weight values in neural network algorithms[30,34]. First, we have characterized the analog current on/off ratio of the epiRAM (5 s etch) by applying a pulse train consisting of set (5 V, 5 µs), reset (−3 V, 5 µs) for potentiation–depression (P–D) pulses and read pulses (2 V, 1 ms) after each P–D pulse. As shown in the P–D plot in Fig. 3a, a remarkably high analog on/off ratio of 240 is measured at 1,000 P–D pulses (500 potentiation/500 depression); the ratio decreases with reduced number of P–D pulses (180 for 400 P–D pulses and 100 for 200 P–D pulses). All 100 epiRAM devices fabricated in one batch exhibit a very high analog on/off ratio of more than 100 with an average value of 132 at 200 P–D pulses (see the analog switching map in Supplementary Fig. 12). EpiRAM also exhibits low temporal and spatial switching threshold variations for both set and reset transitions for the voltage pulse mode, which justifies the suitability of epiRAM for neuromorphic operation (see Supplementary Fig. 13).

While epiRAM exhibits an extremely high analog on/off ratio, the conductance response following P–D is nonlinear, a characteristic that is typical of other filamentary-type switching devices (see Fig. 3a). Such nonlinearity is more prominent when the epiRAM conductance saturates at its maximal value following maximized potentiation pulses and abruptly decays following depression (see Fig. 3a). At maximum potentiation, Ag^+ drift is no longer facilitated in widened dislocations at a given external voltage due to limited spatial capacity within the dislocation. Thus, we attempted to moderate potentiation of epiRAM with mild Ag^+ injection by intentionally decreasing the on/off ratio during the potentiation cycle to avoid saturation (total 200 P–D pulses with a pulse width of 5 µs). As shown in Fig. 3a,b, this successfully allows for a linear conductance response (see Fig. 3b) with an on/off conductance ratio of >100, which is sufficiently high for accurate training in MNIST pattern recognition[34,41]. This trade-off between linearity and the number of P–D pulses is reproducibly observed across the devices and its statistics are shown in Supplementary Fig. 14. This finding suggests that the dislocation has the spatial capacity to accommodate Ag, and that a linear conductance change can be obtained by not reaching its capacity with reduced P–D pulse numbers. Such a trend can also be observed in the P–D curve of epiRAM depending on the dislocation widening time with Schimmel etching in a fixed P–D pulse number. As shown in Supplementary Fig. 15, the linear conductance update observed in epiRAM with Schimmel etching for 200 P–D pulses disappears in epiRAM without Schimmel etching due to extremely limited space in dislocations.

The single-crystalline nature of heteroepitaxial SiGe epiRAM also displays long retention and great endurance. Conductive filaments are structurally retained for 48 h at 85 °C for epiRAM with widened dislocations (etched for 5 s) at the LRS after d.c. sweeps, which proves the long data retention capability of epiRAM (see Fig. 3c). The long data retention further indicates the following: the Ag filament is effectively confined in the dislocation by suppressing Ag diffusion into the bulk SiGe due to the lack of solid solubility of Ag in SiGe; and widened dislocation with 5 s etch provides optimal space to accommodate Ag ions (see Supplementary Fig. 16 for details about optimal dislocation spacing for data retention). This contrasts the case of conventional CBRAMs, where the metal in the filament can pressurize an amorphous switching medium[42] and easily diffuse into the amorphous phase, yielding reduced retention. The filament dynamics in epiRAM (that is, ion

