Memristor-based AI Hardware for Reliable and Reconfigurable Neuromorphic Computing

by

Chanyeol Choi

B.S., Electrical and Electronic Engineering,
Yonsei University (2018)
S.M., Electrical Engineering and Computer Science,
Massachusetts Institute of Technology (2019)

Submitted to the Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

© Massachusetts Institute of Technology 2021. All rights reserved.

Author ....

Department of Electrical Engineering and Computer Science
August 20, 2021

Certified by ....

Jeehwan Kim
Associate Professor of Mechanical Engineering
Thesis Supervisor

Accepted by..

Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Memristor-based AI Hardware for Reliable and Reconfigurable Neuromorphic Computing

by

Chanyeol Choi

Submitted to the Department of Electrical Engineering and Computer Science
on August 20, 2021 in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

In the field of artificial intelligence hardware, a memristor has been proposed as an artificial synapse for creating neuromorphic computer applications. Changes in weight values in the form of conductance must be identifiable and uniform to train a neural network in memristor arrays. Because of the high mobility of metal ions in the Si switching medium, an electrochemical metallization (ECM) memory has shown a high analogue switching capacity. However, switching unpredictability is caused by the extreme stochasticity of ion transport. I demonstrated a Si memristor with alloyed conduction channels that works dependably and enables large-scale crossbar array deployment. In addition, heterogeneously integrated neuromorphic chips have been developed to allow physically reconfigurable neuromorphic computing. This thesis examines alloyed metal-based silicon memristors and stackable neuromorphic chips with heterogeneous integration for reliable and reconfigurable neuromorphic computing.

Thesis Supervisor: Jeehwan Kim
Title: Associate Professor of Mechanical Engineering

Acknowledgements

I would like to begin with a huge thanks to my advisor, Jeehwan Kim. This thesis represents work which would never have been possible without his scientific enthusiasm and astute advice. He helped make my graduate life productive, exciting, and fun. I also would like to thank my PhD committee members, Prof. Jing Kong and Prof. Farnaz Niroui for their invaluable time and precious feedback. I would like to thank Dr. Hyunseok Kim for his unconditional support and for mentally guiding me through my PhD. Thank you to Prof. Sanghoon Bae for being my mentor since 2015. His dedication and passion for research are an inspiration to all around him. I would like to thank Dr. Hanwool Yeon for his instruction and advice on fabrication processes and for sharing his exceptional knowledge on materials science. Thank you to Prof. Peng Lin for advising and inspiring me. He is an exceptional scientist, and I cherish every second of our discussions. I want to thank him for standing by me through the ups and downs of the PhD process. I would like to thank Prof. Kyusang Lee for his support and guide. Thank you to Hoyoung Kim, who has supported me since I worked at Qualcomm in 2014. Thank you to Jiyoung Ahn for always being a source of advice. I would like to mention and recognize my team members, Dr. Scott Tan, Kuan Qiao, Doyoon Lee, Dr. Jaewoo Shim, Dr. Sunmin Suh, Yeongin Kim, Dr. Yunjo Kim, Sangwook Han, Dr. Haneol Lee, Dr. Wei Kong, Dr. Jiho Shin, Dr. Sungkyu Kim, Dr. Hyun S. Kum, Seungju Seo, Jaeyong Lee, Yongmo Park, Jaekang Song, Chansoo Lee. Thank you to my EECS classmates and my friends from the United States, Canada, and France, Soomin Yoon, Maxx Gong, Khiry Kemp, Michael White, Jiwon Lee, Claire Sue Hyon Lee, Lizzie Roberts, Blake Slattengren, Anaïs Capela Teixeira. I am grateful that I am part of Myungdai family, BSK, Baejeong friends. And I would like to acknowledge my dearest friends, Donggyun Lee, Eunho Choi, Yu minjung, Seungyeon Kim. Specially thank you to the people who help me get through Covid-19: Charlotte Reed, Mayla Thompson, Harry

Hartley, Doeon Lee, Caitlin Tierney, Youngmin Baek, Minseong Park, Byungjun Bae, Byunghyun Lee, Janghoon Woo, Jimin Kang, Taehoon Jeong Mina Kim, Kori Groenveld, Eunseok Lee, Jaeyeon Won, and Chulhee Yun. I also would like to thank our team, Dr. Kanghyun Ryu, Dr. Ikbeom Jang, Dr. Yoseob Han, Dr. Young Woo Choi, Subeen Pang, Minyoung Huh, and Guyeong Jung. Lastly, I would like to thank my family. Without the support of my family, I would never have been able to achieve my dreams of studying in the US and receiving an education.

Contents

1 Introduction to Neuromorphic Computing **28**

1.1 Computing systems. 28

1.2 Memristor: resistive random access memory 33

1.3 Research goals and thesis organization...42

2 Alloying Conducting Channels for Reliable Neuromorphic Computing **43**

2.1 Introduction...44

2.2 Contributions and methods 45

2.3 Conclusion 102

3 Stackable Hetero-Integrated Chips for Reconfigurable Edge Neuromorphic Computing **104**

3.1 Introduction...105

3.2 Contributions and methods 106

3.2 Conclusion 138

4 Conclusion **139**

3.1 Conclusion 139

3.2 Future work...140

5 Bibliography **143**

List of Figures

Figure 1-1: Types of bipolar and unipolar two terminal RRAM devices. a, bipolar RRAM device schematic. Depending on the size of filament inside of switching medium, the resistance value of RRAM device is determined. (Orange: OFF, yellow: ON). b, unipolar RRAM device schematic. c, I-V characteristics of bipolar RRAM device. HRS and LRS stand for high resistance value and low resistance value. SET process forms or strengthens a filament and RESET process disrupts or weakens the filament. d, I-V characteristics of unipolar RRAM device

....40

Figure 1-2: Challenges in memristor device performance matrices. Six device matrices are presented. Each number in round brackets represents good device performance from ref [1]–[6]. It is worth noting that none of RRAM devices developed can meet all six requirements at the same time. 41

Figure 2-1: Figures show the d.c. switching performance of a Si memristor using an Ag–Cu alloy. a, Following the forming process, typical current–voltage curves of Ag (left) and silicidable metals (Cu, Ni, Ti, and Cr) (right). b, Retention characteristics of Ag devices with several conductance levels (measured at 0.5 V). c, The effect of the Ag–Cu thickness ratio on alloying during production on d.c. switching uniformity—normalized on/off uniformity over 100 cycles. d, Uniform switching of the Ag–Cu device during 100 cycles. The nominal thicknesses of Ag and Cu are 2 and 1 nm, respectively. e, Histogram for the set voltage distribution of the Ag (black) and Ag–Cu (red) devices illustrated in a and d. f, Improved

retention characteristics of Ag–Cu devices (measured at 0.5 V) with varying compliance currents...47

Figure 2-2: The impact of alloying on d.c. switching performance is kinetically and thermodynamically regulated. 100 switching curves (top), set voltage histogram (middle), and retention characteristics with different compliance currents for Ag–Ti, Ag–Cr, and Ag–Ni (bottom). Ag and the alloying element have nominal thicknesses of 2 and 1 nm, respectively. 49

Figure 2-3: The impact of alloy on the analogue nanosecond switching behavior of Si memristors. a,b, Average and standard deviation of 50P/50D ten-cycle conductance updates in Ag–Cu alloy (a) and pure Ag (b) under the same pulse condition (pulse conditions: potentiation (P) of 50 ns, 4.8 V, and n = 50, depression (D) of 50 ns, –2.9 V, and n = 50, and V_{read} of 1 V, 1 ms. c, PDF of the ANL and G contrast from ten devices (50P/50D, five cycles each) for the Ag–Cu alloy (red) and pure Ag (grey). ANL = (GP(N/2) – GD(N/2)/(G_{max} – G_{min}), where G_{max}, G_{min}, GP(N/2) and GD(N/2) indicate the maximum conductance, minimum conductance, median value of potentiation, and medium value of depression, respectively. G contrast equals G_{max}/G_{min}. d, Endurance test of a Si memristor based on the conduction channel of an Ag–Cu alloy. The conductance was programmed at 50P/50D for 30 cycles. The following pulse conditions were used for the endurance test: potentiation of 50 ns, 5 V, and n = 50, depression of 50 ns, 3 V, and n = 50, and Vread of 1 V, 1 ms...53

Figure 2-4: 32 x 32 Si memristor arrays with Ag and alloy active electrodes. a, Illustration of an Ag–Cu alloy memristor chip. b, An optical micrograph of a single 32 x 32 array. 240 μm scale bar c, SEM picture of a section of the array revealing the crossbar structure. 15 μm scale bar c, Image programming and data retention experiments in Ag (top), Ag–Ni (middle), and

Ag–Cu (bottom) arrays. e, Convolutional kernel processing in the Ag–Cu array, demonstrating the Ag–Cu array's computational capabilities. 55

Figure 2-5: The deposition of Ag-Cu films for active metals. We designed a step-by-step metal deposition method to push Ag-Cu together into a switching medium during the shaping stage. We began by depositing ultrathin Ag islands on switching medium, the thickness of which controls the opening area of Si (step1). Following that, we deposited ultrathin Cu on top of Ag islands-deposited Si, where the Cu islands make direct contact with the switching medium (step2). Finally, we enclosed the Ag-Cu islands in a 15 nanometer thick Ag film (step3). The 'Ag-Cu thickness ratio' determines the quantity of Cu involved in switching. The more Ag film is deposited, the less Cu contributes to switching. During the formation process, the Ag-Cu alloying conduction channels are created in the switching medium.....................................57

Figure 2-6: The development of an Ag-Cu alloy. SEM images and energy dispersive X-ray (EDX) mapping of an Ag-Cu alloying layer on an amorphous Si surface. 400 μm scale bars When the overall thickness was 2 to 4 nanometers, discontinuous metal films were produced with evenly dispersed metal clusters. Metal clusters merged at a 7-nanometer-thick Ag film with a 2-nanometer-thick Cu film. 58

Figure 2-7: The impact of the Ag-Cu alloying ratio on DC switching uniformity and retention. Normalized on/off uniformity of I Ag and (ii) Cu layers with regard to nominal thickness. Following the step-by-step evaporation, a 15-nanometer-thick additional Ag layer was added. For this mapping, the best-performing device for 100 cycles at a compliance current of 5 mA was chosen for each alloying condition. DC switching curves with temporal on/off conductance changes, as well as room temperature retention data, are also provided. When Cu was evaporated even 1 nm before Ag, transistors exhibited irreversible breakdown behavior comparable to pure Cu devices (20-nanometer-thick Cu layer). Switching performance

dynamically varied as Ag thickness grew under fixed Cu thickness (1 nanometer), and Ag (2 nanometer)/Cu (1 nanometer) layers generated optimum switching performance: extremely uniform switching with steady retention behavior at multi-level states. As the Ag-Cu ratio diverged from 2 nanometer-1 nanometer, non-uniform switching with poor retention (7 nanometer-1 nanometer) or on/off degradation occurred, although consistent data retention (2 nanometer-2 nanometer) was found. These findings clearly indicate that Cu additions in Ag active electrodes have a substantial effect on switching performance, even though the quantity of Cu is too tiny to create a continuous layer on the Si surface. In addition, Cu's function may be described as follows. (1) Cu improves the stability of an Ag-based conduction channel while lowering the maximum on/off ratio owing to residual Cu elements linked to the Si switching medium (called backbone of the conduction channel). (2) Excess Cu in the Ag active electrode reduces the on/off window as the cycle number increases. An adjusted Ag-Cu ratio, on the other hand, may promote uniform switching with reasonably reliable data retention....................59

Figure 2-8: Layout of crossbars for alloy arrays with metal capping. In large array implementations, metal capping for p+ Si bottom electrodes is recommended to minimize line resistance. To create an alloy array with a gold capping layer, a novel method was devised. (a) Photolithography and dry etching are used to create isolated a-Si/p+ Si line patterns on an SOI wafer. Active device regions are shown by the protrusions on the line patterns. (b) putting a cap of Au on top of the p+ line patterns to decrease line resistance. The active zones are unaffected. (c) the bottom electrodes are passivated while the active regions are exposed. (d) finishing device arrays by patterning the top electrodes..61

Figure 2-9: Demonstration of an Ag-Cu alloy memristor array for inference. (a) For parallel kernel operation, four convolutional kernels illustrated in (b) were programmed into four columns of the 32 32 array. As a differential pair, two memristors are utilized to represent both positive and negative weights. 62

Figure 2-10: Phase diagram of metal-silicon. a, Ag-Si[7]. b, Ti-Si[8], c, Cr-Si[9]. d, Ni-Si[10]. e, CuSi[11]. With the exception of Ag, the production of silicides is thermodynamically favored for the elements Ti, Cr, Ni, and Cu. 72

Figure 2-11: Ag memristor DC switching uniformity. Temporal change in the set voltage (a) and on/off conductance (b) of an Ag device, as seen in the manuscript's Fig. 2-10a. Variation in set voltage (c) and on-off conductance throughout space (d). Each batch contains 5 devices that have been tested under the identical DC working conditions (compliance current, 5 mA, >100 DC cycles per device). The standard-deviation-to-mean of set voltage and on/off ratio were calculated to be 16.2 percent and 156.2 percent for batch 1 and 18.7 percent and 189.7 percent for batch 2. 73

Figure 2-12: Phase diagram of metal-silver. a, Ti-Ag[12]. b, Cr-Ag[13]. c, Ni-Ag[14]. d, Cu-Ag[15]. Ti produces intermetallic compounds with Ag, showing that Ti and Ag have an attraction force. Despite the fact that Cr, Ni, and Cu form a solid solution system with Ag, a miscible zone exists in the Ag-rich phase of Cu-Ag alloy. Thus, Cu-Ag can form a thermodynamically stable mixed compound (i.e., a solid solution), whereas repulsion force occurs at Cr-Ag and Ni-Ag regardless of the mixing ratio or temperature.

....74

Figure 2-13: Metal diffusivity in Si. a, Diffusion barrier height of Ti[16], Cr[17], Ni[18], Cu[19], and Ag[20]. b, The metal diffusivity Arrhenius plot. Cu, Ni, and Cr diffuse quicker than Ag, whereas Ti is the slowest metal. 75

Figure 2-14: The interfacial energy and relaxed structures of an Ag-Cu layer on a Si switching medium When we look carefully at the interface of metal cluster and Si medium, as shown in the schematic (left), there are three probable possibilities as described: (1) pure silver, (2) an Ag-Cu alloy, and (3) pure copper. The addition of Cu reduces interfacial energy, while the

interfacial energy of pure Ag clusters is 55 meV Å^{-2} higher than pure Cu clusters. This means that the external pressure on Ag-Cu clusters is reduced due to lower interfacial energy when compared to pure Ag clusters, and that alloying Cu with Ag can improve the thermodynamic stability of Ag-based conduction channels in Si switching medium. Supplementary Note 1.1 has the simulation information...76

Figure 2-15: Switching dynamics based on alloying conduction channel: Forming (left), reset (middle), and set (right) states. In this simulation, we look at a conduction channel generated by mixing Ag and Cu atoms in a Si switching medium. Actual alloying has been considered while developing with an incoming uniform mixture of Ag/Cu atoms inside a switching medium. For the initial state, we examined the atomic fractions of Ag and Cu, as well as their activation energies for anode cation dissolution. The 1 s state of formation is the stage at which Ag and Cu atoms begin to dissolve into the Si switching medium based on their activation energy. Furthermore, during reset/set procedures, Ag clusters are dissolved/rejuvenated to a greater extent than Cu clusters. These leftover Cu atoms can serve as the backbone of the conduction channel, improving switching uniformity and conductance state stability. Supplementary Note and Table 2-1 give the simulation circumstances and parameters, respectively.

....77

Figure 2-16: Schematic representation of the KMC simulation. 78

Figure 2-17: Ag-Cu memristor spatial variation. The cumulative probability of (a) the set voltage and (b) the on-off conductance (read voltage, 0.6 V). Each batch contains 5 devices that have been tested under the identical DC working conditions (compliance current, 5 mA, >100 DC cycles per device). The standard-deviation-to-mean (/) of set voltage and on/off ratio

were calculated to be 5.1 percent and 49.4 percent for batch 1 and 4.9 percent and 47.1 percent for batch 2.

..........79

Figure 2-18: Typical forming and subsequent reset processes of pure Ag devices (a) and Ag-Cu alloy devices (b). It should be noted that high formation voltage (>10 V) is undesirable for memristor crossbar array operation because it can cause irreversible breakdown of the devices and, as a result, reduced device yield [21], [22]. The forming voltage of Ag and Ag-Cu devices is roughly 3.7 V, which is 12 V greater than the fixed voltage (c). We believe that this difference is allowable for the array operation, which is strongly supported by the high yield (100 percent) of Si memristor crossbar array. 80

Figure 2-19: Retention test with raised temperature for 1 h. At room temperature (a) and 85 °C (b), conductance levels (over 10 μS) were stable, but lower conductance levels could not be achieved due to poor stability and increased off-state conductance caused by thermal excitation of free carriers from the p+ -Si layer, respectively. However, as temperature rose to 120 °C (c), conductance steadily declined, which is similar to the retention behaviors of pure Ag devices at room temperature. 81

Figure 2-20: The effect of ambient moisture on the memristive performance of Ag-Cu devices. DC switching (a), on- and off-state conductance (b), and retention characteristics (c) as a function of relative humidity (% RH) at ambient temperature. Because moisture (H2O) in a switching medium plays a significant role in redox-based switching dynamics, the level of ambient moisture can influence the switching performance of redox-based memristors [23]–[25]. Despite the fact that the humidity level was altered from 12 to 60 percent RH, Ag-Cu devices demonstrated similar switching behaviors and retention properties. This consistent behavior could be due to the device's passivation layers, which prevent H_2O migration into the

Si switching medium: Cr[26], [27] and silicon nitride[28]–[30] layers, known as excellent materials for preventing water molecules/ions from penetrating (i.e., anti-corrosion), cover Ag-Cu active metal and Si switching medium. 82

Figure 2-21: ANL and G distinguish between temporal and spatial differences in analog switching. For conductance update from 10 Si memeristor devices with three distinct active metals (a) Ag-Cu alloy, (b) pure Ag, and (c) Ag-Ni alloy, 5-cycles of 50 potentiation and 50 depression pulses are used. Pulse condition: Potentiation (50 ns, 4.8 V, n = 50), Depression (50 ns, -2.9 V, n = 50), Vread (50 ns, -2.9 V, n = 50). (1 V, 1 ms).

....83

Figure 2-22: Endurance of Ag-Cu alloyed device. Pulsed voltage stresses (PVS) test is performed under a square pulse condition (non-sinusoidal periodic waveform) on small device cells (< 25 μm^2). Each pulse cycle is made up of two processes: potentiation and depression. While the potentiation process (VP) uses 50 pulses of 4.5 V (amplitude) and 50 ns (duration), the depression process (VD) uses 50 pulses of -2.8 V (amplitude) and 50 ns (duration). Data points representing ON (red dots) and OFF (black dots) states are gathered at a power of 2 DC cycles (2n, n = 0 to 24) with read voltage (0.5 V). Throughout the PVS test (> 109 pulses), the device maintained a high on/off ratio (> 100), which is significantly greater than 5. (considered as device failure) [31], [32]. DC condition: set (4 V), reset (-3.2 V), and compliance current (5 mA).

....84

Figure 2-23: The variation in conductance (ΔG) as a function of pulse amplitude and duration. The conductance change (ΔG) is measured using a nanosecond pulse. At a read voltage of 1 V, we set the initial conductance at 10^{-6} S. (a) Change in conductance with increasing square pulse voltage amplitudes (V = 2, 3, 4, and 5). The pulse condition for conductance measurement is

shown inset in (a). (b) Change in conductance with different square pulse durations at the nanosecond level using 4 V nanosecond pules. The conductance shift with ultrashort pulses is seen in the inset (left) (10 ns, 30 ns, and 50 ns with 5 V nanosecond pulses). The pulse condition for conductance is shown in the inset (right) in (b). In both cases, a 5 μm × 5 μm cell is used.
....85

Figure 2-24: Analog potentiation and depression in 512 steps (512P/512D). For conductance update in an AgCu alloy Si memristor device, three cycles of 512 potentiation and 512 depression pulses are used. Pulse condition: Potentiation (200 ns, 3.7 V, n = 512), Depression (200 ns, -2.75 V, n = 512), Vread (200 ns, -2.75 V, n = 512). (1 V, 200 ns).
....86

Figure 2-25: Statistical examination of the retention measurement for each picture pixel in Fig. 2-4d. Pixel values in 256-level grayscale images are uniformly distributed into 13 groups (e.g. 0-19, 20-39, etc.). For each time step and alloy device (a) Ag, (b) Ag-Ni, and (c) AgCu, the average values of the pixel groups were computed. The grayscale picture values mapped from the conductance values are shown on the left y-axis of the images, while the actual conductance of the device is shown on the right y-axis.
....87

Figure 2-26: The I-V characteristics of an Ag-Cu memristor with varying compliance currents. With a compliance current of less than 100 A, the device can be programmed indefinitely. 88

Figure 2-27: S SPICE simulation configuration. (a) IV Ag-Cu behavioral model parameters compared to measured device performance (b) An Ag-Cu model with a reduced OFF state

conductivity. (c) Schematics of the 1/2V and 1/3V write schemes, as well as the ground read technique, employed in array operation and SPICE simulation.
....89

Figure 2-28: For the write process, SPICE simulation is used. (a) Voltage delivery maps for Ag-Cu arrays with various line resistances and biasing techniques. A 3V bias is given to the source terminals from the left and bottom terminals. Reduced voltage biases applied across device connections in the 32 × 32 array are indicated by color gradients. (b) Simulation of the write process to send 3V to the worst-case cells in arrays, i.e. the cell closest to the source, such as the top right cell in Figure 1. (a). Arrays of various sizes were evaluated. The green and blue lines represent extracted voltage biases at the half-selected cell nearest to the source. When the voltage bias of a half-selected cell exceeds the writing threshold, write disturb occurs (e.g. 3V). The simulations were run for three alternative device conditions: the original Ag-Cu model, the modified Ag-Cu model, and the Ag-Cu model that only employs the lower 10% of conductance ranges. Improved scalability for the write process is obtained by using a better device model or decreasing the conductance ranges of Ag-Cu memristors...90

Figure 2-29: For the read process, SPICE simulation is used. (a) Average read error in 32 × 32 arrays. The read error is defined as the difference between the inferred device conductance and the actual device conductance based on the TIA readout. The simulation findings indicated that when cells got further away from the source, their reading mistakes increased, which was caused by line resistance. Three alternative device conditions, similar to those utilized in write simulation, were also investigated here. Limiting device conductance was discovered to be beneficial in lowering the influence of line resistance. (b) Read error simulation in a 1 × 32 array with the same wire resistance as in (a) to rule out the effect of sneak pathways. The simulations produced comparable results (a). (c) The read error differences between (a) and (b)

were calculated and revealed extremely modest changes, demonstrating that line resistance is the most important element during the read process using the ground read scheme. (d) Simulation of read errors throughout the read process in various array sizes using an Ag-Cu device model with a limited (lower) conductance range.

....92

Figure 2-30: SPICE simulation is used in computing. (a) Simulated multiply-accumulate (MAC) error in 32 32 arrays. The computed MAC value of input voltage vectors and the conductance of programmed devices (values derived from TIAs) are compared to the array's actual MAC output. (b) Simulation of MAC error in various array sizes and device conditions.

....94

Figure 3-1: Integration technologies for sensor-computing systems for use in edge computing applications. a, Schematic representation of 3D heterogeneous integration for a sensor-computing system. A single material system contains physical wires (in yellow) that connect different device layers such as sensors, processors, and memory to one another (for example, silicon). Vertical interlayer vias are used to physically configure the three-dimensional structure. A limitation of functionality and chip stackability can be attributed to the hardwired connections. b, Schematics of stackable hetero-integrated chips with chip-to-chip communication for sensor-computing systems with chip-to-chip communication. This demonstrates the use of light communication between chips, which enables a high degree of freedom for hardware-based reconfigurability at the sensor layer and processor layer, among other places. Optoelectronic devices such as light emitting diodes (LEDs) and photodiodes are used to convey light information between the layers. Each layer is physically in contact with the other. Depending on the application, any functional module from a library of functional modules can be selected and assembled into a heterogeneously integrated chip stack, which is

then tested. The first layer of the chip stack is dedicated to the processing of sensory input, and the input can be processed in parallel over the 3D hetero-integrated chip stack as the chip stack is built. Pre-programmed neuromorphic processing cores in the stacking hetero-integrated circuits enable them to implement various artificial intelligence applications for edge computing, such as letter recognition and objection detection. 108

Figure 3-2: Components of stackable hetero-integrated neuromorphic chips. a, an optical image of a light sensing layer (eye layer) with light input from the bottom. Patterned photomasks and a laser diode with a confocal setup are used to create three different letter patterns: M, I, and T. The eye layer receives the letter images. Scale bar: 1 mm. A photodiode measures the intensity of light passing through a transparent silicon oxide membrane. b, optical photo of a processor layer and an eye layer. To communicate, two chips are physically in contact. The eye layer's light input can be passed on to the next layer. Each layer has a photodiode/LED stack that allows them to receive and transmit light data. I-V characteristics of LED and photodiode devices are presented SEM images with light OFF and ON (~ 0.15 mW/mm2) for the photodiode. At 0.1 V reverse bias, photodiode response to LED light in chip-to-chip communication as a function of LED light intensity. The chip-to-chip communication is depicted in the inset. c, a diagram of a hetero-integrated chip. Illustration and SEM images of optoelectronic device stack (photodiode/LED, scale bar: 100 μm), neuromorphic computing core (Ag-Cu alloy-based Si memristor crossbar array, scale bar: 100 μm), and sideview of chips (scale bar: 1 μm) are presented. We used deep reactive ion etching to align an array of photodiode/LED stacks with backside holes for an optoelectronic device stack (DRIE). We made 32 × 32 memristor crossbar arrays for the neuromorphic computing core, as shown in the schematic and SEM image. On a 128 × 128 pixel image, three different types of 3 x 3 kernel operations (edge detection, sharpen, and soften) are performed.