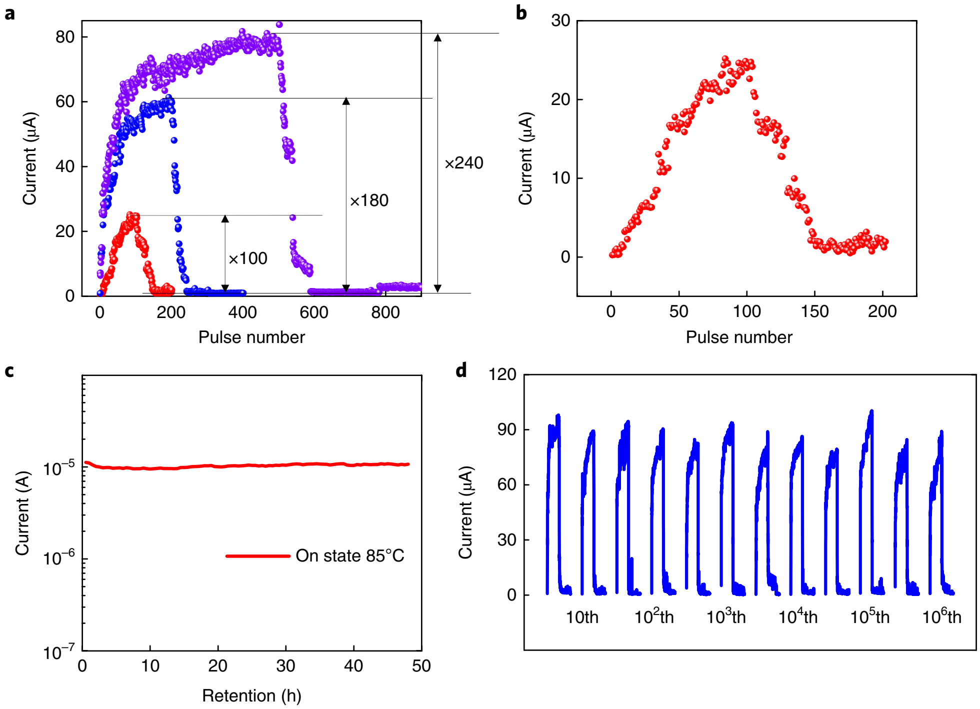

Fig. 3 | The characteristics of epiRAM for neuromorphic computing. a, Potentiation and depression of epiRAM by set/reset training pulses showing the analog on/off conductance ratio. The pulse train consists of consecutive set pulses (5 V, 5 µs) followed by consecutive reset pulses (−3 V, 5 µs). The pulse trains include 100/200/500 set pulses and 100/200/500 reset pulses, respectively. The current was measured at a 2 V read pulse after each set/reset pulse. **b**, The linear potentiation and depression. The pulse train consists of 100 set pulses followed by 100 reset pulses. The current was measured with a 2 V read pulse. **c**, The two-day retention test at 85 °C at the LRS with a read pulse (1.5 V, 1 ms). The device performance remains unchanged after the test. **d**, The endurance test stopped after 10^9 set/reset pulses (10^6 cycles). In the figure, the first and fifth cycles of each order of magnitude are shown. The test cycle has 500 consecutive set pulses and 500 consecutive reset pulses with read pulses after each set/reset pulse. The device performance remains unchanged after 10^6 cycles (10^9 pulses).