....112

Figure 3-3: Hetero-integrated neuromorphic chips that are replaceable and stackable as well as their robust kernel operations. a, Replaceable neuromorphic chips for kernel operations. Patterned images produced by photodiodes in kernel layers with a 0.1 V reverse bias are shown (Top row). The stacking and replacement of hetero-integrated chips is depicted in block diagrams. Different patterned images have had kernel operations performed on them by replacing a kernel layer. Each kernel operation's current sums are shown in the graph. The maximum and minimum values of the current sums are indicated by error bars. b, The multi-layer neuromorphic chip stack is depicted in a block diagram. Pre-programmed memristor crossbar arrays process the input letter images, which are shared across three kernel layers. The current sums are the result of kernel operations on the letters in the input images below (Bottom row). Kernel operations were performed in parallel by stacking three different kernel layers. The maximum and minimum values of the current sums are indicated by error bars. 117

Figure 3-4: The use of stackable neuromorphic chips in a noisy environment: the insertion of a denoising functional layer in the midst of the chaos. a, the addition of Gaussian noise (0.5) to images from the eye layer results in the generation of corrupted letter images. b, block diagram of the letter recognition task that was performed on a corrupted 'T' letter image. On the corrupted letter image of the letter T, we performed three different kernel operations. The current sums are calculated for each kernel individually. The difference between the current sum and the previous sum is indicated by arrows. Because of the noise in the letter "T," the outputs of the current sum from the "I" kernel and the "T" kernel show only a marginal difference, indicating that letter recognition is difficult. c, Denoising functional layer neural network architecture with denoising autoencoder (25 × 5 × 25 neurons) and denoising autoencoder. The denoising layer, which includes memristor crossbar arrays, is added after the letter images have been denoised in step d, as shown in the black diagram. Example of letter recognition task performed on denoised letter image of the letter T, as described in the block

diagram shown in Figure 3b. Following the denoising process, the current sum from the ‘T' kernel yields a significantly higher value than the current sum from the ‘I' kernel. This implies that, following the completion of the denoising process, the task of letter recognition will be more successful.

....120

Figure 3-5: Schematic illustration of three-dimensionally stacked neuromorphic chips for multi-modal sensor fusion. a, Photodetectors and strain sensors can provide several sensory inputs. The depth and width of networks can be changed by simply stacking or changing chips. Furthermore, the size and kind of memristor crossbar array can be changed based on the size and function of the sensor array. b, an example of a basic neural network architecture for sensor fusion. The top left (red) neurons represent early visual processing. Bottom left (blue) neurons represent early haptic processing. Right (purple) neurons exhibit multi-modal sensor fusion with sparse connections.

....124

Figure 3-6: Confocal optical measurement setup for chip-to-chip communication. a, An optical image of a confocal imaging setup. The beam path of light input from a 635 nm laser diode is indicated by a yellow arrow (light source). Hetero-integrated chips are mounted to adapters as shown in the photo in the right-hand side. Purple lines represent the chip for the eye layer, whereas green lines represent the chip for the process layer. Micro-manipulators fine-tune two chips. When more than two chips are communicated, they are replaced and measured sequentially. b, A confocal imaging setup diagram. The light is formed into three letters ('M,' 'I,' and 'T') using a patterned photomask. PM stands for patterned mask; CL stands for free-space collimator lens; BS stands for beam splitter; OL stands for objective lens; and CHIP stands for hetero-integrated chip. To adjust the position of light pattern input and chips,

a confocal setup was employed. We were able to accomplish minimal optical crosstalk between chips by carefully tweaking micro-manipulators and shifting the positions of chip mount adapters with proximity.

....125

Figure 3-7: I-V characteristics of 6 × 6 photodiode array and 6 × 6 LED array in an eye layer. a, I-V curves of photodiodes. Orange curves show the response of photodiodes to the light (0.1 mW/mm^2) while blue curves show the I-V characteristics with no light. b, I-V curves of LED devices in 6 × 6 array. Optoelectronic devices (photodiodes and LEDs) in other layers (e.g. denoising layer and classification layer) show the similar levels of performance. For the demonstration, only 25 devices were used for 5 × 5 pixels images.

....126

Figure 3-8: Schematic illustrations of photodiodes/LEDs array fabrication for a stackable heterogeneously integrated neuromorphic chip. a, Silicon on insulator (SOI) wafer with epitaxially grown Si was prepared. b, SiO_2 layer was deposited on the SOI wafer by plasma enhanced chemical vapor deposition (PECVD). The SiO2 layer was patterned using a photolithography tool and wet etch (BOE 7:1). c, epitaxial Si was patterned and etched by wet etch (KOH solution). d, After epitaxial lift-off (ELO) process, LED/PD stack was transferred to the SiO_2 membrane treated with APTES and polyimide. The detail of LED/PD stack structure is shown in red box. e, LED was patterned by wet etch (Cr etchant, HCl + H_3PO_4). f, PD was patterned by wet etch. g, Ti/Pt/Au was deposited on the p-doped side. h, Ni/Ge/Au was deposited on the n-doped side. The final structure of LED/PD stack is shown in navy box.

....127

Figure 3-9: Field-programmable gate array (FPGA) system for memristor crossbar array measurement. a, Schematic of FPGA system. Parallel programming and reading were

controlled by software on computer. b, Image of the FPGA system. The system consists of a core board, a 64 channels-TIA board, two 64 channels-DAC boards, and a connector board with DUT connectors. DUT connectors were connected to the probe card for programming and reading. We used 5K ohm resistors to initiate the crossbar array. c, Image of the crossbar array mounted to the probe card..128

Figure 3-10: Schematic of 5 x 5 kernel operation on 5 x 5-pixel images. To perform a recognition task of three 5 × 5 letter patterns, we implemented three 5 × 5 kernels into Si memristor crossbar arrays. Fig. 3-10 presents three 5 × 5 kernels represented in software (Top) and programmed in memristor crossbar arrays (Bottom), respectively.

....129

Figure 3-11: Image preprocessing in the layer-to-layer light communication. Patterned Images obtained by photodiode arrays with 0.1 V reverse bias both in eye layer ('A' images) and in classification layer ('B' images). Due to the light diffraction, there are some noises around the letter patterns (Top, 'A' images). These noises are filtered when the light information is processed by a photodiode array in classification layer (Bottom, 'B' images).

....130

Figure 3-12: Flow chart of the optical and electrical measurements in chips. Black boxes indicate the interactions between a laser diode and photodiodes. Blue boxes indicate the interactions between LED/PD stacks. Red boxes indicate the interactions between a memristor crossbar and the FPGA system. In the process of a crossbar, the conductance programming has been performed using a closed-loop scheme before Multiply-Accumulate (MAC) operations.

....131

Figure 3-13: Software simulation results of denoising autoencoder to noisy input as a function of Gaussian noise level. Denoising autoencoder has been implemented to improve the pattern recognition. Gaussian noise has been added to ground truth with different noise levels. Noisy inputs are inserted to the 25 × 5 × 25 denoising autoencoder. Denoised output show the simulation result of noisy input. Learning parameters are as follows. Adam optimizer has been adopted with 0.0003 of learning rate and 100 of epochs. Loss has been calculated by the mean squared error (MSE) and weights have been updated by back-propagation. Different level of gaussian noise has been added to the original data. 132

Figure 3-14: Denoising autoencoder training loss. Training loss from denoising autoencoder. Adam optimizer has been adopted with 0.0003 of learning rate. Loss has been calculated by the mean squared error (MSE) and weights have been updated by back-propagation. Different level of gaussian noise has been added to the original data. Total number of dataset is 180 and 20 % of dataset is used for validation

....133

Figure 3-15: Distribution (cumulative distribution function) of Multiply-accumulate (MAC) operations and current sums of denoising autoencoder in software and memristor crossbar arrays. As described in Fig 3-4c, the 25 × 5 × 25 denoising autoencoder has been implemented into Ag-Cu alloy Si-based memristor crossbar arrays. No bias value has been used in the neural network. For both fully-connected layers, differential pairs of memristors are used to represent positive and negative MAC values. A rectified linear unit (ReLu) has been used to MAC values in the first fully-connected layer. Then, MAC values are normalized and converted to voltage pulses before the second fully-connected layer. Histograms show the MAC values performed by software (floating number) and Ag-Cu alloy Si-based memristor crossbar arrays (current

sums averaged by the number of maximum number pulses = 100, unit - μA). Left shows the MAC values in the first fully-connected layer and right shows the MAC values in the second fully-connected layer. The output current from the crossbar is calculated by memristor differential pairs. The negative MAC values are zeroed for the next kernel layer 134

List of Tables

Table 1-1: Features of manufacturers of hardware systems..31

Table 1-2: Hardware types of AI-driven data processing 32

Table 2-1: Summary of KMC simulation parameters 95

Table 2-2: Summary of simulation parameters 96

Chapter 1

Introduction to Neuromorphic Computing

Neuromorphic computing is the term used to describe computers or components that are inspired by the brain and that mimic an artificial neural network with densely connected parallel neurons and synapses. In order to realize neuromorphic computing, a physical computing platform is required. Designing the appropriate device, circuits, and architecture for neuromorphic computing hardware is essential for putting neural networks in embedded systems into widespread use. Deep learning and artificial intelligence (AI) advances have led to great attention being given to memristor-based AI hardware accelerator, since it can conduct a matrix multiplication in the simplest form to help with AI workloads. Computing systems have a long history, and I will highlight one kind of memory that is called memristor and is also known as resistive random access memory (RRAM) in this thesis. With the alloying technique I developed, I enhanced RRAM, which would be a critical component of neuromorphic computing technology.

1.1 Computing Systems

Once deep neural network (DNN) was implemented, artificial intelligence (AI) began performing better than human beings. Despite DNN-based AI models outperforming human-level recognition and classification, the gap between computer systems and human brain is still wide when it comes to their functionality and energy efficiency. AI algorithms nowadays are

implemented on traditional computer devices such as CPUs and GPUs (GPU). To accelerate the speed of computation for AI tasks, field-programmable gate array (FPGA) and application specific integrated circuit (ASIC) have been developed as well as described in Table 1-1. These computer systems, however, are built on silicon technologies that have beyond the fundamental limits of quantum mechanics. There is thus only a small amount of space to enhance the overall system performance and energy efficiency in traditional computer systems. Performance of the computer system is mostly dependent on the data bandwidth between processor units and memory units in the conventional von Neumann computing design, in which the processor units and memory units are physically separate. This is because the processor units handle data received from memory units and the data handled is then stored in memory units. Because of this, current computer hardware research is mostly interested in addressing bottlenecks at the processor-memory level, as well as artificial intelligence algorithms, peripheral circuits, and computing architecture integrated together. Advanced computational hardware that seeks to achieve extremely efficient operations in neural networks is referred to as AI hardware accelerator. Non-von Neumann computing architecture such as in-memory computing allows for large and concurrent data processing. To use AI hardware in this way, the hardware is capable of handling computer operations for neural networks with a speed and efficiency beyond that of standard computers. Table 1-2 summarizes strengths and weaknesses of hardware used for AI data processing. Unlike CPU, AI hardware is designed to improve the efficiency of computer operations for AI. Due to the parallel data processing required by AI algorithms, GPUs have found widespread usage as AI platform hardware. Until now, GPU has contributed the most to the AI implementation in real-world applications. FPGA is another kind of AI hardware that offers higher degrees of reconfigurability via hardware reconfigurability. In general, FPGA consumes less power but costs more than GPU. Because FPGA is flexible in terms of design and efficient when it comes to energy use, it is an excellent option for people

that need these characteristics. In the meantime, ASIC contains many types of chips such as tensor processing unit (TPU) and neural processing unit (NPU) to implement AI models. Application-specific ASIC design offers rapid data processing and low device footprint. Although ASIC can perform complicated computer operations with the optimized architecture including an arithmetic unit like CPU, it is not possible to reconfigure the architecture once it is designed and it takes long time to manufacture ASIC chips. To solve these problems, neuromorphic computing hardware, which mimics biological neural networks, has been proposed as the next stage of ASIC for accelerating AI workloads and increasing energy efficiency for computer processes. The aim of this thesis is to investigate the use of neuromorphic computer hardware, considered the next generation AI hardware, as it relates to this work.

Hardware type	GPU	FPGA	ASIC	Neuromorphic computing
Feature	• Parallel data processing • Extremely High energy consumption	• Reconfigurable hardware architecture • High energy consumption	• Efficient process for AI • Low energy consumption	• Data processing and memory take place on the same place such as in-memory computing • Extremely low energy consumption
Manufacturer	• Nvidia	• IBM • Microsoft	• Qualcomm • Apple • Samsung	• IBM • Intel

Table 1-1: Features and manufacturers of hardware systems.

Hardware type	Strengths	Weaknesses
CPU	• Flexible software programming	• High power consumption • Low computational power (poor FLOPS)
CPU + GPU	• Tunable hardware usage depending on applications	• Computational power is highly dependent on the number of GPU and their performance
GPU	• Massive training • High data processing parallelism	• High power consumption for inference tasks • High power consumption
FPGA	• Hardware reconfigurability • AI modules can be embedded • Chip design period is shorter than ASIC	• Not proper to low power applications • High cost
ASIC	• Low power consumption for inference tasks • Very efficient in specific applications	• Not reconfigurable after manufacture • Difficult to design ASIC chip for training process • High cost • Long chip design period

Table 1-2: Hardware types for AI-driven data processing.

1.2 Memristor: resistive random access memory

Processing information in human brain has an entirely different paradigm from conventional von-Neumann computing architecture. While von-Neumann computing architecture is optimized for accurate and precise information processing, human brain is controlled by spatial and temporal events called ‘spike’. Neurons can accumulate spikes through synapses and convert them to membrane potential. If many spikes arrive at the neurons within a short time, a membrane potential reaches threshold, and a neuron outputs a spike to other neurons. The spike events are different from digital information in von-Neumann computing architecture in the way that spike only contains information regarding time and the origin of spike. To mimic these biological behaviors like neurons and synapses, many electronic and ionic devices have been invented. The most widely known device is ‘memristor’ which is also known as resistive random access memory (RRAM). For clarity, I will use the term RRAM for memristors from now on. Once it was found that RRAM can represent synapses in neural networks by modulating resistance values with a sandwiched two terminal device, they have been substantially studied to implement neural networks primarily on the simplest structure nano-device ‘RRAM’. In the first generation of neuromorphic computing hardware, synaptic devices were constructed by complementary metal oxide semiconductor (CMOS) transistors. Thus, synaptic strength information between neurons must be stored and processed separately. However, research communities longed for neuromorphic computing hardware that can exhibit more biological behaviors by having both memory function and processing function in a single device. Flash memory, RRAM, phase-change random access memory (PRAM), and magnetic random access memory (MRAM) were suggested for the next generation neuromorphic device. Among them, RRAM is a strong candidate for the next generation neuromorphic computing since it can have multi-values like biological synapses. Compared to flash memory type neuromorphic device, RRAM devices show high switching

speed, high device density, low operating voltage, low energy consumption, and CMOS compatibilities for micro- and nano- fabrication.

Since the first resistance switching is reported in Au/SiO2/Al RRAM devices, many two terminal RRAM devices have been suggested [33]. The switching of RRAM can be achieved by the transformation of the functional materials. In typical, it can have two states ‘OFF’ and ‘ON’ determined by the values of resistance. Current-voltage (I-V) measurement curves are presented depending on a switching medium and an active metal. One is bipolar RRAM and the other is unipolar RRAM as shown in Figure 1-1. When the voltage is scanned during the I-V curve test, the device has an electric field to it such that there is a transition from high resistance to low resistance in the RESET process or from low resistance to high resistance in the SET process. Since the majority of RRAM devices can have analog values not just digital (‘OFF’ and ‘ON’), researchers have attempted to enable this RRAM’s analog memory devices to stimulate biological synaptic functions [34], [35]. In particular, when it comes to the advanced artificial intelligence technique which is deep learning, a crossbar array of RRAMs shows the great potential for accelerating a multiply-accumulate (MAC) operation by simply sending a train of voltages and receiving a current. For reliable MAC operations for deep learning, a great number of RRAM devices is required in a RRAM-based crossbar array. In this thesis, electrochemical metallization (ECM), one type of RRAM device, will be mainly discussed since I improved the property of ECM-based RRAM using metallurgy. I will discuss the working principle of ECM-based RRAM and its challenges for the rest of this chapter.

Unlike other types of RRAM devices, the active electrode of ECM-based RRAM devices is made of a metal that is readily dissolved in and transported through the switching medium layer, which distinguishes them from other types of RRAM devices. When referring

to the switching medium layer, it should be used in a wide sense to encompass materials with both low and high ionic conductivity. This fundamental definition of ECM-based RRAM devices may be fulfilled by a broad range of different materials, but the active metal in most devices studied so far has been Ag or Cu. The ECM-based RRAM serves primarily to illustrate the concept of devices with highly mobile active electrodes such as Ag, Cu, or Ni [36], [37]. Single Ag filaments were initially seen using an optical microscope and the Ag/Ag-As2S3/Au device [38]. Transmission electron microscopy was used to investigate the conducting channels in the Pt/SiO_2/Ag memristor devices under investigation [36]. There is no conductive filaments generated in the SiO_2 layer between the Ag electrode and the Pt electrode due to the novel planar device's manufacturing method. Ag+ is produced when a positive voltage is applied to the Ag electrode, resulting in oxidation. When an electric field is applied, Ag+ flows toward the Pt negative electrode, which also serves as a positive electrode. This process is called 'SET' and occurs under the condition that the active electrode is supplied with a sufficiently high bias voltage V. The SET process is divided into the following steps, which are shown using Ag as an example of the active electrode and Pt as the catalyst: First, anodic dissolution of Ag in the solid electrolyte according to the reaction where Ag+ represents the metal cations. Second, as a consequence of the high electric field, the Ag+ cations move across the solid electrolyte. Third, the amount of Ag+ cations in the solid electrolyte decreases and is removed. When Ag+ comes into contact with the surface of the Pt electrode, the reduction process begins, resulting in the conversion of Ag+ to Ag. The switching middle layer of RRAM devices acquires conductive Ag filaments throughout the Ag accumulation process and this results in a low resistance state. When the metal filament has grown enough to make contact with the opposing Ag electrode, an ECM-based RRAM device is activated. Unless a substantial voltage of opposite polarity is applied, the device will stay in the ON state until the electrochemical breakdown of the metal filament forces it to be reset to the OFF state. During the first phase of

RESET, the electronic current passes through the metallic filament while an electrochemical current dissolves it. The ‘RESET’ finalizes after the conductive filament is broken by the current and the Pt electrode receives a positive voltage, resulting in a high resistance condition for the RRAM device. With a careful design of materials selection, RRAM devices could be controlled with tunable SET and RESET processes. Although ECM-based RRAM can achieve many excellent one of switching properties, none of RRAM devices can meet device requirements simultaneously such as temporal variation, spatial variation, switching endurance, data retention, multi-level capability, and asymmetric nonlinearity for reliable neuromorphic computing as depicted in Figure 1-2. Here, I also would like to cover device performance requirements and why each of them is significant.

First, the difficulty of mass manufacturing RRAM devices is exacerbated by the lack of uniformity in various device characteristics. Several factors, such as switching voltages and high- and low-resistance states, may be very changeable. Resistance switching is subject to temporal and spatial variations. The variation of electrical characteristics of RRAM devices is considered the most challenging factor for real-life applications. It is possible that the variation of RRAM device can be used for security application like a random number generator, but it is required to have less variation in neuromorphic computing which I mainly discuss in this thesis. The random nature of conductive filament formation and rupture is the main source of these variations. The changing nature of RRAM devices between cycles and devices is a major data storage bottleneck. Oxygen vacancy defects, which occur in the conductive filaments from the stochastic behavior of creation and destruction during the switching event, influence impact variability from cycle to cycle. Due to the conductive filament's random nature, it is impossible to predict and precisely regulate its shape. There is also non-uniformity between RRAM devices, which results in the degradation of the window margin between two states and lowers memory performance. According to the researchers, manufacturing process inconsistencies

like as variations in switching layer thickness, etching damage, and electrode surface roughness explain this unpredictability.

Second, endurance is a property of resistant random access memory that is characterized by making transitions between a high resistance state (HRS) and a low resistance state (LRS). As a result, the test for endurance is comprised of finding out maximal number of SET and RESET cycles. This often uses the figure of merit, resistance values for HRS and LRS versus the number of switching cycles. Each resistive state change has the potential to permanently damage the RRAM and degrade its performance. Again, endurance is defined by how many times an RRAM device can be switched between the HRS and LRS states but retain a detectable ratio between the two states. The detectable ratio for two states is commonly suggested above 5 but it is not standardized. The endurance test is performed based on the maximum number of successful set/reset cycles until the HRS and LRS become the value the researchers set tentatively (5 is commonly used). In a resistive switching device, RRAM's endurance properties are tested via a series of I-V sweep measurements through the extraction of HRS and LRS when a read voltage is applied. This technique is dependable since it allows for device switching at the right time throughout each cycle. For a more realistic measurement that is close to the device operation, pulse input can be used instead of I-V sweep.

Third, data retention of an RRAM device is governed by the stability of the LRS and HRS after set and reset transitions. The duration of the device's state after a set/reset operation influences the amount of data that an RRAM device can keep after a set/reset operation. Therefore, the figure of merit for data retention is current versus time for each resistance state. The resistance states can be measured at multi-level values with varying compliance current or number of voltage input. To assess the state retention, one must use a low read voltage and apply continuous voltage stress over time, and then analyze the current versus time curves for both LRS and HRS. Because of the dispersive nature of atomic rearrangements, it is difficult

to achieve a lengthy retention period in RRAM as a consequence of set voltage. In HRS, however, retention time is irrelevant since the device is often in its original state, and RRAM will maintain this state even in the absence of bias. A greater compliance current results in a stronger conducting filament that is more stable over time in an RRAM with conductive filament switching mechanisms, while a lower compliance current results in a weaker conducting filament that is less stable over time. One of the main challenges for RRAM devices is securing the excellent data retention for weakly formed conducting filament.

Fourth, RRAM devices are capable of achieving multilayer resistive states, which enables them to provide the advantages of low-cost, high-density nonvolatile data storage systems. At the present, significant research is being conducted in the area of RRAM with the aim of reducing memory arrays while improving their structural density. Previously, RRAM storage density was improved by shrinking the device; however, the intricacy of the experimental methods made this impossible to apply successfully. Other possible are the use of three-dimensional (3D) crossbar structures and vertical RRAM. However, each of these architectural forms require complicated manufacturing processes, which is undesirable. A far more straightforward method of increasing storage density in RRAM devices is to use multilevel property, which can make the storage of more bits per device without scalability. This layered state is one of the most intriguing features of RRAM, as it has the potential to substantially improve the device density of the memory system. This means we can accomplish more than simply ON and OFF states, and we can do it without increasing the device's size. Precision control of the resistance of the different resistance levels in RRAM is needed for successful multilevel state functioning; otherwise, the device would exhibit variability and reliability issues due to the random nature of conductive filament.

Fifth, non-linearity of resistance value update is substantially critical in RRAM devices. It is desirable that potentiation process (resistance value getting smaller) and depression

process (resistance value getting larger) are symmetric and linear to identical voltage pulses with the opposite polarity. Voltage pulses are required for neural network training because they potentiate or depress each synapse, independent of its initial resistive state. A high linearity of the RRAM characteristics is required, where a fixed applied pulse results in a known amount of potentiation/depression via an additive or multiplicative term, and (ii) a high symmetry of the update is required, where similar update characteristics are obtained for potentiation or depression under positive or negative bias, respectively. Because RRAM devices have distinct set and reset transitions, symmetric features are typically difficult to accomplish with this technology. Linear weight update is needed for synaptic applications; however, real-world RRAM devices often exhibit non-linear behavior. RRAM technologies often exhibit a considerable degree of non-linearity as compared to other memory technologies in terms of both potentiation and depression properties. The asymmetric non-linearity factor is one method for determining the non-linearity of RRAM devices (ANL). It will be utilized in the book's Chapter 2. The high ANL values for Ag-based traditional RRAM devices are readily discovered in Chapter 2, where they are addressed in depth. My goal is to propose an alloying technique for improving device non-linearity, which will have a significant effect on the training process of neural networks on RRAM.