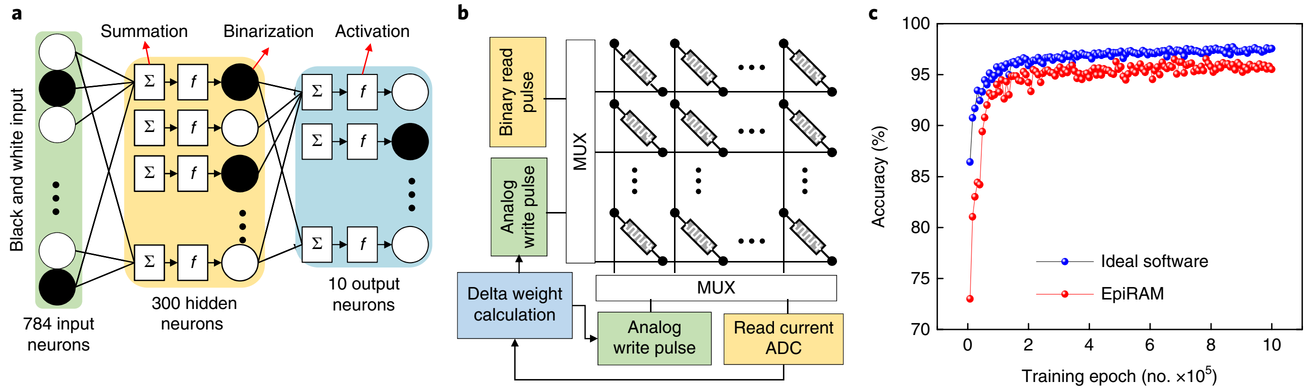

Fig. 4 | The image recognition simulation. a, A three-layer MLP neural network with a black-and-white input signal for each layer in the algorithm level. The inner product (summation) of the input neuron signal vector and the first synapse array vector is transferred after activation and binarization as the input vector for the second synapse array. **b**, Circuit block diagram of hardware implementation showing a synapse layer composed of epiRAM crossbar arrays and the peripheral circuit. MUX, multiplexer; ADC, analog-to-digital converter. **c**, Evolution of accuracy with training epochs for an ideal device and an epiRAM device.

and electron transport at elevated temperatures) are described in Supplementary Fig. 17. The immiscibility of Ag into SiGe in epiRAM also substantially enhances the endurance as shown in Fig. 3d (analog operation is capable of switching for over 10^9 set/reset pulses with a stable switching performance). On the other hand, the filament metal in conventional CBRAMs can progressively diffuse into porous amorphous switching medium and can be piled up into the amorphous switching medium as the switching cycle increases, yielding limited endurance.

On the basis of measured characteristics of epiRAM, we have simulated an artificial neural network to perform supervised learning with the MNIST handwritten recognition data set[43]. For the simulation, we utilized a three-layer neural network with 28×28 pre-neurons, 300 hidden neurons and 10 output neurons[44]. The multilayer perception (MLP) algorithm with stochastic gradient descent weight update is used in the simulation based on our epiRAM properties including non-ideal factors: finite on/off ratio, finite number of conductance levels, device-to-device variation, cycle-to-cycle variation, wire resistance and read noise. The 784 neurons of the input layer correspond to a 28×28 MNIST image, and the 10 neurons of the output layer correspond to 10 classes of digits (0–9)[44]. The detail of the three-layer MLP is shown schematically in Fig. 4a. The inner product (summation) of the input neuron signal vector and the first layer of the synapse matrix is transferred after activation and binarization as the input vector for the second synapse array. The schematic in Fig. 4b shows a synapse layer composed of epiRAM crossbar arrays and the peripheral circuit. Based on the inner product outcome of the input signal vector and the synapse vector by the read current, the amount of delta weight is calculated, which provides feedback into arrays to adjust the synapse weights by write pulses. After training with one million patterns randomly selected from 60,000 images from a training set, the recognition accuracy is tested with a separate set of 10,000 images from the testing set[45,46]. Our simulation proves that the neural network formed with epiRAM can achieve, on average, 95.1% recognition accuracy (96.5% as a maximum), which is comparable to the accuracy of 97% obtained by the software baseline as shown in Fig. 4c[44,47]. The impact of various device parameters on the recognition accuracy considered for our simulation and specific values for epiRAM are displayed in Supplementary Fig. 18.

One of the additional key advantages in epiRAMs is the layer-by-layer controllability of films involved in the devices during epitaxial growth. For example, the Schottky barrier height between Ag and Si can be precisely controlled by modulating the doping concentration of the Si epilayer right before SiGe epitaxy. As shown in Supplementary Fig. 19, the set voltage and read current can be modulated by varying the Schottky barrier height at the Ag/Si interface. Tunability of *I*–*V* characteristics with Schottky barrier heights in epiRAM could allow optimization of recognition accuracy and power consumption, and prevention of sneak paths. In addition, the layer-by-layer growth of p–i–p back-to-back diodes in SiGe switching medium permits self-selection behaviour as shown in Supplementary Fig. 20.