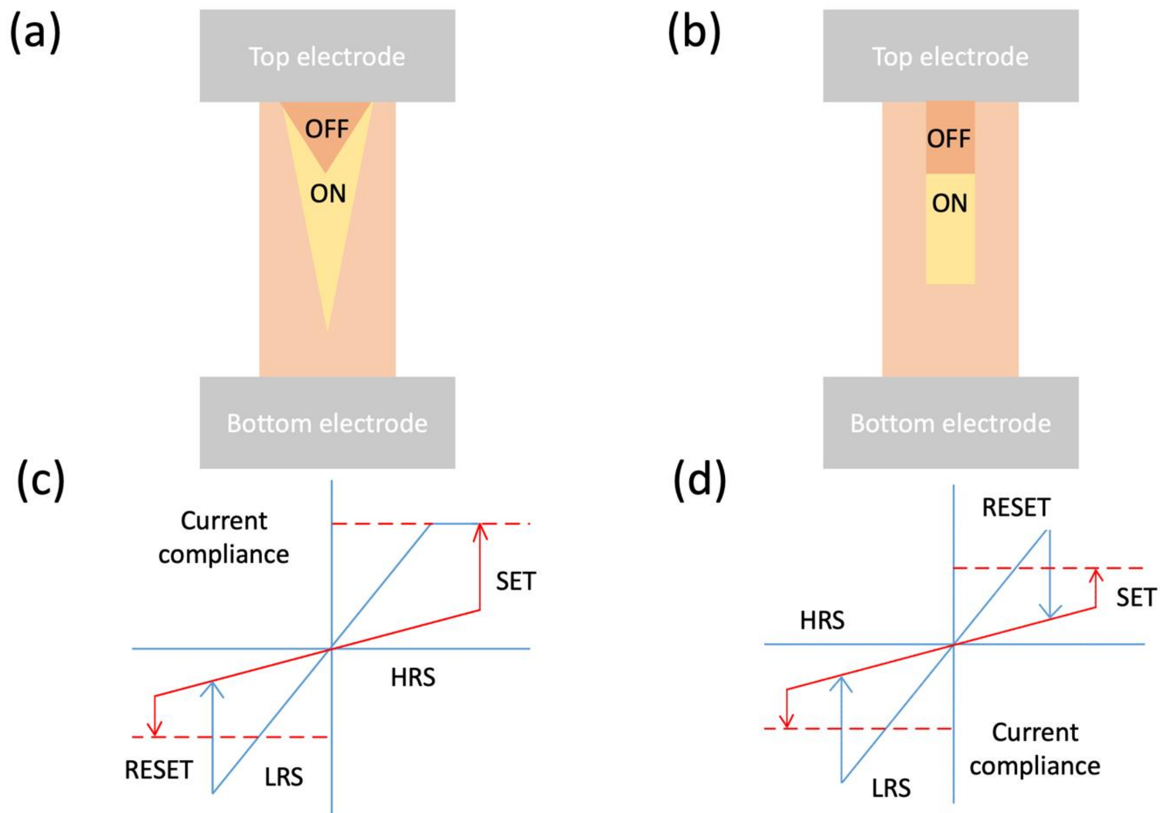

Figure 1-1: Types of bipolar and unipolar two terminal RRAM devices. a, bipolar RRAM device schematic. Depending on the size of filament inside of switching medium, the resistance value of RRAM device is determined. (Orange: OFF, yellow: ON). b, unipolar RRAM device schematic. c, I-V characteristics of bipolar RRAM device. HRS and LRS stand for high resistance value and low resistance value. SET process forms or strengthens a filament and RESET process disrupts or weakens the filament. d, I-V characteristics of unipolar RRAM device.

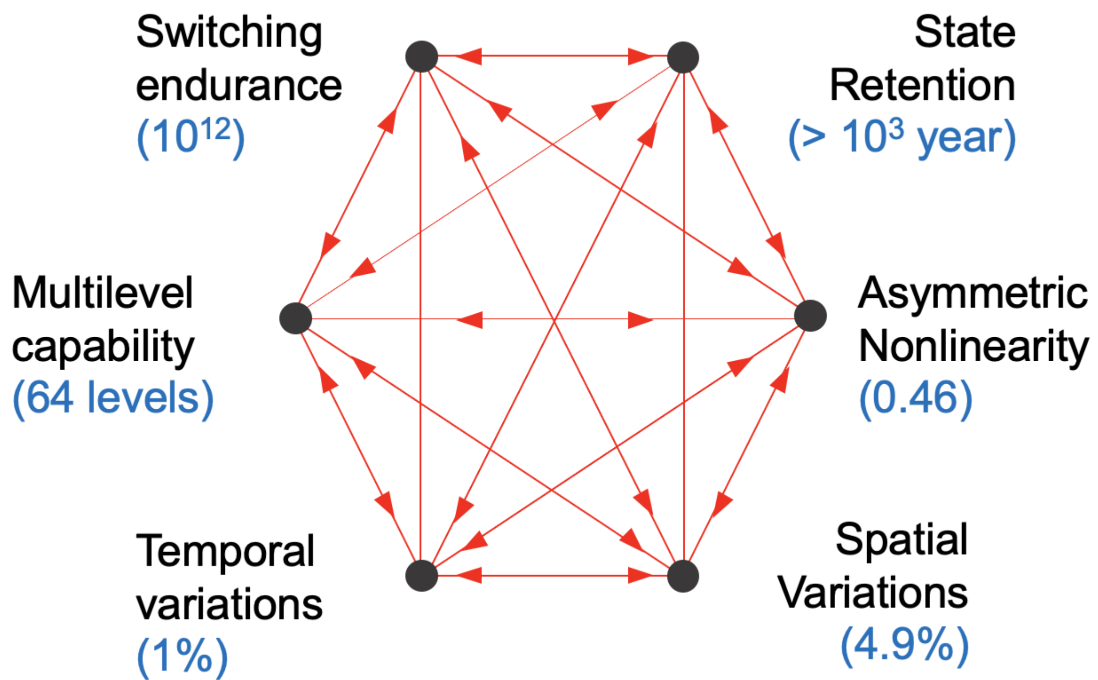

Figure 1-2: Challenges in memristor device performance matrices. Six device matrices are presented. Each number in round brackets represents good device performance from ref [1]–[6]. It is worth noting that none of RRAM devices developed can meet all six requirements at the same time.

1.3 Research Goals and Thesis Organization

The purpose of this thesis is to investigate and build a reliable RRAM device as well as crossbar arrays for reliable neuromorphic computing applications. When copper (Cu) and silver (Ag) are alloyed, silicon-based RRAM devices show superior switching characteristics such as linearity, symmetry, and uniformity when compared to silicon-based RRAM devices with silver active metal. Additionally, this thesis describes the development of heterogeneously integrated neuromorphic devices for reconfigurable neuromorphic computer hardware. We can physically reassemble the chips using chip-to-chip optical connectivity for a variety of applications. It is anticipated that hetero-integrated neuromorphic circuits would enable easy reconfiguration of various sensory devices and multi-functional computers. The following is the substance of this thesis. To begin, Chapter 2 discusses the device solution, which consists of Ag-Cu alloy silicon-based RRAM devices. The alloying metal method enables us to build well-balanced RRAM devices in terms of device performance matrices suggested in Fig. 1-2. Second, Chapter 3 adds to the hardware-level reconfigurability that heterogeneously integrated stackable chips provide in order to adapt to a variety of situations (i.e. multi-modality like light, touch, temperature, and high noisy environment). At the conclusion of each chapter, the work's contributions will be addressed.

Chapter 2

Alloying Conducting Channels for Reliable Neuromorphic Computing

This chapter covers Ag-Cu alloyed memristor device and crossbar arrays. This chapter has been published with minor changes as 'Alloying conducting channels for reliable neuromorphic computing', Nature Nanotechnology, 15, 574-579 (2020) [39]. Hanwool Yeon[†], Peng Lin[†], Chanyeol Choi[†], Scott Tan, Yongmo Park, Doyoon Lee, Jaeyong Lee, Feng Xu, Bin Gao, Huaqiang Wu, He Qian, Yifan Nie, Seyoung Kim, and Jeehwan Kim. ([†]co-first authors) The abstract of the paper is shown below.

For emerging neuromorphic computing applications, a memristor has been suggested as an artificial synapse [40]–[42]. With regards to device conductance, alterations to the weight values should be distinct and homogeneous while training a neural network in memristor arrays. Because metal ions are highly mobile in the Si switching medium, good analog switching performance has been demonstrated [1], [6], [43] by a primarily-silicon-based electrochemical metallization (ECM) memory [44], [45]. Alternatively, switching variability is caused by the high stochasticity of ion transport. We present our controllable, stable Si-based memristor that incorporates alloyed conduction channels for large-scale crossbar array

implementation. The principal mobile metal in the conduction channel is conventional silver (Ag), which is alloyed with silicidable copper (Cu) to stabilize switching. Cu effectively governs Ag movement at an appropriate alloying ratio. This results in three major improvements: first, there is significantly greater uniformity in spatial and temporal switching; second, across many levels of conductance there is stable data retention; and, third, in analogue conductance states the programming symmetry is ameliorated. With our alloyed memristor, we are able to produce large crossbar arrays featuring precise analogue programming and high device yield. These discoveries represent an essential step in the progression past von Neumann computing.

2.1 Introduction

Device performance matrices must meet a variety of requirements in order to run large-scale memristor arrays. Metal-oxide-based memristors have previously shown promise for image classification and signal processing in the context of large-scale arrays [3], [43], [46]–[48]. In the majority of these implementations, however, an extra transistor, acting as the selection device, was required to connect each memristor in a series and to control the device's analog switching properties [3], [49], [50]. Although the gate voltage of the choosing transistor can modulate a reasonable analogue weight update [3], by adding a transistor, the memristor's stackability and scalability are limited while the design complexity and peripheral overheads increase significantly. However, a memristor array without a transistor has significant limitations because it loses its ability for fine-grained conductance tuning and it becomes increasingly sensitive to switching variations. [51]. Even without transistors, ECM memory [52], [53] exhibits excellent multi-level switching properties towards a more linear and symmetric weight update. For above cases, Ag is commonly used as an active metal because

of its high mobility [44], [45]. However, because of the weakly formed conduction channels in low conductance states, such mobility causes weight variation and lowers long-term reliability [54]. Although these devices can modulate precise analogue states, they cannot be maintained for computing purposes. As a result, we need to develop a novel memristor that retains stability at all conductance levels while simultaneously demonstrating great analogue tunability.

2.2 Contributions and Methods

In this portion of my thesis, I will outline my first independent project. In order to solve the problems of stochastic switching uniformity and tunable weights across a wide range of conductance levels, we pioneered a new hardware approach that tackles both of these issues at the same time in a single device. Prior to developing our ECM devices, we first established a metallurgical approach. This was accomplished via the formation of and tailoring of conduction channels inside a switching medium in the order to discover a combination of metals and switching medium that would solve the problems of uniformity and tunable weights. We selected Si as our switching medium because of the large number of studies that have been done on the interactions between silicon and metals. On the basis of this information, we investigated how various combinations of metals would react with one another and with the Si switching media. We concluded that Ag and Cu were the most thermodynamically compatible combination, and this determination was confirmed when we built our memristors. Because of a combination of three factors, the conduction channels produced by Ag and Cu in Si exhibited substantially superior memristor performance. First, the conduction channels demonstrated uniform switching gradients. Second, when multilevel conductance states were tested, the channels showed reliable retention. Third, the analog conductance updates showed enhanced

linearity and/or symmetry. Upon observing this promising performance, we fabricated 32 x 32 transistor-less alloyed memristor crossbar arrays to test our hardware. We had a 100% yield, were able to program and operate the memristor reliably, and performed interference tasks thanks to the significantly improved data retention. This work will facilitate the field’s move towards efficient analogue computing without the need for transistors in the memristor arrays.

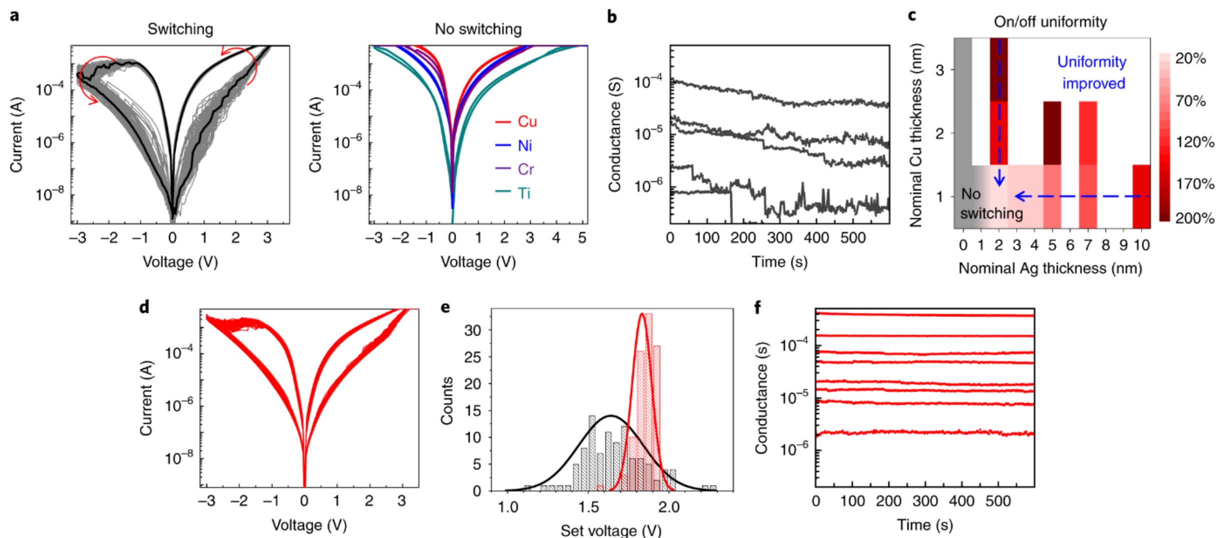

Figure 2-1: Figures show the d.c. switching performance of a Si memristor using an Ag–Cu alloy. a, Following the forming process, typical current–voltage curves of Ag (left) and silicidable metals (Cu, Ni, Ti, and Cr) (right). b, Retention characteristics of Ag devices with several conductance levels (measured at 0.5 V). c, The effect of the Ag–Cu thickness ratio on alloying during production on d.c. switching uniformity—normalized on/off uniformity over 100 cycles. d, Uniform switching of the Ag–Cu device during 100 cycles. The nominal thicknesses of Ag and Cu are 2 and 1 nm, respectively. e, Histogram for the set voltage distribution of the Ag (black) and Ag–Cu (red) devices illustrated in a and d. f, Improved retention characteristics of Ag–Cu devices (measured at 0.5 V) with varying compliance currents.

We began by attempting to understand the switching behaviors of our Si ECM memory, which are dependent on the active metals used in its construction. A 6-nanometer thick amorphous Si (a-Si) film atop a p-type Si substrate (0.01 centimeter) containing several active metals, including Ag, Cu, Ni, Cr, and Ti, is used as a device layer (see Methods for details about the device fabrication). Under a d.c. operation mode, a substantial difference in resistive switching performances was detected among the active electrodes, based on their reactivity with Si, as illustrated in Fig. 2-1a. Reversible resistive switching was seen in Ag devices, with the device conductance increasing under a positive biased condition (that is, set) and decreasing under a negative biased condition (that is, reset). After the forming process, however, irreversible conductance alterations were detected in Cu, Ni, Ti, and Cr devices. These metal-dependent switching behaviors are pertinent to recent work on Si ECM memory [21], [55]–[57], and their phase diagram with Si can be understood (Fig. 2-10). Ag is thermodynamically unstable in Si, as seen in the phase diagram, implying that it can be electrochemically mobile and resulting in resistive switching behavior. Cu, Ni, Ti, and Cr, on the other hand, have a significant interaction with Si, promoting the formation of a thermodynamically stable interface between the conduction channel and Si (associated to silicides), which might result in irreversible switching. Although the thermodynamic instability of Ag compared to Si permits resistive switching, it also results in significant switching variation and poor data retention. The set voltage and on/off ratio of the Ag device's associated 100 d.c. switching show temporal fluctuations of 16.4 and 97.6%, respectively (Fig 2-11a,b). For different fabrication batches, the device also shows a lot of spatial variance (Fig 2-11c,d). Furthermore, as illustrated in Fig. 2-10b, the device's conductance levels degrade significantly at each conductance level.

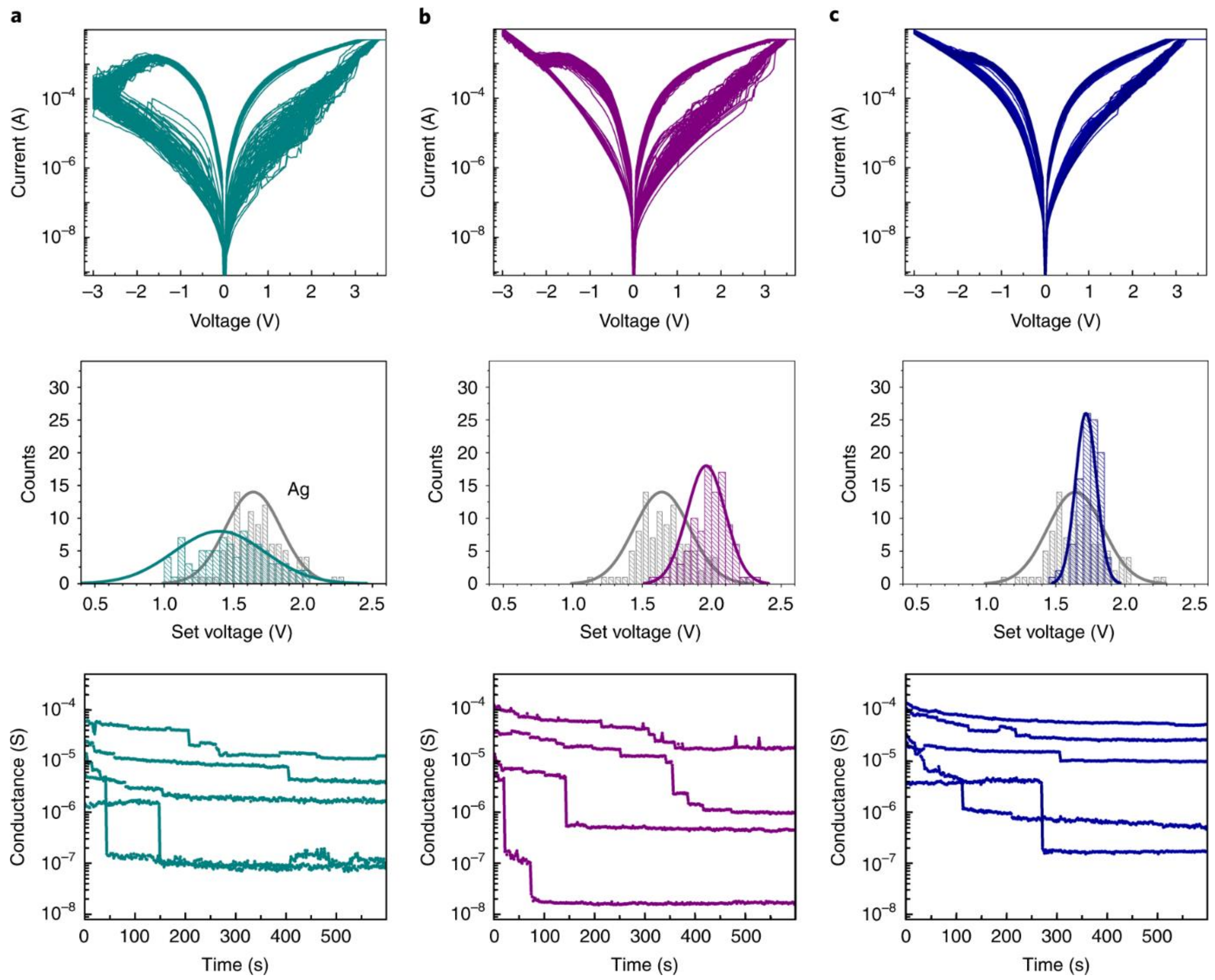

Figure 2-2: The impact of alloying on d.c. switching performance is kinetically and thermodynamically regulated. 100 switching curves (top), set voltage histogram (middle), and retention characteristics with different compliance currents for Ag–Ti, Ag–Cr, and Ag–Ni (bottom). Ag and the alloying element have nominal thicknesses of 2 and 1 nm, respectively.

In order to address issues such as non-uniform switching and short data retention, it may be necessary to improve the interactivity of active metals with their switching medium. We could not choose active metals that would form irreversible conduction channels by reacting too strongly with the Si switching medium. Keeping the requirements for an active metal in mind, we developed Ag alloys with silicidable metals that could perform resistive switching with augmented uniformity and data retention. We centered our design research around satisfying the subsequent three requirements. First, the silicidable metals and Ag must be thermodynamically stable together to increase Ag's stability in Si. Next, the silicidable metals' drift mobility [58] needs to be equal or higher than Ag's drift mobility in order for them to migrate into Si before or at least at the same time as the Ag migration so that the silicidable metal create the Ag conduction channel scaffold. Finally, the proportion of silicidable metal to silver must be minimized to maintain the dominant switching characteristics of Ag. In order to find the ideal silicidable metal, we investigated the phase diagrams of Ag with each silicidable metal (Fig. 2-12) and considered the diffusivity of each silicidable metal in silicon (Fig. 2-13). We chose copper as the complementary alloying element for a variety of reasons. First, as previously stated, Cu has greater diffusivity in Si than Ag (ref.[59]), so it is able to form a backbone or scaffold for the Ag (Fig. 2-13). Additionally, while Ni and Cr are not miscible in Ag, Cu is partially miscible, which forms a bridge to stabilize Ag in Si (Fig. 2-12). We calculated the interfacial energy and stability of the conduction channel between Si and Ag–Cu using density functional theory, and we employed a kinetic Monte Carlo simulation to predict switching dynamics centered on the Ag–Cu alloy to assess our hypothesis about the establishment of the Ag–Cu alloying conduction channel (See Supplementary Note 1, Fig. 2-14–2-16). These simulations indicate that while alloying Ag with Cu appears to help stabilize the conductance channel, Ag can still diffuse in and out during set and reset. We determined the optimal Ag–Cu ratio by assessing the d.c. switching performance of a nominal thickness

control of Ag and Cu during evaporation at various Ag–Cu mixing ratios (see Fig. 2-10c and Fig. 2-5–2-7 for more details). The best Ag–Cu ratio was obtained by comparing the performance of d.c. switching with various Ag–Cu mixing ratios. According to Fig. 2-1d, when Ag–Cu devices were utilized in place of pure Ag devices, a considerable improvement in d.c. switching uniformity was attained (Fig. 2-1a) (Fig. 2-1a). The remarkable change in the temporal variation of the set voltage from 16.4 to 3.3 percent with enhanced spatial uniformity (Fig. 2-1e and Fig. 2-17), even though the formation voltage remained practically identical, was statistically quantified (Fig. 2-1e and Fig. 2-17). (Fig. 2-18). However, as indicated in Fig. 2-1f, when switching from low to high conductance states for Ag–Cu devices, a considerable improvement in data retention was found. This advancement was especially noteworthy (see Fig. 2-19 for details of the retention properties with increased temperatures and extended evaluation times). Additionally, uniform switching properties with consistent data retention were shown to be only marginally dependent on the ambient moisture level (Fig. 2-20), which has been shown to have a substantial effect on switching characteristics [23], [25]. It is worth noting that alloying Ag with silicidable metals does not necessarily result in increased memristive performance, as the alloying process must follow the principles stated above. For example, despite their higher cost, Ag–Ti devices obey the law of miscibility with Ag but have a lower diffusivity than Ag in Si (Fig. 2-13). Because of its low diffusivity, it may not be possible to form a backbone before Ag has migrated, despite the fact that Ti forms stable compounds with Si. As a result of this, Fig. 2-2a demonstrates the non-uniform distribution of d.c. switching in Ag–Ti devices and a low data retention rate, similar to Ag devices. As a result, although the diffusivity of Ag–Cr and Ag–Ni devices is higher than that of Ag (Fig. 2-13), they are not miscible with Ag. In contrast to Ag devices, Ag–Cr–Ni devices exhibit improved d.c. switching uniformity as a result of the formation of backbones in comparison to Ag devices (Fig. 2-2b,c). However, due to the immiscibility of the Ag–Cr and Ag–Ni devices with Ag,

long-term stability of conductance levels was not attainable with these devices. Through these design criteria, we observed that an Ag–Cu alloying is helpful for driving a Si memristor's uniform switching and stable multilevel data retention.

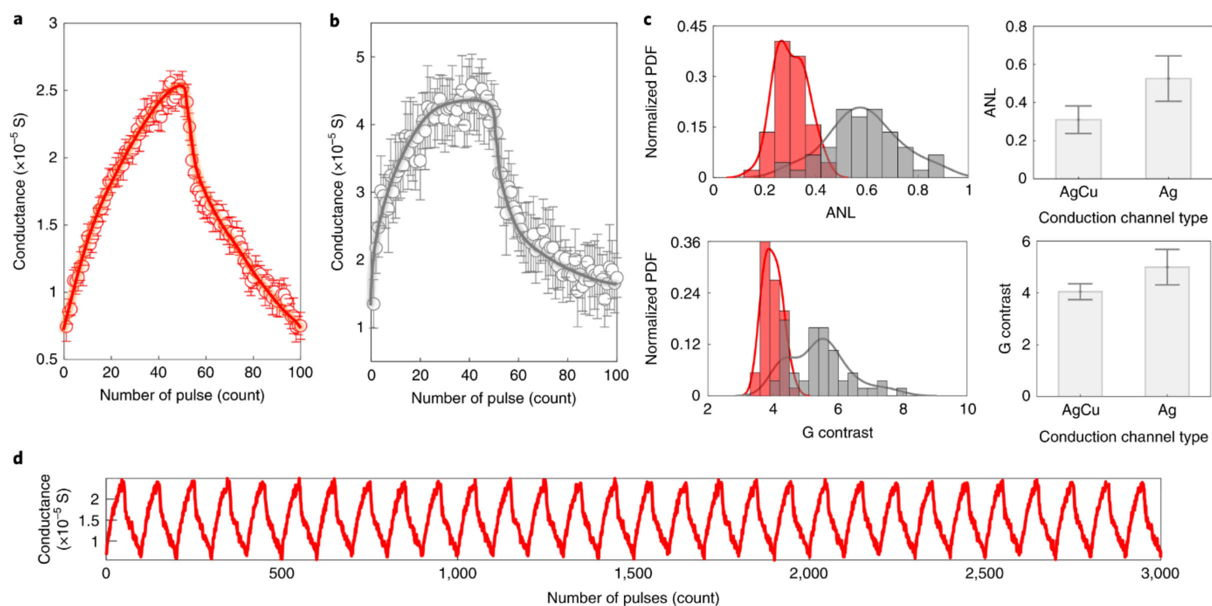

Figure 2-3: The impact of alloy on the analogue nanosecond switching behavior of Si memristors. a,b, Average and standard deviation of 50P/50D ten-cycle conductance updates in Ag–Cu alloy (a) and pure Ag (b) under the same pulse condition (pulse conditions: potentiation (P) of 50 ns, 4.8 V, and n = 50, depression (D) of 50 ns, –2.9 V, and n = 50, and V_{read} of 1 V, 1 ms. c, PDF of the ANL and G contrast from ten devices (50P/50D, five cycles each) for the Ag–Cu alloy (red) and pure Ag (grey). ANL = (GP(N/2) – GD(N/2)/(G_{max} – G_{min}), where G_{max}, G_{min}, GP(N/2) and GD(N/2) indicate the maximum conductance, minimum conductance, median value of potentiation, and medium value of depression, respectively. G contrast equals G_{max}/G_{min}. d, Endurance test of a Si memristor based on the conduction channel of an Ag–Cu alloy. The conductance was programmed at 50P/50D for 30 cycles. The following pulse conditions were used for the endurance test: potentiation of 50 ns, 5 V, and n = 50, depression of 50 ns, 3 V, and n = 50, and Vread of 1 V, 1 ms.

In order to demonstrate the computational capacity of our alloyed ECM memristors, 32 x 32 transistor-less arrays of Ag–Cu devices were fabricated, as depicted in Figs. 2-4a–c, in order to show that they are capable of storing information. The fabrication technique was developed in such a way that the utilization of Si bottom electrodes could be increased to the greatest extent possible (Methods, Supplementary Note 2 and Fig. 2-8). For the purpose of determining the effect of alloying on the crossbar arrays, we also created arrays with silver and silver–nickel devices. For all electrical actions on the array, such as electroforming, programming, and inferences, among other things, a customized measuring system was utilized to ensure that they were carried out correctly (Methods). In order to detect poor retention, it was determined that we would program conductance ranges between zero and fifteen seconds (Fig. 2-25). This is the range in which poor retention can be detected. To demonstrate the weight storing capabilities of the materials, a 256-level greyscale image was encoded into the Ag–Cu–Ag and Ag–Ni arrays, as shown in Figure 2-4d, and the resulting image was displayed in Figure 2-4e. As a result, all of the arrays achieved 100 percent device yield, and the observable evolution of weight values revealed that the Ag–Cu alloy array maintained the intended picture due to the significantly improved data retention when compared to the other arrays. However, the contrast and quality of the images encoded into the Ag–Ni and Ag arrays deteriorated dramatically, with practically all of the recorded information being lost as a result of this degradation. These findings demonstrate the efficacy of the alloy strategy in terms of enhancing the long-term stability of ECM memristors over time, as demonstrated by the results. This makes it possible to operate on arrays with lower conductance ranges, which can aid in the reduction of programming power (see Supplementary Note 3 and Fig. 2-26) as well as the alleviation of sneak path and line resistance difficulties (see Supplementary Note 4 and Supplementary Figs. 2-27–2-30).

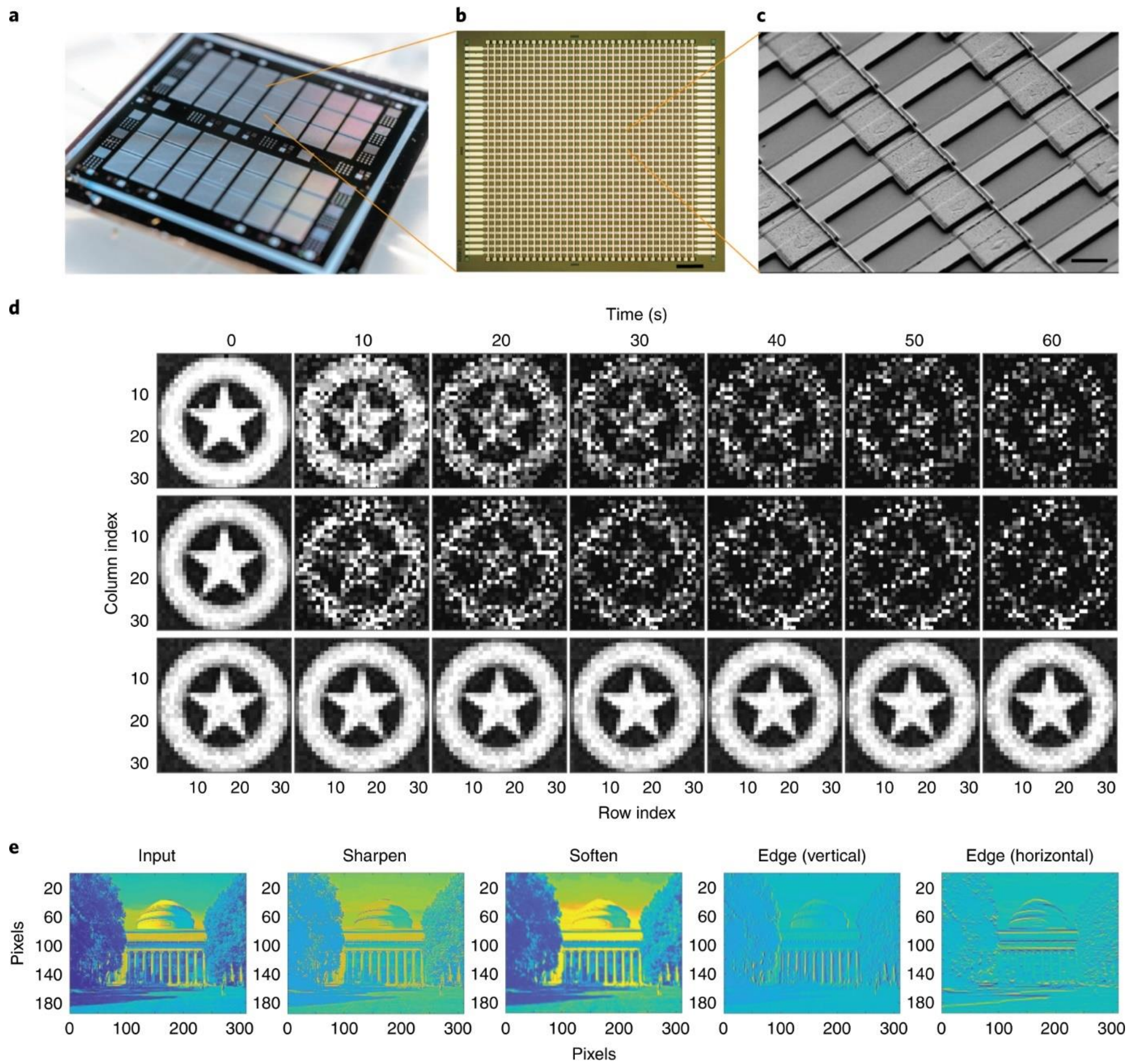

Figure 2-4: 32 x 32 Si memristor arrays with Ag and alloy active electrodes. a, Illustration of an Ag–Cu alloy memristor chip. b, An optical micrograph of a single 32 x 32 array. 240 µm scale bar c, SEM picture of a section of the array revealing the crossbar structure. 15 µm scale bar c, Image programming and data retention experiments in Ag (top), Ag–Ni (middle), and Ag–Cu (bottom) arrays. e, Convolutional kernel processing in the Ag–Cu array, demonstrating the Ag–Cu array's computational capabilities.

We programmed convolutional kernels into the Ag–Cu transistor-less arrays and used them to conduct image processing tasks in order to demonstrate that they were functionally operational. When a kernel (with a tiny matrix of weights) is applied to each pixel and its local neighbors in an image, the convolutional process creates output pixels from the weighted sum operation between the kernel weights and the input pixel values, which is performed on the input pixel values. In order to take advantage of the enhanced data preservation given by the Ag–Cu memristor, faithful image processing based on convolutional kernels was demonstrated, as illustrated in Figs. 2-4e and 2-4f. Parallel processing was achieved by programming four image kernels into four columns of the array (sharpen, box blur, vertical and horizontal edge detection, and vertical and horizontal edge detection) (Methods and Fig. 2-9). To implement negative weights, two memristors in the same output column were employed as a differential pair to accept either positive- or negative-valued input pixels (1 x 18 pixels total for each input vector). It has been demonstrated that our alloy approach is effective for computing applications through the successful image processing performed with Ag–Cu memristor arrays. Large array sizes may be required in order to accommodate more sophisticated jobs. In addition, increasing the array dimension can result in a quadratic rise in the parallelism of multiply-accumulate operations when the array size is increased. In order to lower the line resistances and to overcome more severe sneak-path concerns in large-scale arrays, further scaling of the array dimension needs optimization of the array architecture (Supplementary Note 4) as the array dimension grows larger (for example, integrating selector devices). The precision of the analogue tuning should also be carefully examined because the programming voltage may be disrupted by different weight patterns and increasing line resistances in big arrays, which can cause the programming voltage to fluctuate.

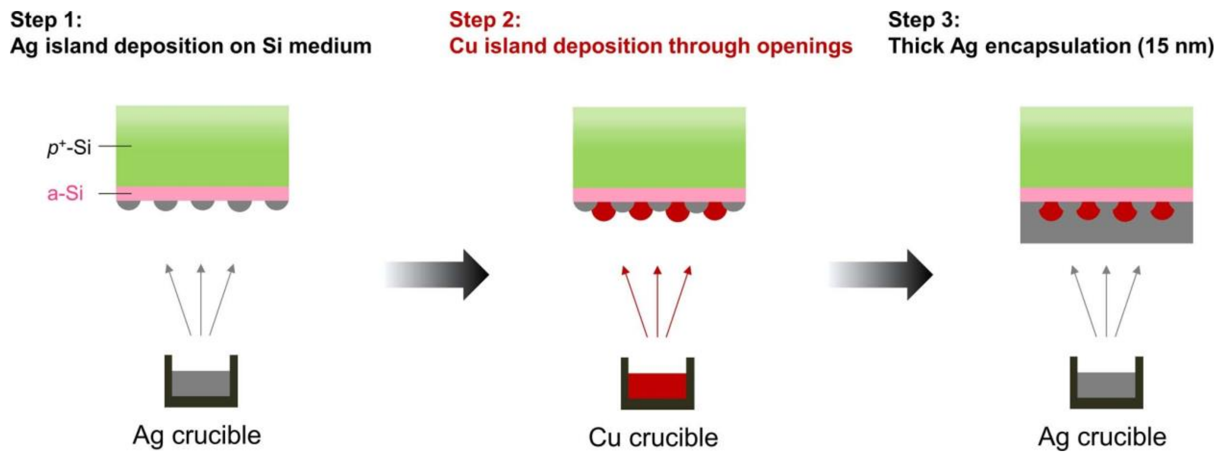

Figure 2-5 depicts the deposition of Ag-Cu films for active metals. We designed a step-by-step metal deposition method to push Ag-Cu together into a switching medium during the shaping stage. We began by depositing ultrathin Ag islands on switching medium, the thickness of which controls the opening area of Si (step1). Following that, we deposited ultrathin Cu on top of Ag islands-deposited Si, where the Cu islands make direct contact with the switching medium (step2). Finally, we enclosed the Ag-Cu islands in a 15 nanometer thick Ag film (step3). The ‘Ag-Cu thickness ratio' determines the quantity of Cu involved in switching. The more Ag film is deposited, the less Cu contributes to switching. During the formation process, the Ag-Cu alloying conduction channels are created in the switching medium.

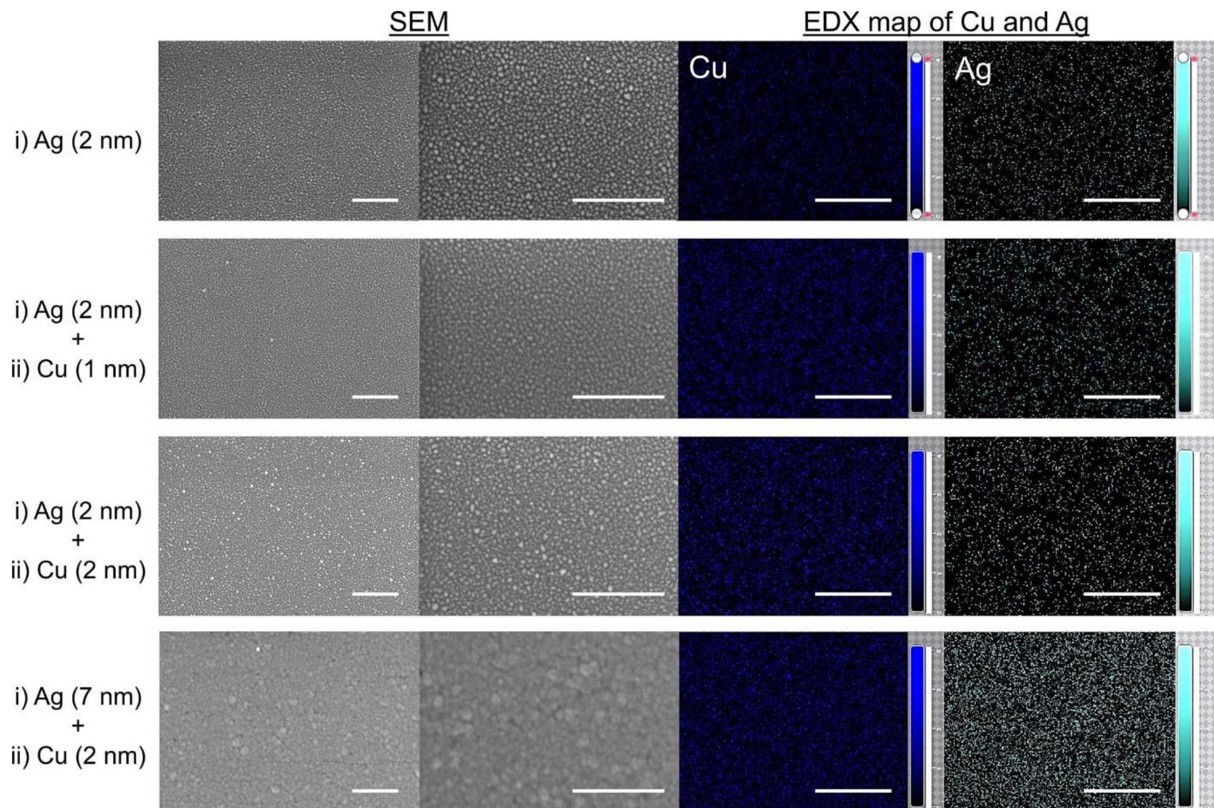

Figure 2-6: The development of an Ag-Cu alloy. SEM images and energy dispersive X-ray (EDX) mapping of an Ag-Cu alloying layer on an amorphous Si surface. 400 μm scale bars When the overall thickness was 2 to 4 nanometers, discontinuous metal films were produced with evenly dispersed metal clusters. Metal clusters merged at a 7-nanometer-thick Ag film with a 2-nanometer-thick Cu film.

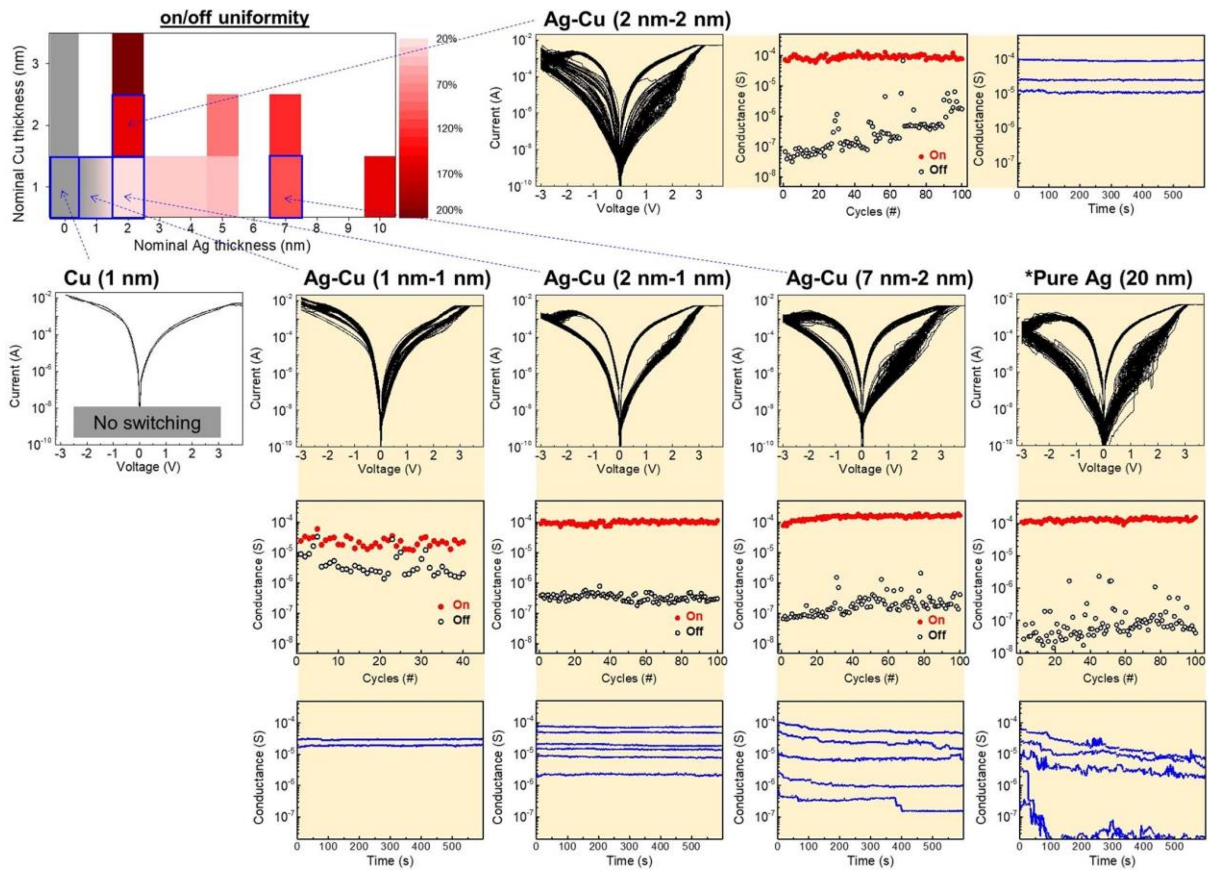

Figure 2-7: The impact of the Ag-Cu alloying ratio on DC switching uniformity and retention. Normalized on/off uniformity of I Ag and (ii) Cu layers with regard to nominal thickness. Following the step-by-step evaporation, a 15-nanometer-thick additional Ag layer was added. For this mapping, the best-performing device for 100 cycles at a compliance current of 5 mA was chosen for each alloying condition. DC switching curves with temporal on/off conductance changes, as well as room temperature retention data, are also provided. When Cu was evaporated even 1 nm before Ag, transistors exhibited irreversible breakdown behavior comparable to pure Cu devices (20-nanometer-thick Cu layer). Switching performance dynamically varied as Ag thickness grew under fixed Cu thickness (1 nanometer), and Ag (2 nanometer)/Cu (1 nanometer) layers generated optimum switching performance: extremely uniform switching with steady retention behavior at multi-level states. As the Ag-Cu ratio diverged from 2 nanometer-1 nanometer, non-uniform switching with poor retention (7 nanometer-1 nanometer) or on/off degradation occurred, although consistent data retention (2

nanometer-2 nanometer) was found. These findings clearly indicate that Cu additions in Ag active electrodes have a substantial effect on switching performance, even though the quantity of Cu is too tiny to create a continuous layer on the Si surface. In addition, Cu's function may be described as follows. (1) Cu improves the stability of an Ag-based conduction channel while lowering the maximum on/off ratio owing to residual Cu elements linked to the Si switching medium (called backbone of the conduction channel). (2) Excess Cu in the Ag active electrode reduces the on/off window as the cycle number increases. An adjusted Ag-Cu ratio, on the other hand, may promote uniform switching with reasonably reliable data retention.

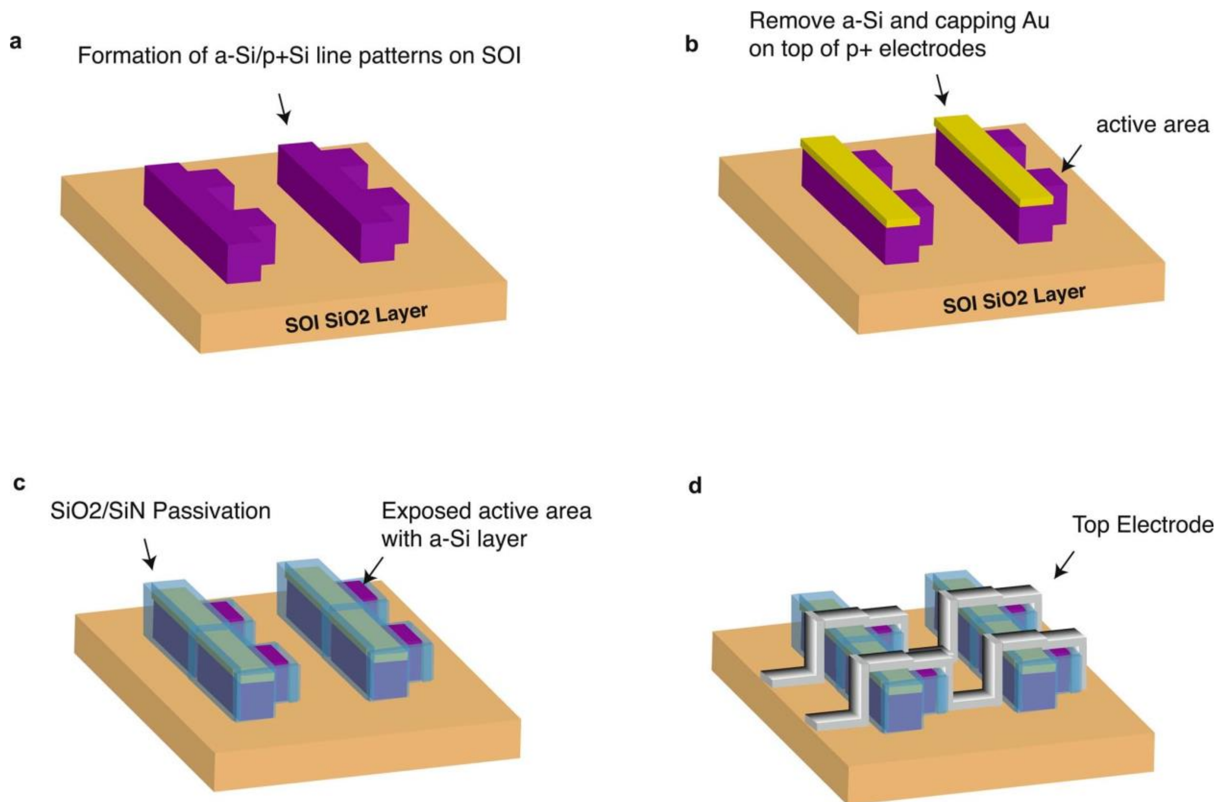

Figure 2-8: Layout of crossbars for alloy arrays with metal capping. In large array implementations, metal capping for p+ Si bottom electrodes is recommended to minimize line resistance. To create an alloy array with a gold capping layer, a novel method was devised. (a) Photolithography and dry etching are used to create isolated a-Si/p+ Si line patterns on an SOI wafer. Active device regions are shown by the protrusions on the line patterns. (b) putting a cap of Au on top of the p+ line patterns to decrease line resistance. The active zones are unaffected. (c) the bottom electrodes are passivated while the active regions are exposed. (d) finishing device arrays by patterning the top electrodes.