In conclusion, our unique design for resistive switching devices demonstrates characteristics required for implementing neuromorphic hardware. The confinement of the conducting filament into widened dislocations in SiGe offers superior spatial and temporal uniformity, long retention, excellent endurance, a high on/off ratio and linear weight update. In addition, the epitaxial growth of single-crystalline switching medium benefits from technologies in bandgap engineering to improve self-selection and low power consumption of the device. In possessing these properties, epiRAM-based neuromorphic arrays can achieve 95.1% learning accuracy. The development of epiRAM opens an avenue to realize fully functioning large-scale neural network beyond the conventional von Neumann computing algorithm. In addition, epiRAM meet properties required for digital non-volatile memory (see Supplementary Fig. 21).

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41563-017-0001-5.

Received: 11 July 2017; Accepted: 27 November 2017;
Published online: 22 January 2018

References

1. Jo, S. H. et al. Nanoscale memristor device as synapse in neuromorphic systems. *Nano Lett.* **10**, 1297–1301 (2010).
2. Burgt, Y. et al. A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing. *Nat. Mater.* **16**, 414–418 (2017).
3. Burr, G. W. et al. Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element. In *2014 IEEE Int. Electron Devices Meeting* 29.5.1-29.5.4 (IEEE, 2014).
4. Wang, Z. et al. Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nat. Mater.* **16**, 101–108 (2016).