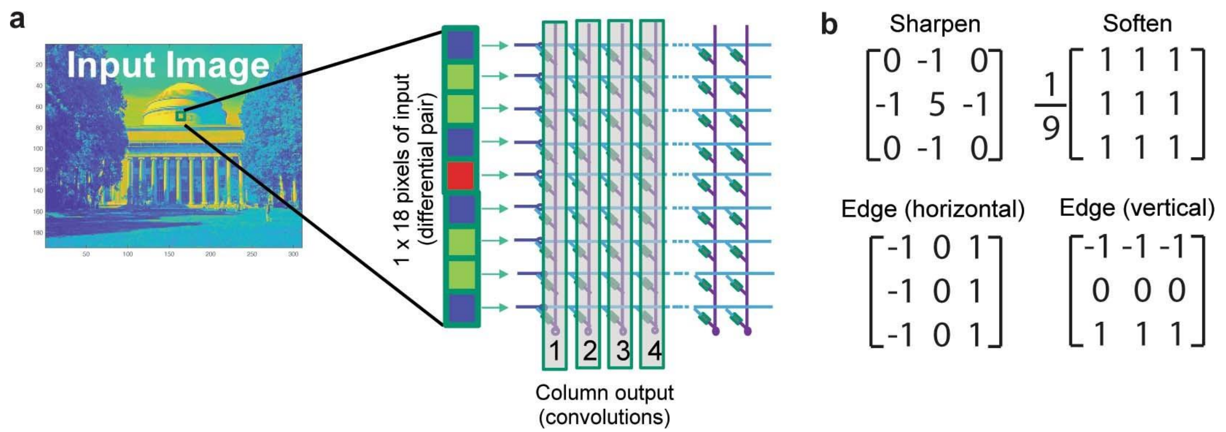

Figure 2-9: Stability demonstration of an Ag-Cu alloy memristor array for inference. (a) For parallel kernel operation, four convolutional kernels illustrated in (b) were programmed into four columns of the 32 32 array. As a differential pair, two memristors are utilized to represent both positive and negative weights.

Below is dedicated to Methods for research work presented in Chapter 2.

Trench-type Si memristor fabrication

A PECVD SiO_2/SiN_x isolation layer (120 nanometer/20 nanometer) was formed at 300 °C and a 6-nanometer-thick intrinsic a-Si film was prepared on a p-type (100)-oriented Si wafer (0.01 ohm centimeter, boron doping concentration of ~10^{19} centimeter^{-3}) by plasma-enhanced chemical vapour deposition (PECVD) at 200 °C. After through-hole patterning in the isolation layer (25–800 micrometer^2), 20-nanometer-thick active metals were deposited into the hole to create contact with the a-Si and a Cr/Au layer (20 nanometer/100 nanometer) was generated on top of the active metals to act as a passivation layer. As the counter electrode of the ECM memory, an ohmic contact between Au and p-Si was created. Cr and SiN_x (refs. [26]–[30]) were chosen to inhibit moisture penetration into the Si switching medium, hence increasing the device's resistance to environmental fluctuations.

Ag alloy deposition

For the deposition of the Ag alloy, a step-by-step evaporation procedure was used. Following the evaporation of the Ag, the alloying metal was evaporated (Figs. 2-5 and 2-6). The switching medium's Ag alloying ratio was adjusted by the nominal thickness of the metals as determined using a quartz crystal microbalance.

Si memristor crossbar array fabrication

As the substrate, a 150-millimeter (100)-oriented silicon-on-insulator wafer was employed. After the initial wafer cleaning, a stack of a 500-nanometer-thick heavily doped p-type Si layer (0.001 ohm centimeter) and a 200-nanometer-thick p-type Si layer (0.01 ohm centimeter) was

homoepitaxially grown on a silicon-on-insulator wafer at 940 °C by ultrahigh vacuum chemical vapour deposition, followed by the deposition of a 6-nanometer-thick intrinsic a-Si thin film using PECVD. The Si bitlines were patterned and isolated from each other using photolithography and dry etching. Additional Cr/Au layers (1 nanometer/100 nanometer thick) were designed and deposited on top of the bitlines using photolithography and electron-beam evaporation to reduce wire resistance and serve as measurement input/output pads, as illustrated in Fig. 2-8. After that, a PECVD layer of SiO2/SiNx (120 nanometer/20 nm) was deposited to cover the entire wafer as a passivation layer for the bitlines. Following that, the active device region and input/output contact pads for the bitlines were formed using a photolithography etch that included reactive ion etching and wet etching with a buffered oxide solution to remove the capping layer selectively. The Ag alloy layer (20 nanometer) and Cr/Au capping layer (20 nanometer/50 nanometer) were deposited over the active area to serve as the memristor's active top electrodes. The fabrication process was completed by patterning and depositing the wordline using tilted Au sputtering.

Device d.c. measurements

The B1500A semiconductor device parameter analyser, in conjunction with a B1517A high-resolution source measure unit, was used to perform quasi-static d.c. current–voltage measurements as well as room-temperature state-retention studies. During the forming and setting process, bidirectional current–voltage sweep measurements were performed on the Si memristors with a compliance current to ensure proper operation.

Ultrafast pulse measurement

An oscilloscope (DSOX3024T, Keysight), a pulse generating unit (PGU 33600A, Keysight), and a transimpedance amplifier (DHPCA-100, Edmund 59–179) were used to conduct an

ultrafast (nanoscale) analogue switching test. Weight update was carried out using a series of programming pulses for potentiation and depression in the same amplitude but opposite polarity as the original pulses. After each programming pulse, read pulses were delivered to our memristive devices in order to track the change in conductance over time. To ensure that the conductance values were stabilized, a current from the read pulse was integrated and averaged for one millisecond. Because wave reflections are likely to occur in the radio-frequency domain, the impedance value of the oscilloscope was set to 50 ohm, and the load impedance of the pulse generator unit was set to infinite to prevent reflections.

Array measurement

The array measurement was carried out using a board-level peripheral system that had been designed to have parallel access and programming capability. The system's specifics have been previously published. [60]. The memristor arrays were accessed using a 32 x 32 probe card that was linked to the peripheral system and measured the voltage across them. In order to facilitate selective programming while also minimizing sneak-path issues, a 1/2 voltage biasing scheme was implemented (selected rows were biased at 1/2 of the operating voltage, and selected columns were biased at 1/2 of the operating voltage, whereas all of the unselected rows and columns were grounded). To combat sneak current during computation, the ground scheme (in which all column outputs were essentially grounded by the transimpedance amplifiers) was utilized for inference to suppress it. Supplementary Note 4 provides a detailed examination of the array operations performed on our passive Ag–Cu memristor array. Each device in the produced array must undergo an initial electroforming procedure before it can be used. It was necessary to perform the forming process in a series by applying a train of ramping voltage pulses to the device until its conductance exceeded the one microsecond threshold. Following the formation of the gadget, it was reset to its initial off state by the manufacturer. The forming voltages (which were picked between 4 and 8 V) were tiny enough that they did not cause

problems with devices that were not selected being set to 1/2 voltages. Each device in the array was cycled between 5 and 20 μS at least five times after electroforming to ensure that it was ready for application.

Image programming and retention test

A closed-loop technique was used to program greyscale images into different alloy arrays, which were then tested. The conductance values between the maximum and minimum conductance values defined for each task were linearly transferred to the 256-scale pixel values using linear mapping. The conductance tuning of pixels and devices was carried out in parallel within each column to allow for rapid programming. It was necessary to perform the retention test several times at an interval of 10 seconds by repeatedly checking the device conductance. When comparing different alloy memristors, the raw conductance value as well as the reconstructed image based on 256-scale pixel values were both employed to make the comparison. It was necessary to reverse and linearly translate the conductance values back to the greyscale pixel values in order to obtain the final result.

Convolutional kernel operation

In order to demonstrate the feasibility of using Ag–Cu alloy memristor arrays for reliable inference applications, four convolutional kernels were selected for use in a proof-of-concept presentation. Each kernel consisted of three pixels by three pixels, and two memristors were utilized to represent both positive and negative weight values in each kernel. The 3 x 3 x 2 memristors were mapped into an 18 x 1 vector and programmed into a column of the array using the MATLAB programming language. The 3 × 3 input pixels were applied accordingly to the input rows, with both positive- and negative-valued pixel amplitude for the differential

memristor pair. After programming the kernels, 3×3 input matrices from a 310×194 pixel image were fed to the array in series for iterative convolutional kernel operations. Through the use of a transimpedance amplifier, the column outputs from each cycle were converted into voltage amplitude, which was then read out by analog-to-digital converters and recorded as a single pixel in filtered images.

Below is dedicated to Supplementary Notes for research work of the chapter 2.

Supplementary Note 1. DFT and KMC simulation for conduction channel formation

1.1. DFT calculation of stability of the interface between metal clusters and Si medium

When metallic clusters are formed in a solid electrolyte, extra pressure (ΔP) is applied on the clusters due to the surface tension [61], [62]. Δp is given approximately by $\Delta p = 2 \gamma / r$, where γ is the interfacial energy and r is the radius of the clusters. As a result, extra pressure can rupture metal clusters, and interfacial energy should be reduced to improve metal cluster thermodynamic stability. The use of Ag clusters has been used to show volatile resistive switching devices (also known as diffusive memristors) that are based on the high interfacial energy of the Ag/switching medium. [63]. This is consistent with our findings that pure Ag produces unstable conductance weight, which is explained by the thermodynamic immiscibility of Ag in Si. As a first-order approximation, the interfacial energy between Si surface and metal (alloy) layers were investigated using DFT calculation with Vienna ab-initio Simulation Package (VASP). With a cut-off energy of 450 eV, the valence electrons were extended into projected enhanced waves. The exchange-correlation effect is described using

the Perdew-Burke-Ernzerh technique. The electronic and ionic convergence requirements were 10^{-6} eV and 10^{-5} eV, respectively. The Si-metal interface was built using 2x2 Si-3x3 Cu, 3 x 3 Si-4 x 4 Ag, and bi-axial strain of 0.4 percent and -1.5 percent on the metal layers, respectively, to reduce lattice mismatch. A 4 x 4 x 1 Monkhorst-Pack mesh was used for integration across the first Brillouin zone. The interfacial energy (int) between Si and the metal layer is defined as follows: E_{Si} is the energy of the silicon substrate, E_{metal} is the total energy of the metal (alloy) layer, E_{system} is the total energy of the metal/Si stacked system, A is the interface area, Si is the surface energy of Si, and metal is the surface energy of metal, respectively [64], [65]. Two Cu-Ag alloying scenarios are investigated over the Si surface: (1) Si in contact with metal layers (Ag and Cu) and (2) Si in contact with a complete Cu-Ag stoichiometric mixture (Si/Cu-Ag). The interfacial energy between the metal layer and the Si switching medium is seen in Fig 2-14. Cu incorporation into Ag clusters reduces interfacial energy, implying that alloying Ag with Cu improves the stability of Ag-based conduction channels by lowering surface tension. As a consequence, consistent and reliable switching with symmetric analog weight update is achieved.

1.2. KMC simulation for conduction channel formation

The simulation results of (1) Forming, (2) Reset, and (3) Set operations are shown in Fig 2-15. Supplementary Note 1.3 contains the details of the simulation conditions and parameters.

(1) Forming process

Anodic oxidation occurs at the active metal when a positive bias is provided, and metal cations move into the Si matrix. Because of the low ionic conductivity of the Si matrix, cations are reduced and metallic clusters are generated at the Si bulk (i.e., conduction channel creation)

rather than on the surface of the inert electrode during migration, as seen in the 3s recorded image. Cu and Ag form metallic clusters at the same time because their mobility is equivalent to or quicker than that of Ag. The interfacial energy of Si matrix/metal clusters determines the stability of metallic clusters.[61], [66]. Whether the dominant phase of Cu clusters is a silicide or not, thermodynamically stable clusters are driven by attractive interfacial contact between Si and Cu. However, Ag clusters are unstable in the Si matrix. Our DC sweep results clearly imply that the Cu-based conduction channel is too steady to induce resistive switching. As a result, the amount of Cu-based clusters should be kept to a minimum to avoid irreversible breakdown, and Ag clusters should be the major component of the conduction channel. Because Cu improves the thermodynamic stability of Ag in Si, it can function as a nucleation promoter for Ag clusters.

(2) Reset

Joule-heating-assisted electrochemical oxidation occurs in the conduction channel when negative bias is applied [67]. Because of the thermodynamic instability, when the conduction channel is exclusively made of Ag, the channel is easily eroded: high Ag/Si interfacial energy increases Ag cluster contraction in addition to thermo-electrochemical stressors. Ag clusters are predominantly dissolved in the Ag alloying channel, while Cu clusters are largely preserved at the Si matrix based on a 7 s captured image. Cu residual clusters could be the source of the stable off-state conductance level.

(3) Set

Ag devices have non-uniform SET performance because channel re-formation is stochastic by nature. However, uniform switching is enabled in Ag-Cu alloy memristors because Ag clusters

are re-formed around Cu clusters, which act as the skeletal structure for the conduction channels.

1.3. Configuration of KMC simulation

We have included different physical and chemical processes in our KMC simulation since it is based on electrochemical metallization (ECM) RRAM: (1) Ag^+ /Cu^+ cation dissolution from the anode (dissolution), (2) Ag^+ /Cu^+ cation diffusion in the dislocated Si (diffusion), (3) Ag^+/Cu^+ cation reduction at the nucleation site (electro-crystallization), (4) Ag-Cu cluster growth from a single nucleation atom (metal clustering), and (5) Ag/Cu atom oxidation from Ag-Cu cluster and Ag/Cu atom oxidation from the nucleation site (oxidation). All reduction, oxidation, and diffusion rates are expressed as $P = f \cdot \exp(-E_a/k_BT)$, where f is the vibration frequency, E_a is the activation energy that depends on each process, k_B is Boltzmann constant, and T is the temperature. Table 2-1 contains a comprehensive list of all important parameters. According to Ref [68], [69], the activation energy barrier for cation dissolution from the anode is investigated in this study. The cation diffusion in Si is referred from Ref [19]. The computational approach estimates the cation reduction and oxidation activation energies at the nucleation location. When the metal-Si bonding energy is higher, reduction is more likely to occur whereas oxidation is more difficult [62]. This implies that the activation energy of Cu for reduction is low, whereas Ag's activation energy for reduction is large. The thermodynamic nucleation model is used to calculate the activation energy of Ag-Cu cluster growth. The equilibrium thermodynamics-based classical nucleation model [63], both homogeneous and heterogeneous nucleation are mainly governed by Gibbs free energy (ΔG) associated with the surface energy of cluster (Φ) as expressed as $\Delta G = \Phi - \Delta\mu$, where $\Delta\mu$ represents super-saturation which indicates the electrochemical potential difference between metal cations and

fixed metal atoms in cluster. Since Δμ is proportional to bonding energy, Cu is estimated to have lower activation energy for cluster growth than Ag [64]. Other factors such as diffusion or redox activity energy barrier are also assessed using the DFT computation. The activity energy associated with the oxidation process of the AgCu cluster is proportional to the number of bonds linked to the atom. The oxidation of cluster atoms with more linked bonds necessitates a higher activation energy. When an external voltage is applied, Ea can be altered through both physical and chemical processes. The diffusion barrier is decreased along the electrical field direction, causing cation migration from anode to cathode. For the redox reaction, Ea for the forward and reverse transitions can be described as $-\alpha q\eta$ and $(1-\alpha)q\eta$, where α is typically 0.5 and η represents the electrochemical overpotential [62], [70]. A random resistor network based on percolation theory is used to replicate the above microscopic process [71], [72]. In this model, the resistance value of the bond connecting Ag/Cu metal atom site is the lowest denoted by r_l^{metal} , and the resistance value of the Si-Si bond is the highest denoted by r_h^{bulk}. The resistance value of the bond connecting Ag/Cu atom and Si atom falls in between r_l^{metal} and r_h^{bulk}. All of the resistance's I-V characteristics follow the laws of ohmic behavior. This results in an electric potential distribution being derived from the Kirchhoff equation, while the temperature distribution is supplied by the Fourier heat equation as a consequence of the Kirchhoff equation: $CdT/dt = \nabla(k \cdot \nabla T) + Q$, where C is the heat capacity per unit volume of Si bulk, k is the thermal conductivity of Si bulk and Q is the Joule heat power density. On the basis of the electric potential distribution and the local temperature distribution, the calculations for all microscopic processes are made. Fig 2-16 depicts a flowchart of the KMC simulation procedure.

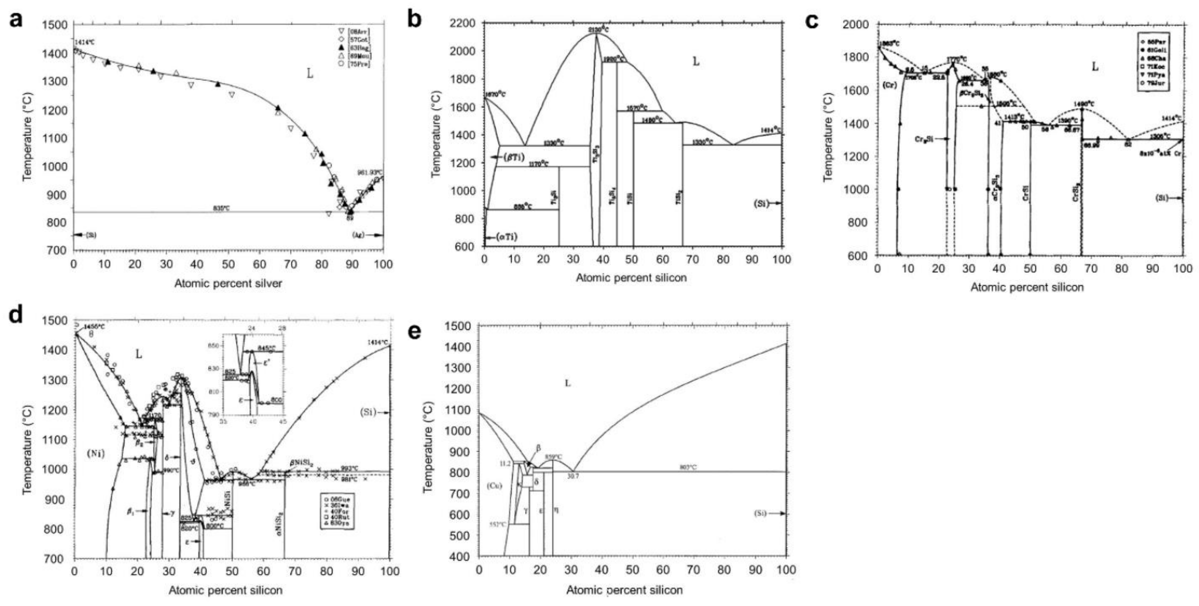

Figure 2-10: Phase diagram of metal-silicon. a, Ag-Si[7]. b, Ti-Si[8], c, Cr-Si[9]. d, Ni-Si[10]. e, CuSi[11]. With the exception of Ag, the production of silicides is thermodynamically favored for the elements Ti, Cr, Ni, and Cu.

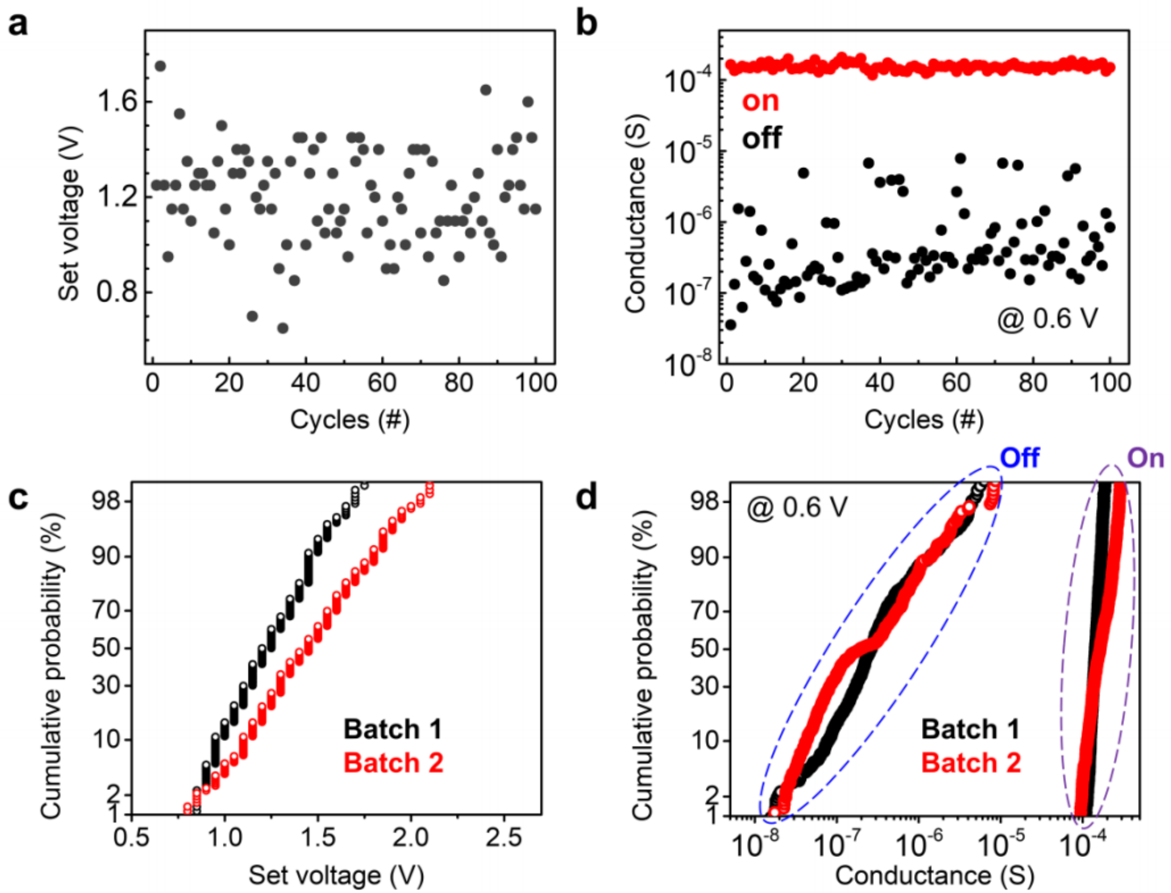

Figure 2-11: Ag memristor DC switching uniformity. Temporal change in the set voltage (a) and on/off conductance (b) of an Ag device, as seen in the manuscript's Fig. 2-10a. Variation in set voltage (c) and on-off conductance throughout space (d). Each batch contains 5 devices that have been tested under the identical DC working conditions (compliance current, 5 mA, >100 DC cycles per device). The standard-deviation-to-mean of set voltage and on/off ratio were calculated to be 16.2 percent and 156.2 percent for batch 1 and 18.7 percent and 189.7 percent for batch 2.

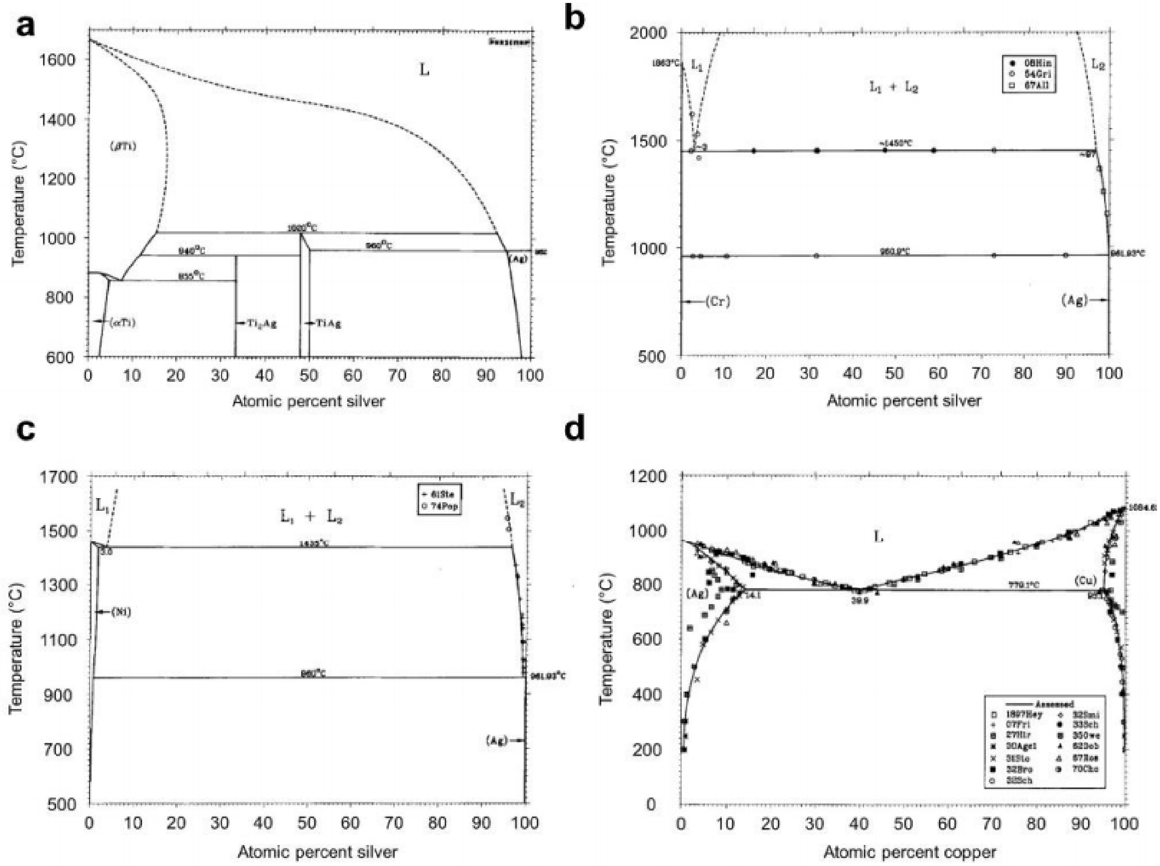

Figure 2-12: Phase diagram of metal-silver. a, Ti-Ag[12]. b, Cr-Ag[13]. c, Ni-Ag[14]. d, Cu-Ag[15]. Ti produces intermetallic compounds with Ag, showing that Ti and Ag have an attraction force. Despite the fact that Cr, Ni, and Cu form a solid solution system with Ag, a miscible zone exists in the Ag-rich phase of Cu-Ag alloy. Thus, Cu-Ag can form a thermodynamically stable mixed compound (i.e., a solid solution), whereas repulsion force occurs at Cr-Ag and Ni-Ag regardless of the mixing ratio or temperature.