5. Kim, S. et al. Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity. *Nano Lett.* **15**, 2203–2211 (2015).
6. Lee, M.-J. et al. A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structures. *Nat. Mater.* **10**, 625–630 (2011).
7. Gokmen, T. & Vlasov, Y. Acceleration of deep neural network training with resistive cross-point devices: Design considerations. *Front. Neurosci.* **10**, 333 (2016).
8. Shibuya, K., Dittmann, R., Mi, S. & Waser, R. Impact of defect distribution on resistive switching characteristics of Sr_2TiO_4 thin films. *Adv. Mater.* **22**, 411–414 (2010).
9. Park, G.-S. et al. In situ observation of filamentary conducting channels in an asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structure. *Nat. Commun.* **4**, 495707 (2013).
10. Yu, S., Guan, X. & Wong, H.-S. P. Conduction mechanism of TiN/HfO_x/Pt resistive switching memory: A trap-assisted-tunneling model. *Appl. Phys. Lett.* **99**, 063507 (2011).
11. Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. *Nature* **453**, 80–83 (2008).
12. Szot, K., Speier, W., Bihlmayer, G. & Waser, R. Switching the electrical resistance of individual dislocations in single-crystalline $SrTiO_3$. *Nat. Mater.* **5**, 312–320 (2006).
13. Kim, K.-H. et al. A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications. *Nano Lett.* **12**, 389–395 (2012).
14. Yang, Y. et al. Electrochemical dynamics of nanoscale metallic inclusions in dielectrics. *Nat. Commun.* **5**, 377–383 (2014).
15. Jo, S. H., Kim, K. H. & Lu, W. High-density crossbar arrays based on a Si memristive system. *Nano Lett.* **9**, 870–874 (2009).
16. Jo, S. H. & Lu, W. CMOS compatible nanoscale nonvolatile resistance switching memory. *Nano Lett.* **8**, 392–397 (2008).
17. Waser, R., Dittmann, R., Staikov, G. & Szot, K. Redox-based resistive switching memories - nanoionic mechanisms, prospects, and challenges. *Adv. Mater.* **21**, 2632–2663 (2009).
18. Yang, Y. et al. Observation of conducting filament growth in nanoscale resistive memories. *Nat. Commun.* **3**, 732 (2012).
19. Ielmini, D. & Waser, R. *Resistive Switching: From Fundamentals of Nanoionic Redox Processes to Memristive Device Applications* (Wiley-VCH, Weinheim, Germany, 2016).
20. Yang, J. J., Strukov, D. B. & Stewart, D. R. Memristive devices for computing. *Nat. Nanotechnol.* **8**, 13–24 (2013).
21. Krishnan, K., Tsuruoka, T., Mannequin, C. & Aono, M. Mechanism for conducting filament growth in self-assembled polymer thin films for redox-based atomic switches. *Adv. Mater.* **28**, 640–648 (2016).
22. Alibart, F., Zamanidoost, E. & Strukov, D. B. Pattern classification by memristive crossbar circuits using ex situ and in situ training. *Nat. Commun.* **4**, 2072 (2013).
23. Speck, J. S., Brewer, M. A., Beltz, G., Romanov, A. E. & Pompe, W. Scaling laws for the reduction of threading dislocation densities in homogeneous buffer layers. *J. Appl. Phys.* **80**, 3808 (1996).
24. Porter, D. A., Easterling, K. E. & Sherif, M. Y. *Phase Transformations in Metals and Alloys* (CRC Press, Boca Raton, USA, 2009).
25. Houghton, D. C. Strain relaxation kinetics in $Si_{1-x}Ge_x$/Si heterostructures. *J. Appl. Phys.* **70**, 2136–2151 (1991).
26. Romanov, A. E., Pompe, W., Beltz, G. & Speck, J. S. Modeling of threading dislocation density reduction in heteroepitaxial layers I. Geometry and crystallography. *Phys. Status Solidi* **198**, 599–613 (1996).
27. Rollert, F., Stolwijk, N. A. & Mehrer, H. Solubility, diffusion and thermodynamic properties of silver in silicon. *J. Phys. D* **20**, 1148 (1987).
28. Effenberg, G., Aldinger, F. & Prince, A. *Ternary Alloys* 211–221 (VCH, Weinheim, Germany, 1988).
29. Al-Joubori, A. A. & Suryanarayana, C. Synthesis of metastable $NiGe_2$ by mechanical alloying. *Mater. Des.* **87**, 520–526 (2015).
30. Yu, S. et al. Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect. In *2015 IEEE International Electron Devices Meeting (IEDM)* 17.3.1–17.3.4 (INSPEC, London, 2015).
31. Hull, R. *Properties of Crystalline Silicon* (Institution of Electrical Engineers, 2006).
32. Wells, A. F. *Structural Inorganic Chemistry* (Oxford University Press, New York, USA, 2012).
33. Schimmel, D. G. Defect etch for <100> silicon evaluation. *J. Electrochem. Soc.* **126**, 479–483 (1979).
34. Chen, P.-Y., Gao, L. & Yu, S. Design of resistive synaptic array for implementing on-chip sparse learning. *IEEE Trans. Multi-Scale Comput. Syst.* **2**, 257–264 (2016).
35. Prezioso, M. et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61–64 (2015).
36. Lee, J., Du, C., Sun, K., Kioupakis, E. & Lu, W. D. Tuning ionic transport in memristive devices by graphene with engineered nanopores. *ACS Nano* **10**, 3571–3579 (2016).
37. You, B. K., Byun, M., Kim, S. & Lee, K. J. Self-structured conductive filament nanoheater for chalcogenide phase transition. *ACS Nano* **9**, 6587–6594 (2015).
38. Liu, Q. et al. Improvement of resistive switching properties in ZrO_2-based ReRAM with implanted Ti ions. *IEEE Electron Device Lett.* **30**, 1335–1337 (2009).
39. Chang, W. Y., Lin, C. A., He, J. H. & Wu, T. B. Resistive switching behaviors of ZnO nanorod layers. *Appl. Phys. Lett.* **96**, 242109 (2010).
40. Yoon, J. H. et al. Highly improved uniformity in the resistive switching parameters of TiO_2 thin films by inserting Ru nanodots. *Adv. Mater.* **25**, 1987–1992 (2013).
41. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R. & Bengio, Y. Quantized neural networks: training neural networks with low precision weights and activations. Preprint at https://arxiv.org/abs/1609.07061 (2016).
42. Ambrogio, S., Balatti, S., Choi, S. & Ielmini, D. Impact of the mechanical stress on switching characteristics of electrochemical resistive memory. *Adv. Mater.* **26**, 3885–3892 (2014).
43. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2323 (1998).
44. Kataeva, I., Merrikh-Bayat, F., Zamanidoost, E. & Strukov, D. Efficient training algorithms for neural networks based on memristive crossbar circuits. In *Proc. Int. Joint Conf. Neural Networks* 1–8 (IEEE, 2015).
45. Chen, P.-Y., Peng, X.C. & Yu. S. *User Manual of MLP Simulator (+NeuroSim)* (accessed 1 January 2017); https://github.com/neurosim/MLP_NeuroSim
46. Chen, P.-Y., Peng, X. & Yu, S. NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures. *IEEE Int. Electron Devices Meeting (IEDM)* (IEEE, San Francisco, USA, 2017).
47. Prezioso, M. et al. Modeling and implementation of firing-rate neuromorphic-network classifiers with bilayer Pt/Al_2O_3/TiO_{2-x}/Pt memristors. In *Technical Digest - International Electron Devices Meeting, IEDM*, 17.4.1–17.4.4 (IEEE, 2016).

Acknowledgements

This work is partially supported by NSF-SRC-E2CDA under contract no. 2018-NC-2762B. We thank J. J. Yang, Q. Xia, P. Lin, Y. Li, M. Rao and Y. Zhuo of the University of Massachusetts for valuable help and fruitful discussion. We also thank S. Kim of the IBM T.J. Watson Research Center for valuable suggestions for experiments. This work was performed in part at the Micro Technology Laboratories (MTL) at the Massachusetts Institute of Technology, and in part at the Harvard University Center for Nanoscale Systems (CNS), supported by the National Science Foundation under NSF ECCS award no. 1541959.

Author contributions

S.C. and J.K. conceived this work and J.K. directed the team. S.C., S.H.T. and J.K. designed experiments. S.C., S.H.T., Z.L. and J.K. prepared the manuscript. S.H.T. and Y.K. carried out the epitaxial growth experiments and characterization. S.C., S.H.T., C.C. and H.Y. performed the device fabrication and electrical measurements of epiRAM devices and TEM/SEM characterization. Z.L., P.-Y.C. and S.Y. performed the simulation work. All authors discussed and contributed to the discussion and analysis of the results regarding the manuscript at all stages.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41563-017-0001-5.

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

Correspondence and requests for materials should be addressed to J.K.

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

Methods

Epitaxial growth. SiGe was epitaxially grown on p^+ Si(001) substrates in a low-pressure chemical vapour deposition (LPCVD) reactor equipped with a close-coupled showerhead. Silane (SiH_4) and germane (GeH_4) were used as precursors for Si and Ge sources. Growth of 60-nm-thick SiGe films was performed at 100 Torr at 750 °C. Diborane was used as the precursor gas for p-type-doped Si and SiGe layer deposition. Hydrogen was used as a carrier gas.

EpiRAM fabrication. Proceeding epitaxial growth, an etching technique, called the Schimmel etch, was performed. Devices were immersed in a mixture of 44% 32 M Cr solution and 64% hydrofluoric acid for 5 s. Then, 100 nm of Al was evaporated and annealed to form an ohmic contact. A 100-nm-thick SiO_2 layer was deposited by plasma-enhanced chemical vapour deposition and etched back by buffered oxide etch (5:1) solution after traditional photolithography (active area: 5 µm by 5 µm unless specified). Ag (100 nm)/Pd (20 nm) was deposited for the top electrodes and 5 nm Ti/100 nm of Au was deposited for the contact pads.