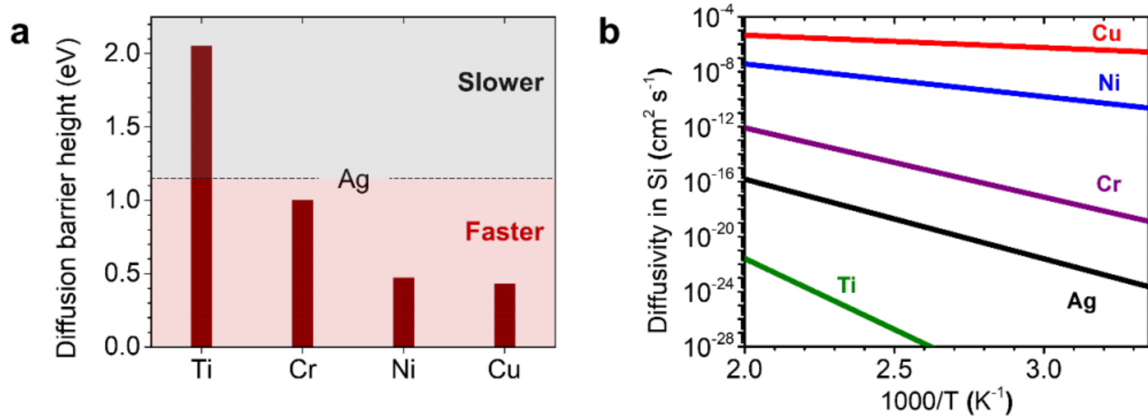

Figure 2-13: Metal diffusivity in Si. a, Diffusion barrier height of Ti[16], Cr[17], Ni[18], Cu[19], and Ag[20]. b, The metal diffusivity Arrhenius plot. Cu, Ni, and Cr diffuse quicker than Ag, whereas Ti is the slowest metal.

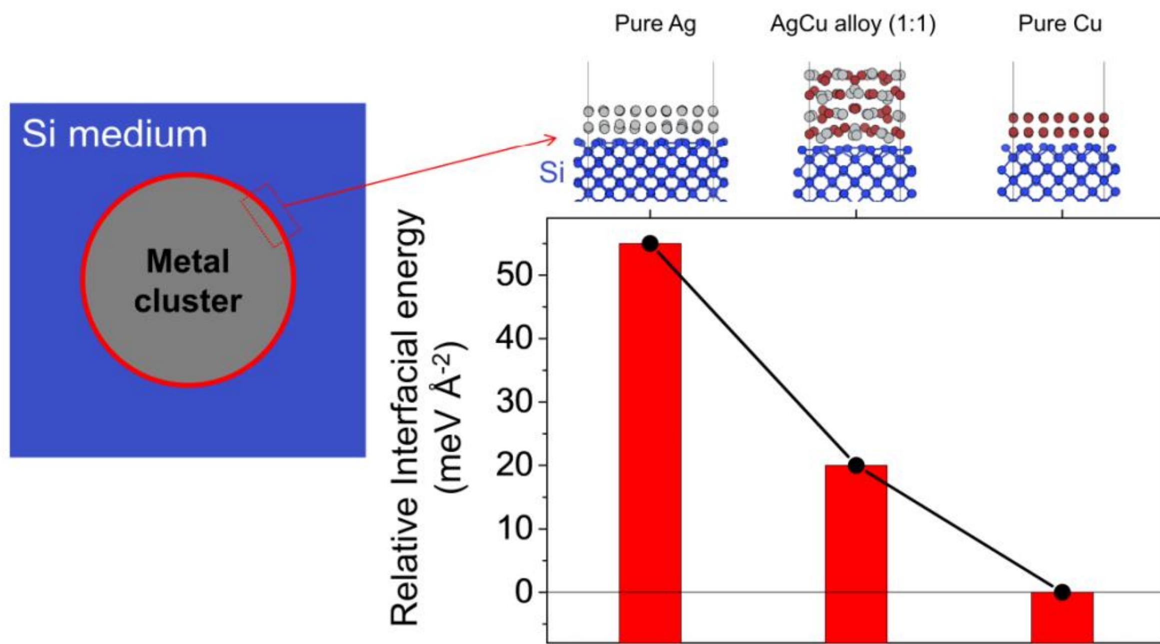

Figure 2-14: The interfacial energy and relaxed structures of an Ag-Cu layer on a Si switching medium When we look carefully at the interface of metal cluster and Si medium, as shown in the schematic (left), there are three probable possibilities as described: (1) pure silver, (2) an Ag-Cu alloy, and (3) pure copper. The addition of Cu reduces interfacial energy, while the interfacial energy of pure Ag clusters is 55 meV Å^{-2} higher than pure Cu clusters. This means that the external pressure on Ag-Cu clusters is reduced due to lower interfacial energy when compared to pure Ag clusters, and that alloying Cu with Ag can improve the thermodynamic stability of Ag-based conduction channels in Si switching medium. Supplementary Note 1.1 has the simulation information.

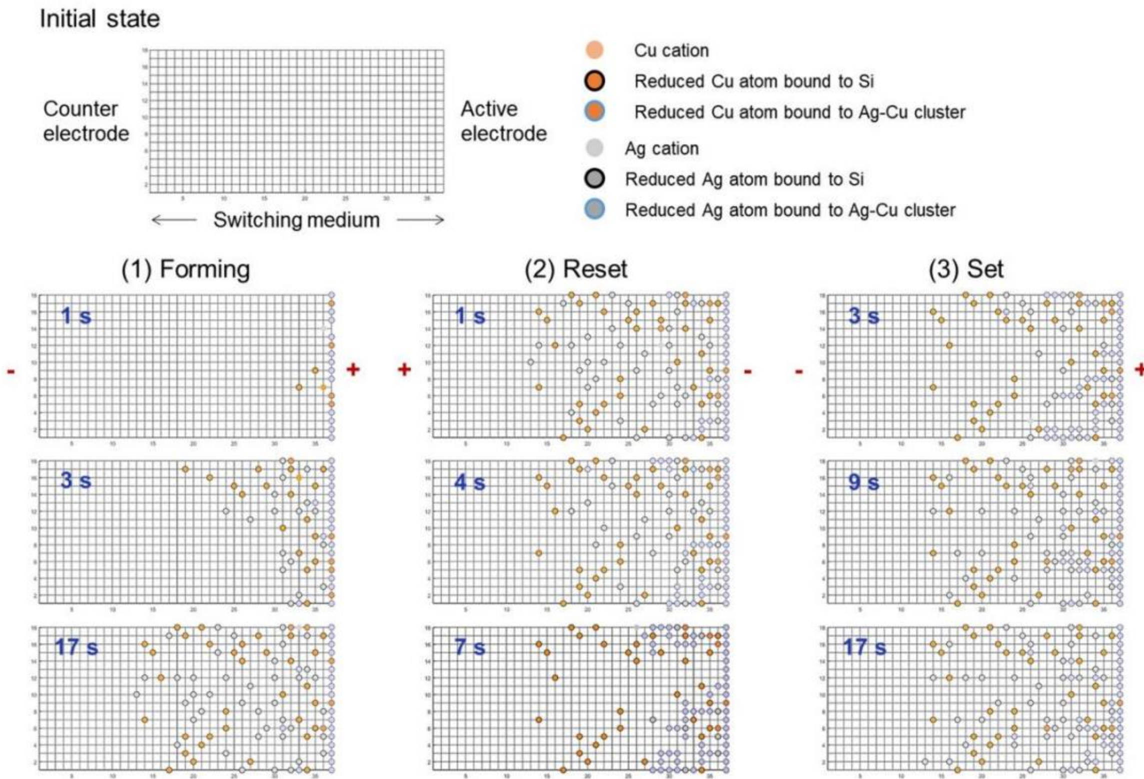

Figure 2-15: Switching dynamics based on alloying conduction channel: Forming (left), reset (middle), and set (right) states. In this simulation, we look at a conduction channel generated by mixing Ag and Cu atoms in a Si switching medium. Actual alloying has been considered while developing with an incoming uniform mixture of Ag/Cu atoms inside a switching medium. For the initial state, we examined the atomic fractions of Ag and Cu, as well as their activation energies for anode cation dissolution. The 1 s state of formation is the stage at which Ag and Cu atoms begin to dissolve into the Si switching medium based on their activation energy. Furthermore, during reset/set procedures, Ag clusters are dissolved/rejuvenated to a greater extent than Cu clusters. These leftover Cu atoms can serve as the backbone of the conduction channel, improving switching uniformity and conductance state stability. Supplementary Note and Table 2-1 give the simulation circumstances and parameters, respectively.

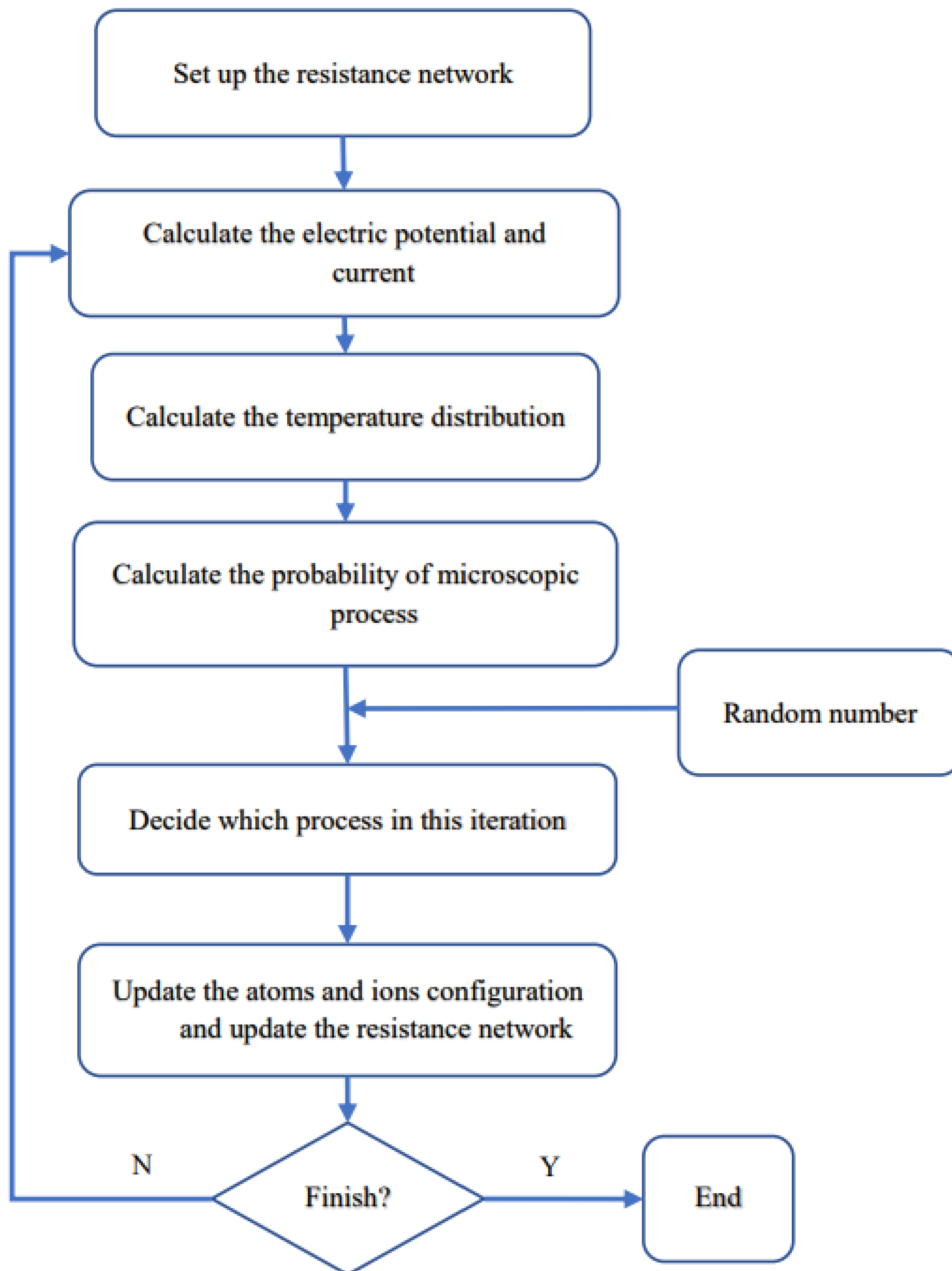

Figure 2-16: Schematic representation of the KMC simulation.

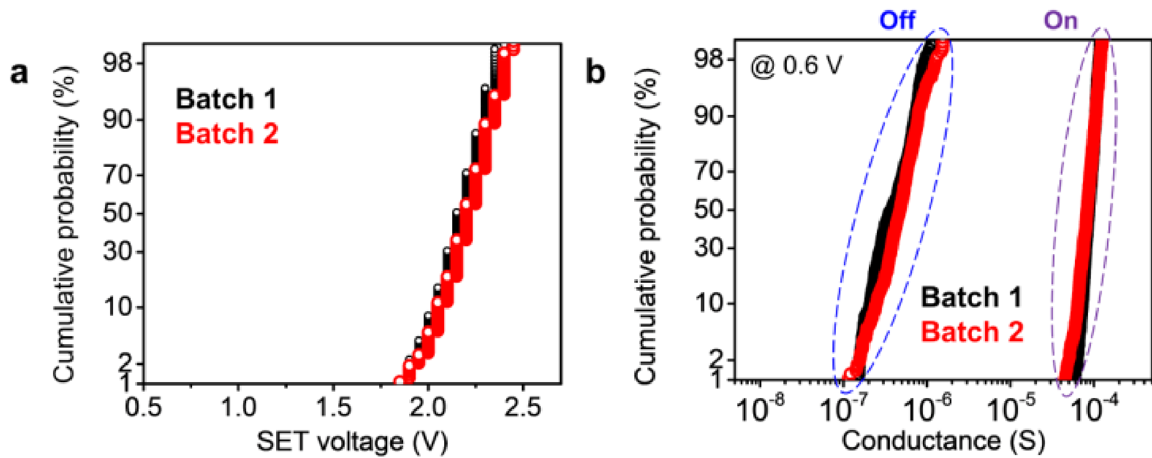

Figure 2-17: Ag-Cu memristor spatial variation. The cumulative probability of (a) the set voltage and (b) the on-off conductance (read voltage, 0.6 V). Each batch contains 5 devices that have been tested under the identical DC working conditions (compliance current, 5 mA, >100 DC cycles per device). The standard-deviation-to-mean (/) of set voltage and on/off ratio were calculated to be 5.1 percent and 49.4 percent for batch 1 and 4.9 percent and 47.1 percent for batch 2.

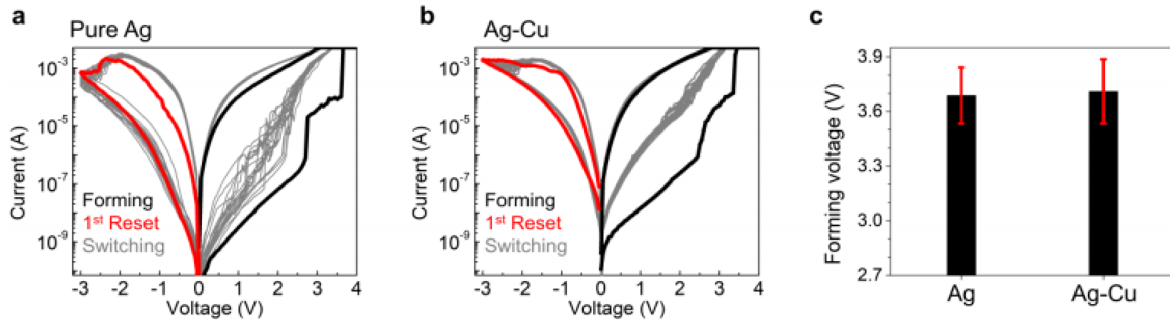

Figure 2-18: Typical forming and subsequent reset processes of pure Ag devices (a) and Ag-Cu alloy devices (b). It should be noted that high formation voltage (>10 V) is undesirable for memristor crossbar array operation because it can cause irreversible breakdown of the devices and, as a result, reduced device yield [21], [22]. The forming voltage of Ag and Ag-Cu devices is roughly 3.7 V, which is 12 V greater than the fixed voltage (c). We believe that this difference is allowable for the array operation, which is strongly supported by the high yield (100 percent) of Si memristor crossbar array.

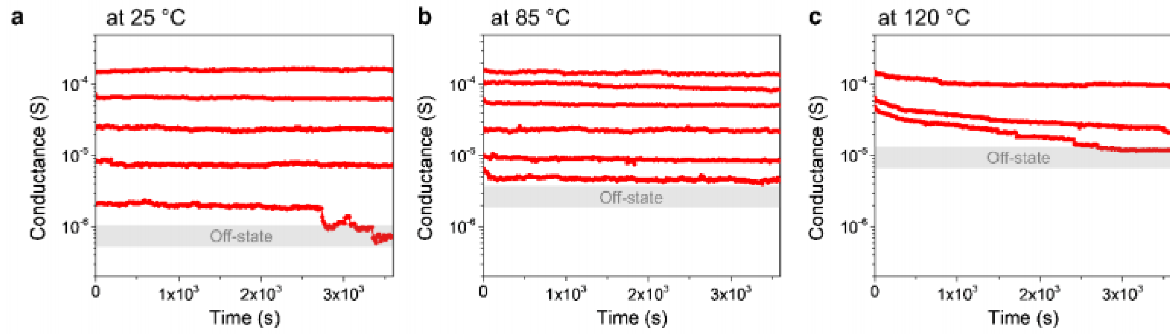

Figure 2-19: Retention test with raised temperature for 1 h. At room temperature (a) and 85 °C (b), conductance levels (over 10 μS) were stable, but lower conductance levels could not be achieved due to poor stability and increased off-state conductance caused by thermal excitation of free carriers from the p+ -Si layer, respectively. However, as temperature rose to 120 °C (c), conductance steadily declined, which is similar to the retention behaviors of pure Ag devices at room temperature.

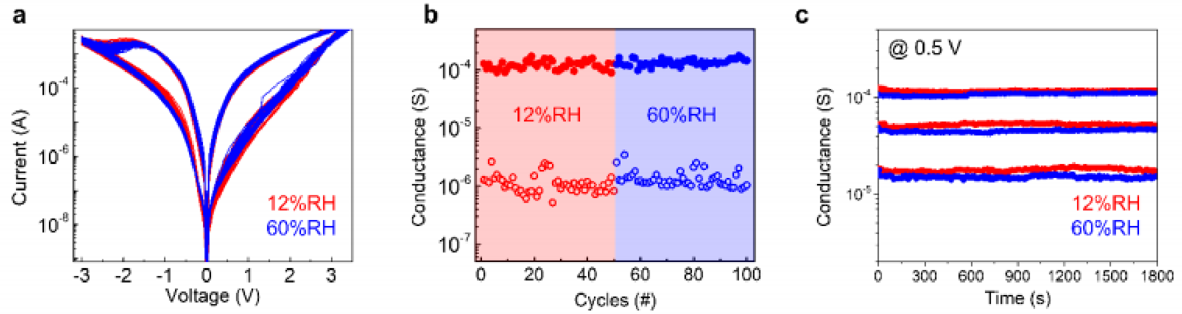

Figure 2-20: The effect of ambient moisture on the memristive performance of Ag-Cu devices. DC switching (a), on- and off-state conductance (b), and retention characteristics (c) as a function of relative humidity (% RH) at ambient temperature. Because moisture (H2O) in a switching medium plays a significant role in redox-based switching dynamics, the level of ambient moisture can influence the switching performance of redox-based memristors [23]–[25]. Despite the fact that the humidity level was altered from 12 to 60 percent RH, Ag-Cu devices demonstrated similar switching behaviors and retention properties. This consistent behavior could be due to the device's passivation layers, which prevent H_2O migration into the Si switching medium: Cr[26], [27] and silicon nitride[28]–[30] layers, known as excellent materials for preventing water molecules/ions from penetrating (i.e., anti-corrosion), cover Ag-Cu active metal and Si switching medium.

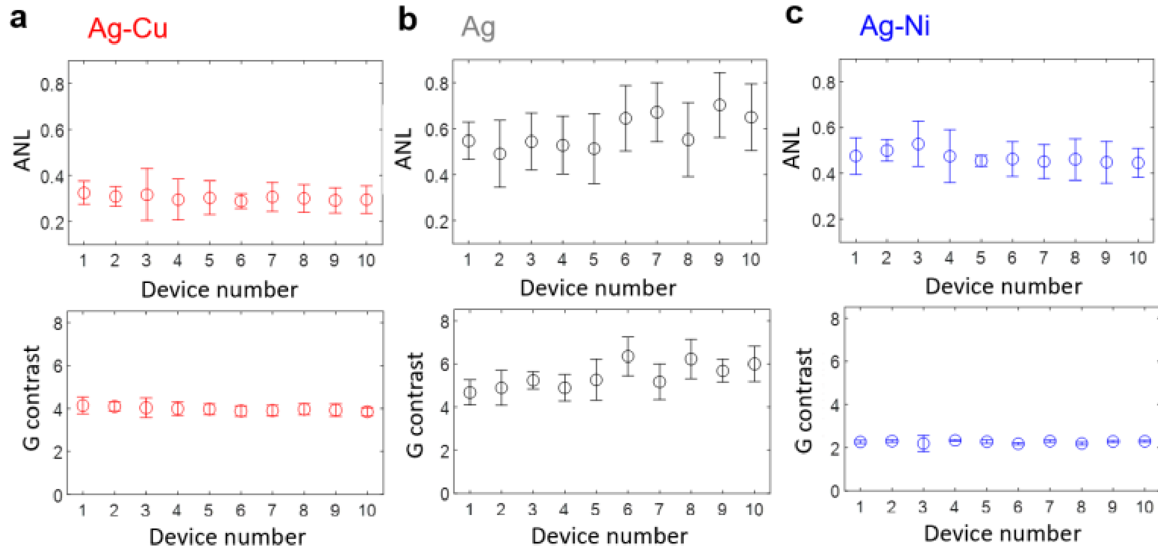

Figure 2-21: ANL and G distinguish between temporal and spatial differences in analog switching. For conductance update from 10 Si memeristor devices with three distinct active metals (a) Ag-Cu alloy, (b) pure Ag, and (c) Ag-Ni alloy, 5-cycles of 50 potentiation and 50 depression pulses are used. Pulse condition: Potentiation (50 ns, 4.8 V, n = 50), Depression (50 ns, -2.9 V, n = 50), Vread (50 ns, -2.9 V, n = 50). (1 V, 1 ms).

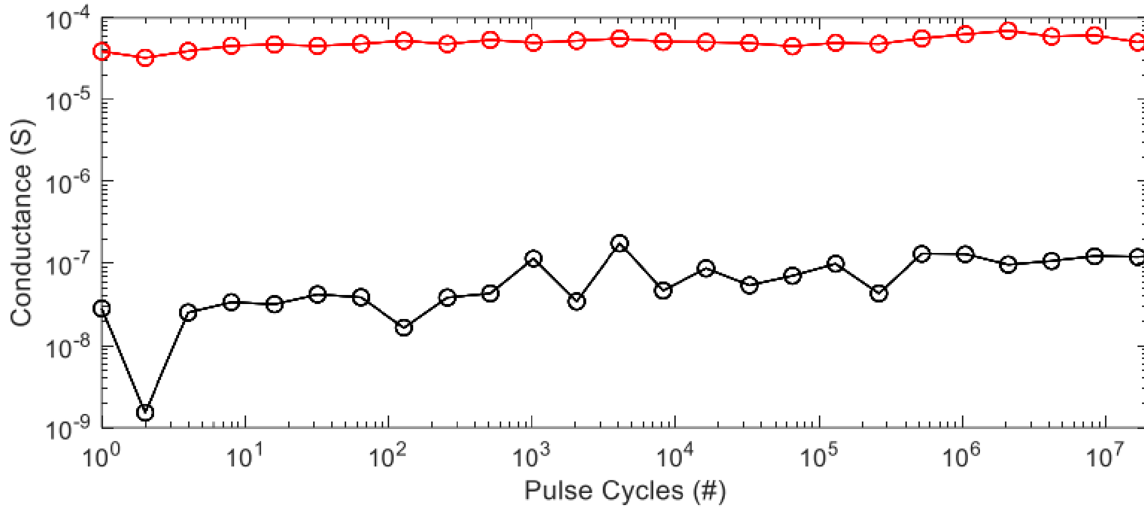

Figure 2-22: Endurance of Ag-Cu alloyed device. Pulsed voltage stresses (PVS) test is performed under a square pulse condition (non-sinusoidal periodic waveform) on small device cells (< 25 μm^2). Each pulse cycle is made up of two processes: potentiation and depression. While the potentiation process (VP) uses 50 pulses of 4.5 V (amplitude) and 50 ns (duration), the depression process (VD) uses 50 pulses of -2.8 V (amplitude) and 50 ns (duration). Data points representing ON (red dots) and OFF (black dots) states are gathered at a power of 2 DC cycles (2n, n = 0 to 24) with read voltage (0.5 V). Throughout the PVS test (> 109 pulses), the device maintained a high on/off ratio (> 100), which is significantly greater than 5. (considered as device failure) [31], [32]. DC condition: set (4 V), reset (-3.2 V), and compliance current (5 mA).