Device measurements. Quasi-static d.c. current–voltage (I–V) measurements were executed with a B1500A semiconductor device parameter analyser with a B1517A high-resolution source/measure unit. EpiRAM devices were tested with bidirectional I–V sweep measurements with current compliance ranging from 10 µA to 50 mA. Retention measurements were performed in a vacuum at elevated temperature with a LakeShore model TTP4-1.5K probe station. Pulse data were collected by a custom-built measurement system.

Analog measurements. For reproducible training, stabilization steps composed of repeating 30 sets of 200/400/1,000 P–D pulses were applied to stabilize Ag within dislocations. We measured the linearity of the conductance change as this can eliminate the need for peripheral circuits to compensate for nonlinearity in neuromorphic circuits. Linearity of the P–D plots is determined after the stabilization step as the training proceeds to the stabilization. The potentiation slope changes depending on the stabilization history because the strength of the post-stabilized Ag filament in the dislocation can vary depending on the number of depression pulses.

Transmission electron microscopy. Cross-section TEM samples were prepared by an NVision 40 dual-beam focused-ion-beam imaging system with an OmniProbe nanomanipulator. Platinum was deposited using the electron beam and ion beam. The cross-section area of interest was milled and glued to a lift-out grid, and then thinned and polished down to ~100 nm. Cross-sectional imaging was performed with a JEOL 2100 transmission electron microscope.

Threshold voltage definition. We measured the threshold voltage where the current passed 300 µA during the set process (constant-current threshold voltage), which is a popularly accepted method to measure threshold voltage[48].

Array simulation. The simulation is conducted on the basis of the platform "+NeuroSim". The source code is written with C++ programming language and is able to run on LINUX operation systems. A three-layer MLP neural network (784 × 300 × 10) is used with 28 × 28 MNIST images. After training with one million patterns (randomly selected from the 60,000-image training set), the system is used to recognize a separate 10,000-image testing set with non-ideal factors including finite on/off ratio, spatial/temporal variation, read noise, wire resistance, and quantization of read currents. The original patterns from the MNIST database are converted to black-and-white patterns with a threshold of 128 for pixel values ranging from 0 to 255. A logistic function is used as the activation function. The optimized learning rate for the first and second layer of the synapse is 0.4 and 0.2, respectively. The read voltage is 2 V and the read-out current is quantified to 8 bits by the analog-to-digital conversion circuit. The cycle-to-cycle variation describes the variation of conductance values at each level. We assume the conductance at each level obeys a normal distribution. The cycle-to-cycle variation is defined as the standard deviation divided by the maximum conductance[34,49]. The device-to-device variation describes the variation of the parameter A, which is explained in the Supplementary Information and ref. [49]. We assume that the fitting parameter A obeys a normal distribution. Device-to-device variation is defined as the standard deviation divided by the average value of A. Wire resistance between each crosspoint is 5Ω based on standard 14 nm CMOS technology. Read noise is chosen as 5%. The read-out current is quantified, normalized, and transferred to subsequent controlling logic circuits to calculate the delta weight. The conductance update is implemented with half-voltage operation and the entire array is written line-by-line. The peripheral circuit and most of the neuron circuit have been verified by HSPICE simulation; only the delta weight calculation is performed by software. The recognition accuracy is calculated every 8,000 images during each training process and the accuracy at each data point for the analysis in the paper is the average value for the last ten accuracy points.

Data availability. The data from this study are available from the corresponding author on reasonable request.

References

48. Ortiz-Conde, A. et al. A review of recent MOSFET threshold voltage extraction methods. *Microelectron. Reliab.* **42**, 583–596 (2002).
49. Gao, L. et al. Fully parallel write/read in resistive synaptic array for accelerating on-chip learning. *Nanotechnology* **26**, 455204 (2015).