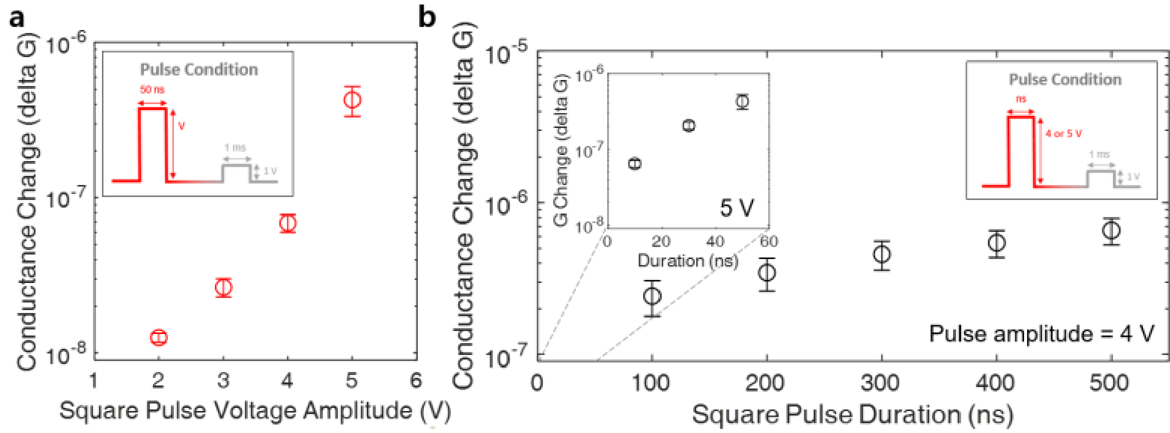

Figure 2-23: The variation in conductance (ΔG) as a function of pulse amplitude and duration. The conductance change (ΔG) is measured using a nanosecond pulse. At a read voltage of 1 V, we set the initial conductance at 10^{-6} S. (a) Change in conductance with increasing square pulse voltage amplitudes (V = 2, 3, 4, and 5). The pulse condition for conductance measurement is shown inset in (a). (b) Change in conductance with different square pulse durations at the nanosecond level using 4 V nanosecond pules. The conductance shift with ultrashort pulses is seen in the inset (left) (10 ns, 30 ns, and 50 ns with 5 V nanosecond pulses). The pulse condition for conductance is shown in the inset (right) in (b). In both cases, a 5 µm × 5 µm cell is used.

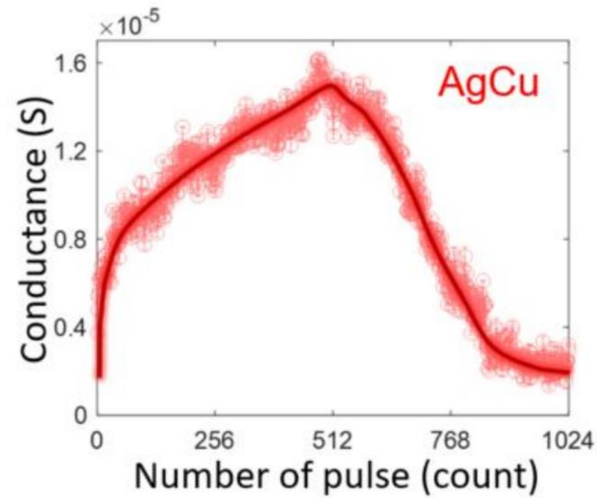

Figure 2-24: Analog potentiation and depression in 512 steps (512P/512D). For conductance update in an AgCu alloy Si memristor device, three cycles of 512 potentiation and 512 depression pulses are used. Pulse condition: Potentiation (200 ns, 3.7 V, n = 512), Depression (200 ns, -2.75 V, n = 512), Vread (200 ns, -2.75 V, n = 512). (1 V, 200 ns).

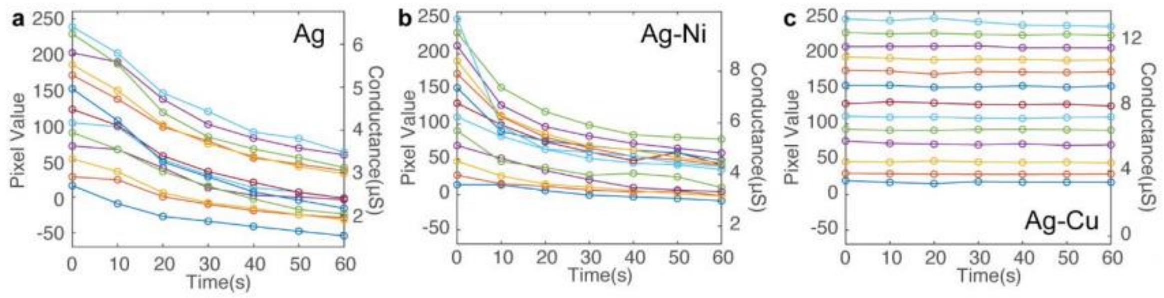

Figure 2-25: Statistical examination of the retention measurement for each picture pixel in Fig. 2-4d. Pixel values in 256-level grayscale images are uniformly distributed into 13 groups (e.g. 0-19, 20-39, etc.). For each time step and alloy device (a) Ag, (b) Ag-Ni, and (c) AgCu, the average values of the pixel groups were computed. The grayscale picture values mapped from the conductance values are shown on the left y-axis of the images, while the actual conductance of the device is shown on the right y-axis.

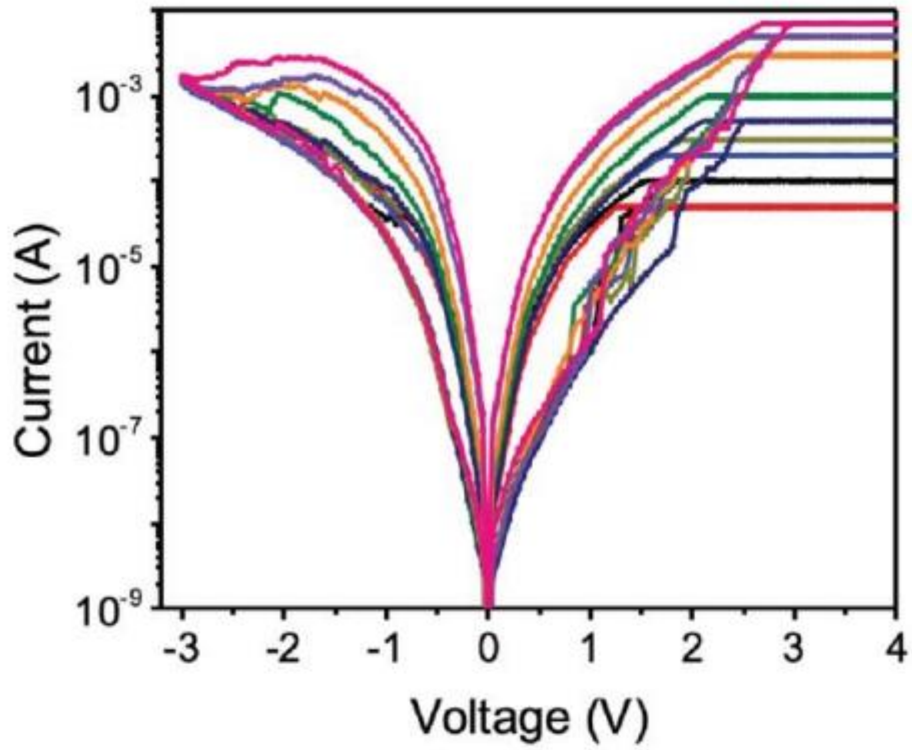

Figure 2-26: The I-V characteristics of an Ag-Cu memristor with varying compliance currents. With a compliance current of less than 100 A, the device can be programmed indefinitely.

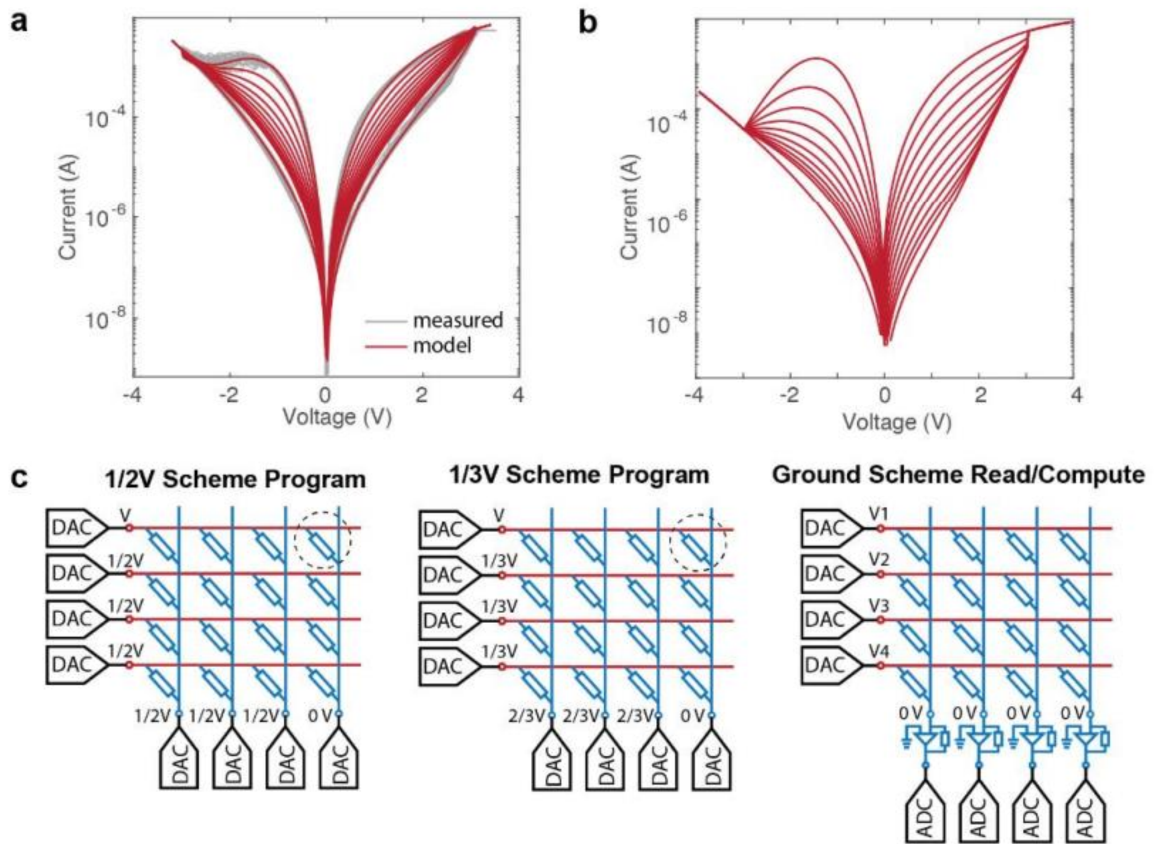

Figure 2-27: SPICE simulation configuration. (a) IV Ag-Cu behavioral model parameters compared to measured device performance (b) An Ag-Cu model with a reduced OFF state conductivity. (c) Schematics of the 1/2V and 1/3V write schemes, as well as the ground read technique, employed in array operation and SPICE simulation.

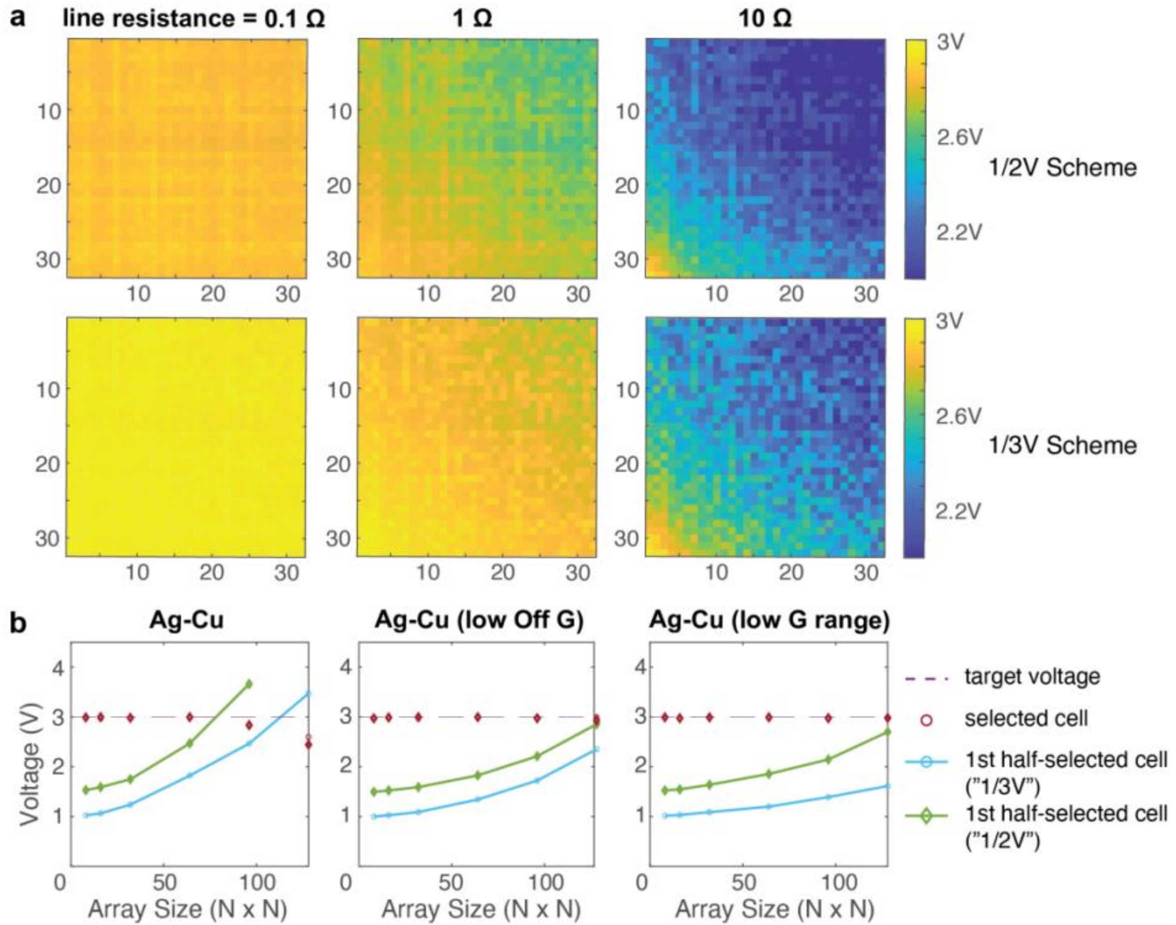

Figure 2-28: For the write process, SPICE simulation is used. (a) Voltage delivery maps for Ag-Cu arrays with various line resistances and biasing techniques. A 3V bias is given to the source terminals from the left and bottom terminals. Reduced voltage biases applied across device connections in the 32 × 32 array are indicated by color gradients. (b) Simulation of the write process to send 3V to the worst-case cells in arrays, i.e. the cell closest to the source, such as the top right cell in Figure 1. (a). Arrays of various sizes were evaluated. The green and blue lines represent extracted voltage biases at the half-selected cell nearest to the source. When the voltage bias of a half-selected cell exceeds the writing threshold, write disturb occurs (e.g. 3V). The simulations were run for three alternative device conditions: the original Ag-Cu model, the modified Ag-Cu model, and the Ag-Cu model that only employs the lower 10% of conductance

ranges. Improved scalability for the write process is obtained by using a better device model or decreasing the conductance ranges of Ag-Cu memristors.

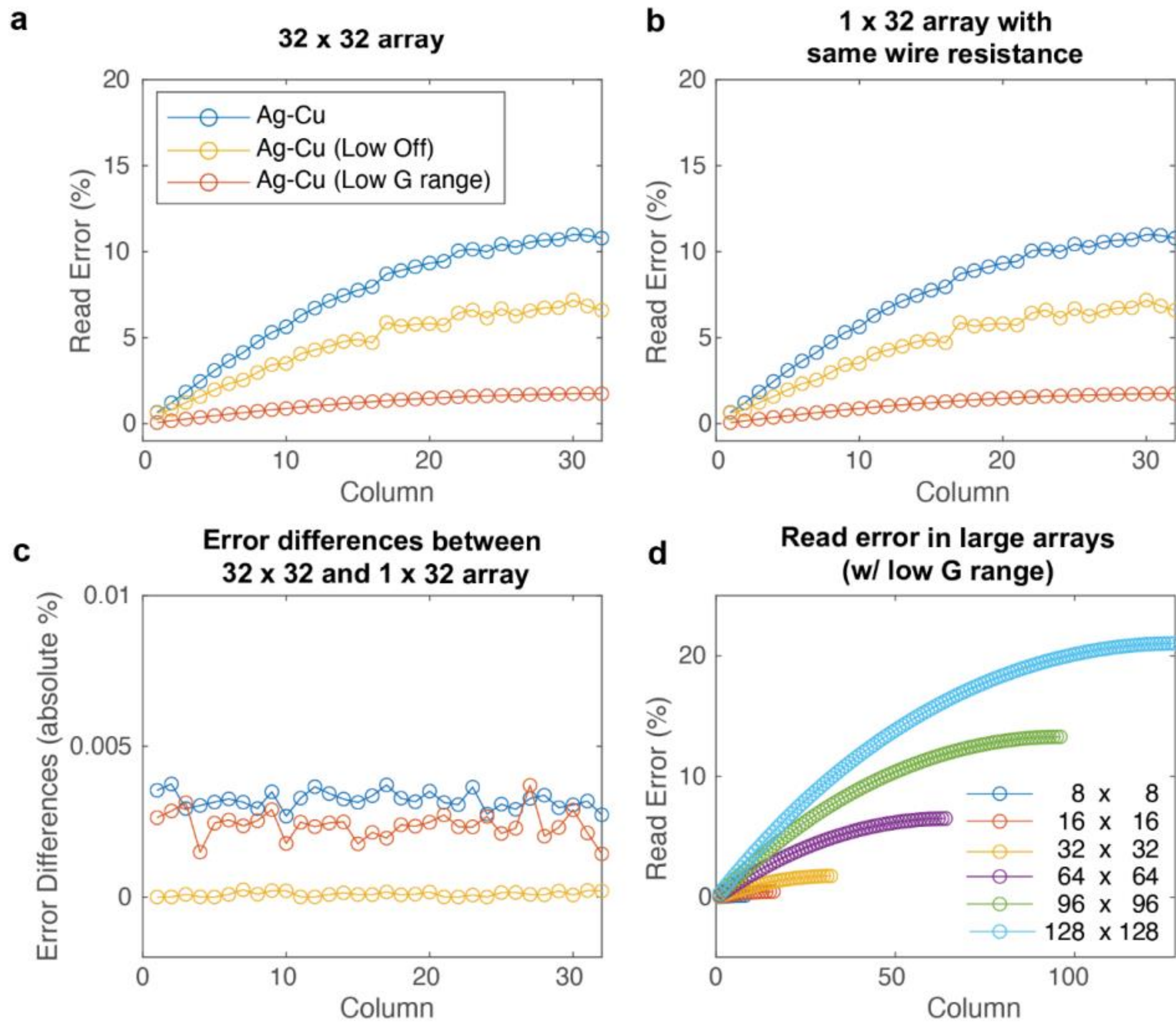

Figure 2-29: For the read process, SPICE simulation is used. (a) Average read error in 32 × 32 arrays. The read error is defined as the difference between the inferred device conductance and the actual device conductance based on the TIA readout. The simulation findings indicated that when cells got further away from the source, their reading mistakes increased, which was caused by line resistance. Three alternative device conditions, similar to those utilized in write simulation, were also investigated here. Limiting device conductance was discovered to be beneficial in lowering the influence of line resistance. (b) Read error simulation in a 1 × 32 array with the same wire resistance as in (a) to rule out the effect of sneak pathways. The simulations produced comparable results (a). (c) The read error differences between (a) and (b) were calculated and revealed extremely modest changes, demonstrating that line resistance is

the most important element during the read process using the ground read scheme. (d) Simulation of read errors throughout the read process in various array sizes using an Ag-Cu device model with a limited (lower) conductance range.

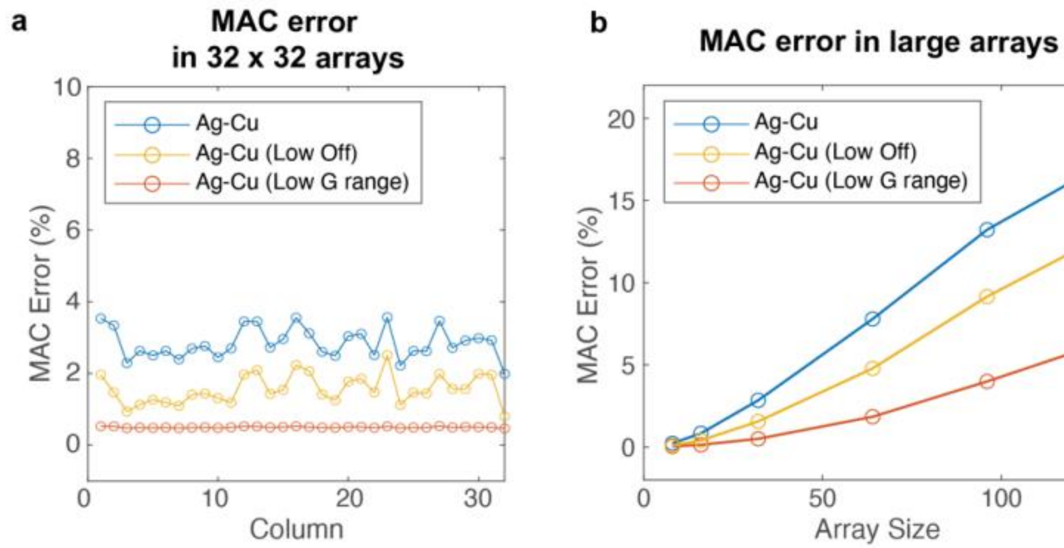

Figure 2-30: SPICE simulation is used in computing. (a) Simulated multiply-accumulate (MAC) error in 32 32 arrays. The computed MAC value of input voltage vectors and the conductance of programmed devices (values derived from TIAs) are compared to the array's actual MAC output. (b) Simulation of MAC error in various array sizes and device conditions.

Process	Cu E_a value (eV)	Ag E_a value (eV)
Cation dissolution from anode (Cu^+ and Ag^+)	0.52[27,28]	0.8[27,28]
Cation diffusion in Si (Cu^+ and Ag^+)	0.43[13]	1.15[13]
Cation reduction at nucleation site	0.45	0.5
Ag-Cu cluster growth	0.55	0.65
Atom oxidation from Ag-Cu cluster	0.98, 1.08, 1.18, 1.5 respectively related to the atom with 1, 2, 3, 4 bonds	1.0, 1.1, 1.2, 1.5 respectively related to the atom with 1, 2, 3, 4 bonds
Atom oxidation from nucleation site	1.45	1.25

Table 2-1: Summary of KMC simulation parameters.

Parameters	**Values**
Line Resistance	0.1Ω, 1Ω, 10Ω (cell-to-cell)
Array Dimensions	8 × 8, 16 × 16, 32 × 32, 64 × 64, 96 × 96, 128 × 128
I/O biasing schemes	Write: 1/3 V, 1/2 V, Read/VMM: Ground
Array Weight Distributions	Device states ∈ U(0,1) or U(0,1/10) (normalized)
Selected Device Weights	0, 1/2, 1 or 0, 1/20, 1/10 (normalized)

Table 2-2: Summary of simulation parameters

Supplementary Note 2. Design considerations for array fabrication with Si electrode

Our alloy memristor array made use of a single crystal Si electrode, which provides good switching performance and a high device yield. However, the Si electrode is more resistive than metal, making it unsuitable for use as long interconnects. For large array operations, ensuring low line resistance is crucial (see supplementary Note 4 for more details). To address this issue, we employed a modified crossbar architecture for our alloy memristor array, as shown schematically in Fig. 2-8. To lower line resistance, a metal capping layer of 100 nm thick Ti/Au layer was placed on top of the array's long p^+ Si bottom electrodes (see Methods). As a result, we were able to reach 1.5 cell-to-cell line resistance, which is sufficient for our 32 × 32 arrays' operations. This method may be even more effective in dense designs with 10 nm features, where nanometer silicon line patterns would be exceedingly resistive. The usage of this arrangement necessitates a somewhat bigger size (~6-8 F^2 cell), yet it is still a dense design with numerous performance benefits. When using a more complex foundry process, the arrangement could be further optimized. Meanwhile, additional engineering efforts, such as narrowing device conductance ranges, as demonstrated in Supplementary Note 4, may be useful in compensating for high line resistance, particularly in nano-scale arrays. In the meantime, the epitaxial formed p+ Si bottom electrode employed in this work is not a prerequisite for the alloy memristor. Bottom electrodes can also be made of other types of single crystal p^+ Si films with corresponding doping concentrations. We have, for example, manufactured devices using as-received p^+ Si wafers and demonstrated equivalent performance. We chose high temperature epitaxial deposition for the p^+ Si layer deposition because it was the most convenient option for us to deposit a specific p^+ Si layer on top of the SOI wafer with different doping concentration for the device layer. Any Front-End-Of-Line (FEOL) compatible doping approach, such as ion implantation, can be used to replace the procedure.

As a result, our memristor's fabrication technique is compatible with integration with CMOS circuits, which is essential for a completely integrated system.

Supplementary Note 3. Array operation at reduced conductance ranges

While wide dynamic ranges may imply more resolvable states for computing, the high current required during programming will use a significant amount of power. To address this issue, we chose to limit array operations to the lower conductance ranges. The DC characterisation of our device with varied current compliance is shown in Fig. 2-26. The device can be stably programmed at less than 100 μA while maintaining a sufficient operation window. Operating devices with a limited conductance range can nonetheless achieve good programmability for computing, as evidenced by the consistent analog switching behavior in Fig. 2-3 and extremely reliable array demonstrations in Fig. 2-4. Reduced programming current/power may be advantageous for future on-chip memory and computer applications. In addition, our SPICE simulation (see Supplementary Note 4) reveals that integrating low device conductance in arrays is more advantageous in dealing with sneak path and line resistance difficulties. Nonetheless, the real conductance ranges can be adjusted further because low device conductance and less dynamic range may introduce greater latencies and impair MAC accuracies due to non-ideal circuit conditions such as readout circuitry input offsets and thermal disturbances.

Supplementary Note 4. Analysis of the impact of line resistance and sneak paths in alloy memristor arrays

SPICE (Simulation Program with Integrated Circuit Emphasis) simulation was used to analyze the performance of alloy memristors in large-scale arrays utilizing the behavioral device model of Ag-Cu memristor. Figure 2-27a depicts the measured device's I-V characteristics (gray color) as well as the SPICE model (red color). In the investigation, additional device models and

conditions were also used. First, a device model with similar I-V characteristics to an AgCu memristor but lower OFF state conductance was derived from the Ag-Cu model (i.e. lower mean conductance and higher On/Off ratio); the modified model's IV-plot is illustrated in Fig. 2-27b. On the contrary, we investigated a situation in which the devices were only operated at the lowest 10% of their complete conductance ranges (i.e., lower mean conductance and lower On/Off ratio). In addition to multiple device models, different electrode line resistances, array dimensions, and I/O biasing techniques were used to analyze their effects during the Write, Read, and Multiply-Accumulate (MAC) operations. Table 2-2 summarizes the parameters utilized in simulations. We began by evaluating the array's performance during the Write operation. Because of sneak paths and nonzero line resistances, the voltage bias across the device junction may deviate significantly from the voltage bias applied between the selected row/column source terminals, resulting in reduced voltage delivery efficiency, defined as: Voltage delivery efficiency = $V_{junction(i,j)} / V_{source} = V_{junction(i,j)} / (V_{row(i)} - V_{column(j)})$ To simulate the write operation, all devices in a 32 × 32 Ag-Cu memristor array were programmed with a randomly generated weight distribution. All devices were written in succession, and the row/column source terminals were suitably biased using either a 1/3V or a 1/2V method (as schematically illustrated in Fig. 2-27c). As shown in Fig. 2-28a, the actual voltage biases across the selected device junctions were recovered and used to plot the voltage delivery map. The color gradient seen in the figures implies a significant voltage drop across the arrays. According to the simulation results, lower line resistance is crucial for ensuring strong array programming capabilities. The presence of non-zero line resistance would not only result in a direct voltage drop across the electrodes, but would also have a significant impact on the biasing precision of the 1/3V and 1/2V methods, limiting their effectiveness in suppressing sneak current. Sneak pathways had an impact as well, as evidenced by the color variation related with the randomly created conductance map for 1/3V and 1/2V schemes. Higher source voltages are necessary to

compensate for the voltage drop due to the reduced voltage delivery efficiency. As the source voltage biases increased, so did the voltage biases across those half-selected cells (cells that share at least one electrode with the selected cell). When any half-selected cell obtains voltage bias that above the switching threshold and is unintentionally programmed, the write disturbance occurs. To quantify the scalability of an Ag-Cu array, we simulated the Write process with various array dimensions, aiming to supply 3V (the device's set threshold) to the worst-case cell. From each simulation, the junction voltages of the selected cell (the one closest to the source) and the first half-selected cell (the one closest to the source with the least voltage drop) were collected. The graphs of actual junction voltages of selected cells and first half-selected cells over varied array dimensions are shown in Fig. 2-28b. The write failure occurs when the voltage bias of the first half-chosen cell exceeds the voltage bias of the selected cell. Based on existing array circumstances, our simulation reveals that Ag-Cu memristor arrays with dimensions up to 64 × 64 can be operated with enough margins below probable write failure. To boost array scalability even further, memristors with low conductance would be preferable. This might be accomplished through device engineering to create more resistive memristors, or by limiting the conductance of the memristors to its lower conductance range. Further simulations were performed to support our hypothesis, using either the modified Ag-Cu model with low off state conductance (high ON/OFF ratio and lower mean conductance) or solely the bottom 10% conductance range of the current Ag-Cu model (low ON/OFF ratio and lower mean conductance). Both simulations, as shown in Fig. 2-28b, demonstrate reduced write disturbances independent of On/Off ratio, implying that a high contrast between line resistance and device resistance is required to ensure robust programming in passive selector-less arrays. Meanwhile, a simulation based on the "Ground" read technique was run to evaluate the array performance during the Read and MAC procedures. The "Ground" read scheme is a realistic way for suppressing the sneak path current, and it was used for reading and calculating here.

The "ground" system is depicted schematically in Fig. 2-27c. Because all of the columns were grounded and had the same voltage potential, the leakage current that flowed between them was considerably reduced. However, for the ground read scheme to work properly, low line resistance is required so that the ground biases can be appropriately applied to all of the cells in the array. In this simulation, we chose a line resistance of 1 Ω, which was near to our actual array circumstances. As shown in Fig. 2-29a, the read accuracy of the alloy array was investigated by measuring the read errors from all cells in the 32 × 32 array. The read operation was simulated by applying a read voltage (i.e. 1V) to the array row by row. The column output current can be used to deduce and estimate the conductance of the devices on the specified row at each read cycle. The error % was calculated by comparing the estimated and real conductance of the devices. The average read error of each column in the 32 x 32 array with different device models is shown in Fig. 2-29a. The simulation results show that when the columns become further away from the source, the read error increases due to line resistance. To reduce read error, use device types with low Off state conductance or limit the conductance range. On the other hand, we investigated the efficacy of the ground strategy in suppressing the sneak path current. This is investigated further by simulating the read process in a 1 x 32 array with the same row and column line resistance as the 32 × 32 array, as seen in Fig. 2-29b. The absolute difference in error % between the 32 × 32 array and the 1 × 32 array was calculated and presented in Fig. 2-29c, revealing only very tiny variations. The results indicated that the nonzero line resistance of the electrode was the primary cause of read error in our process. As a result, lowering line resistance or employing low device conductance are favored ways for improving reading accuracies, particularly in large arrays. The read errors from a varied array size in the optimum low conductance region of an Ag-Cu memristor were simulated further and are illustrated in Fig. 2-29d. Finally, we assessed the computation accuracy of multiply-accumulate (MAC) processes. The MAC procedure includes introducing arbitrary voltage

vectors to the array, which are then multiplied by the array's programmed conductance map. However, because of the significant impact of line resistance, the programmed conductance of each cell deviates from the target conductance, resulting in the MAC error. The MAC error was defined as the difference between the MAC values detected from the peripheral output and the intended MAC values estimated based on the devices' input voltage amplitudes and desired conductance values. It is important to note that the target conductance in this case should be the conductance value sensed from the peripherals during the programming step, not the cell's genuine conductance value. Figure 2-30a depicts the simulated MAC faults in a 32 x 32 array. Our simulation reveals that using devices with low Off state conductance or employing a low conductance technique are both effective methods for achieving high accuracy computation. The average MAC error for various array sizes is further simulated and depicted in Fig. 2-30b. Larger array implementations may be achievable based on our simulation results, but they would necessitate considerable modifications of the operating circumstances, such as lowering line resistance, employing a more durable 1/3V biasing method, and utilizing a low conductance range. Improving the device and array fabrication processes is also critical for increasing array size and/or computing precision.

2.3 Conclusion

We believe that alloying the conduction channels in ECM devices will fundamentally solve the ‘tunability–stability dilemma' between robust weight adjustment and long-term stability in our quest to identify the ultimate device for neuromorphic computing. The proper engineering of interaction and migration of alloying elements in conducting channels allows for a great deal of flexibility in tailoring the electrical performance of the devices. This allowed us to create large-scale transistor-free crossbar arrays capable of storing and inferencing neural network

weights. We believe that our alloy design guideline for reliable memristor performance may be extended to other material systems in order to optimize conduction channels and switching dynamics for improved performance in neuromorphic computing applications.

Chapter 3

Stackable hetero-integrated chips for reconfigurable neuromorphic computing

This chapter introduces the stackable hetero-integrated chips for hardware-wise neuromorphic computing. This chapter will be published with a temporary title 'Stackable Hetero-Integrated Chips for Reconfigurable Edge Neuromorphic Computing', Chanyeol Choi[†], Hyunseok Kim[†], Ji-Hoon Kang[†], Hanwool Yeon[†], Celesta S. Chang, Junmin Suh, Jiho Shin, Yeongin Kim, Haneol Lee, Doyoon Lee, Sang-Hoon Bae, Hyun Kum, Jaeyong Lee, Ikbeom Jang, Peng Chen, Wenqiang Zhang, Peng Yao, Subeen Pang, Kanghyun Ryu, Yifan Nie, Hang Chi, Jagadeesh Moodera, Huaqiang Wu[*], Peng Lin[*], and Jeehwan Kim*. ([†]co-first authors.) The abstract of work is presented below.

Due to the increasing number of sensors and the quantity of sensory information being collected at an exponential pace in this age of edge computing, modern sensor computing systems have been receiving greater attention. Additionally, as deep learning advances, parallel data processing becomes increasingly necessary to manage the massive amounts of data generated by artificial intelligence. It has been proposed to use 3D heterogeneous integration in conjunction with improved packaging technologies in order to decrease time delay and increase data bandwidth, which have been hampered by data transfers between sensors, memory, and CPUs. Despite the fact that 3D integration technologies can provide a multi-

functional hardware platform for edge sensor-computing systems, dealing with a large amount and a variety of sensory information in real-world scenarios remains difficult due to a lack of hardware-wise reconfigurability and hardware constraints such as 2D dataflow. In this paper, we propose stackable hetero-integrated chips with light communication for reconfigurable modularity and neuromorphic core for parallel data processing. We present three different materials systems for hetero-integrated chips, each with its own advantages. In particular, silicon-based memristor crossbar arrays for neuromorphic computing, gallium arsenide-based photodiodes (PD), and indium gallium phosphide-based light emitting diodes (LED) for chip-to-chip communication are among the technologies being developed. Using replaceable and stackable hetero-integrated circuits, we show the robust kernel operations using optoelectronic devices (such as PD and LED) and memristor crossbar arrays to illustrate the robust kernel operations. For a more in-depth demonstration of our sensory input processing in a noisy environment, we develop and integrate a denoising module to deal with noisy pictures, which leads to improved letter recognition accuracy. As a result of our hetero-integrated chips' hardware-wise reconfigurability, vertical scalability of diverse functional layers, and non-von-Neumann computing, we expect sensor-computing systems to be more versatile in their ability to accommodate the complexity of any neural network in edge computing.

3.1 Introduction

Edge computing is made possible by an almost endless number of edge devices located near data sources, each of which produces an enormously vast amount of data. The demand for edge computing, which combines sensors and processors located in close physical proximity to one another, is rapidly expanding in order to decentralize data processing [73]–[75]. Furthermore, data-driven artificial intelligence (AI) applications have fundamentally altered

the landscape of computer designs, with the elimination of unnecessary data movement and the promotion of parallel data processing becoming critical steps in achieving low latency and high energy efficiency for real-time AI applications while maintaining high performance [76]–[78]. Because it allows for the smooth integration of various functional layers such as sensors, processors, and memory, 3D heterogeneous integration technology makes the deployment of the edge computing paradigm possible [79]–[82]. Despite this, there are three key issues with the present generation of 3D hetero-integrated chips for edge computing, which are discussed below. Firstly, due to the physical connection of functional layers such as sensors and processors to one another, they are not replaceable, and a new chip must be created whenever the other functional layers are required. Second, because to the high temperature techniques that are used in series on 3D hetero-integrated circuits for edge computing, they may result in low device yield. Lastly, the employment of the von-Neumann technique in 3D integrated systems has a limited ability to accelerate data transfer because of a memory bottleneck and a lack of parallel data processing capability.

3.2 Contributions and Methods

It is demonstrated in this paper that reconfigurable heterogeneous integration can be achieved by using stackable chips that include embedded artificial intelligence. Each chip layer comprises a hetero-integrated optoelectronic device-based communication unit, which makes it possible to change, insert, stack, and restack the layers in a way that is not conceivable with any other method of fabrication. Different chips might be combined into a vertically aligned chain of processors, taking full advantage of the area communication interfaces and seamless connections that are possible when chips are placed in close physical proximity to one another. Because of this, the chips can be reconfigured when the external inputs change, if additional

processing capabilities are required, or when the input modality changes. This flexibly reconfigurable architecture can be utilized to execute a variety of jobs and change over the course of a person's lifelong learning process, among other things. In addition, we implemented neuromorphic computing processors into the chip by embedding memristor crossbars into each layer of the chip's architecture. As a deep learning hardware accelerator, they convey analog data quickly and accurately to photo-sensors nearby, allowing for faster deep learning training. In a free space, we were able to demonstrate successful communication between each of the chip layers thanks to our innovative chip architecture. The eye layer of the chip generated the initial data, which was then processed using the memristor-based hardware accelerators. The eye layer of the chip reacted to external optical stimuli to generate the initial data. It was also possible to test and validate the reconfigurability of our hetero-integrated circuits using our architecture because of the way they were designed. Sensor input patterns from a variety of sensors were efficiently detected and processed with significant flexibility by replacing or stacking pre-trained chip layers. In addition, we were successful in proving an additive feature by developing and demonstrating a denoising processor on the chips, which allowed us to accurately classify images that had been distorted. We are convinced that the reconfigurability of varied functional layers will provide sensor-computing systems with greater versatility in order to adapt to the complexity of any sensing data or artificial intelligence assignment.

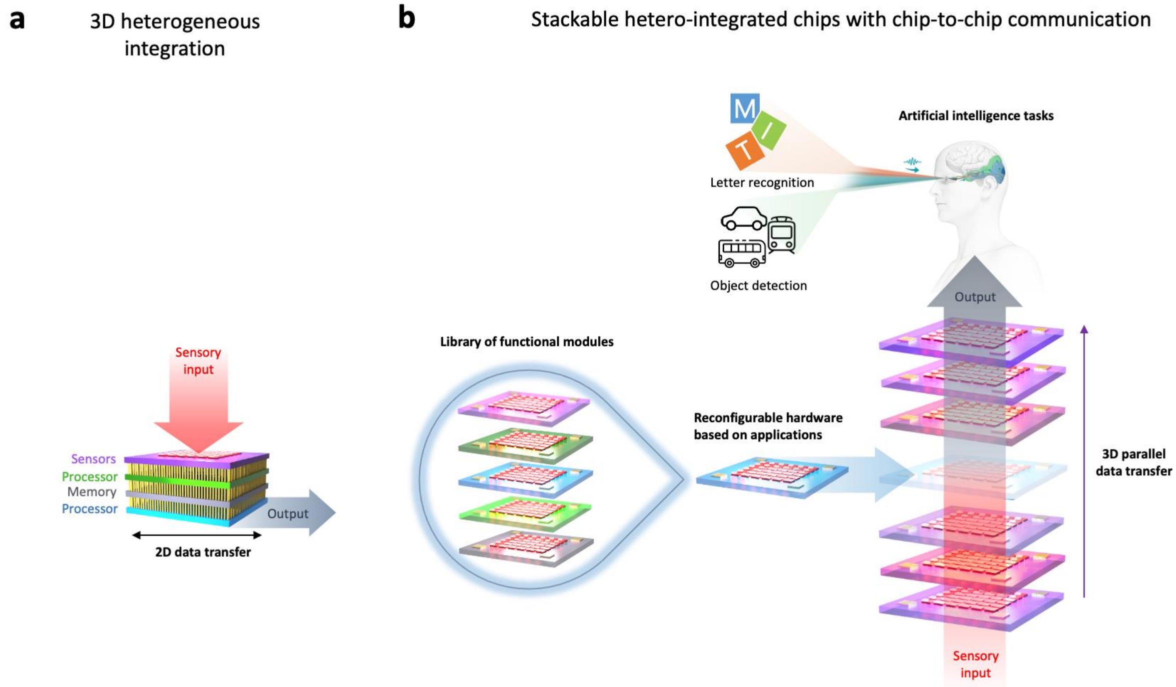

Figure 3-1: Integration technologies for sensor-computing systems for use in edge computing applications. a, Schematic representation of 3D heterogeneous integration for a sensor-computing system. A single material system contains physical wires (in yellow) that connect different device layers such as sensors, processors, and memory to one another (for example, silicon). Vertical interlayer vias are used to physically configure the three-dimensional structure. A limitation of functionality and chip stackability can be attributed to the hardwired connections. b, Schematics of stackable hetero-integrated chips with chip-to-chip communication for sensor-computing systems with chip-to-chip communication. This demonstrates the use of light communication between chips, which enables a high degree of freedom for hardware-based reconfigurability at the sensor layer and processor layer, among other places. Optoelectronic devices such as light emitting diodes (LEDs) and photodiodes are used to convey light information between the layers. Each layer is physically in contact with the other. Depending on the application, any functional module from a library of functional modules can be selected and assembled into a heterogeneously integrated chip stack, which is then tested. The first layer of the chip stack is dedicated to the processing of sensory input, and

the input can be processed in parallel over the 3D hetero-integrated chip stack as the chip stack is built. Pre-programmed neuromorphic processing cores in the stacking hetero-integrated circuits enable them to implement various artificial intelligence applications for edge computing, such as letter recognition and objection detection.

Traditional heterogeneous-integrated chips and our advanced, reconfigurable heterogeneous system are depicted in Fig. 3-1, which also summarizes the designs of both systems. However, as previously stated, the existing state of the art heterogeneous integration system (Figure 3-1a) is not yet mature enough to solve the key flaws listed below [75], [83], [84]. For starters, standard heterogeneously-integrated circuits are unable to respond to changing external stimuli since the sensors and processors are fixed after heterogeneous integration and cannot be changed to accommodate new or changing environments. For the second time, device dependability is susceptible to degradation because to the close spacing between layers and high process temperature during post-processing. Third, because von-Neumann computing makes use of both the processor and the memory for data processing, the response time and data bandwidth are both constrained by the superfluous data transfer of sensing data. Fig. 3-1b shows how our stacking hetero-integrated chips may totally overcome the difficulties mentioned above in conventional hetero-integration systems. First, with the standalone hetero-integrated chips, the chips are now removable and stackable, allowing them to respond efficiently to a variety of scenarios. Second, we bound optoelectronic systems in each layer so that the individual freestanding chip layers may be produced separately, allowing light communication across layers to be enabled without affecting the chip production process flow. Third, we also consolidated the CPU and memory into a single AI component by embedding an artificial intelligence system consisting of memristor arrays in each layer of the hetero-integration units. This allowed for efficient and speedy data transport. As shown in Figure 3-1b, we have configurable hetero-integrated systems with artificial intelligence and optoelectronics, which include photodiodes and light emitting diodes (LED). Each AI- and optoelectronic-embedded layer, similar to a Lego block, enabled us to (i) replace either the processor or the sensor, depending on sensing or computation requirements, (ii) stack layers to enhance neural network tasks, and (iii) add or delete different layers to enhance the function.

The following are examples of the above cases: (i) different types of sensors could be easily replaced or different pre-trained computing layers could be replaced to recognize various sensory input; (ii) trained computing layers could be continuously stacked for highly parallel kernel operations to recognize varying and various input from sensor layers; and (iii) computing layers could be added for heavy processor unit usage. In order to meet specific requirements, hetero-integrated chips can be tailored according to their intended use and the sensing modality employed and one can expect following outcomes: (i) Due to the absence of wire interconnects, it is possible to easily reconfigure functional layers; (ii) the limitations on fabrication processes, such as thermal budget, can be alleviated; and (iii) the spectrum of neural network functions, such as increasing/reducing kernels and modulating depth of neural networks with cross-modality sensory information, can be broadened by different modality sensor devices for multifocal sensors as described in Figure 3-5.

In this study, we demonstrate a separate kernel operation by replacing computing layers with pretrained crossbar arrays in the computing layers, as well as a parallel kernel operation by stacking layers with crossbar arrays in the computing layers. It is reasonable to predict that when the kernel (designed conductance of the memristor, G) is matched to the input picture (voltage input, V), we will have the highest possible current total value ($I = \Sigma (V \cdot G)$). All of the various layers are assembled and disassembled in order to demonstrate a range of chip stacks in different combinations.

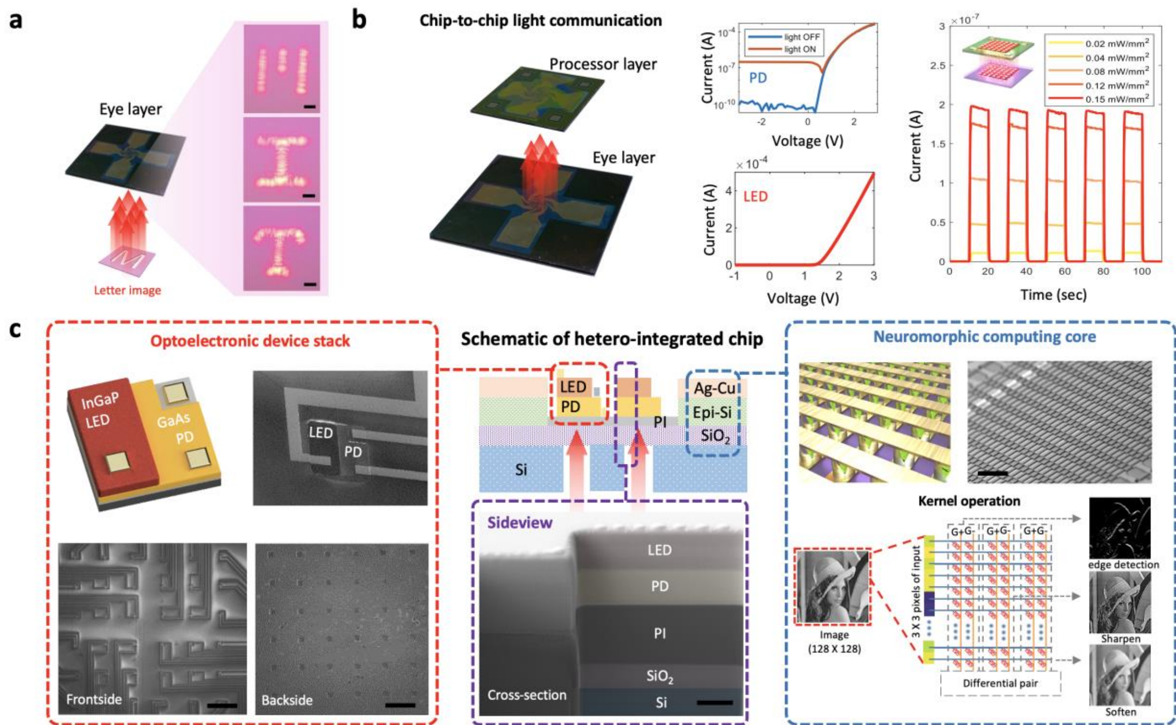

Figure 3-2: Components of stackable hetero-integrated neuromorphic chips. a, an optical image of a light sensing layer (eye layer) with light input from the bottom. Patterned photomasks and a laser diode with a confocal setup are used to create three different letter patterns: M, I, and T. The eye layer receives the letter images. Scale bar: 1 mm. A photodiode measures the intensity of light passing through a transparent silicon oxide membrane. b, optical photo of a processor layer and an eye layer. To communicate, two chips are physically in contact. The eye layer's light input can be passed on to the next layer. Each layer has a photodiode/LED stack that allows them to receive and transmit light data. I-V characteristics of LED and photodiode devices are presented SEM images with light OFF and ON (~ 0.15 mW/mm2) for the photodiode. At 0.1 V reverse bias, photodiode response to LED light in chip-to-chip communication as a function of LED light intensity. The chip-to-chip communication is depicted in the inset. c, a diagram of a hetero-integrated chip. Illustration and SEM images of optoelectronic device stack (photodiode/LED, scale bar: 100 µm), neuromorphic computing core (Ag-Cu alloy-based Si memristor crossbar array, scale bar: 100 µm), and sideview of chips (scale bar: 1 µm) are presented. We used deep reactive ion etching to align an array of

photodiode/LED stacks with backside holes for an optoelectronic device stack (DRIE). We made 32 × 32 memristor crossbar arrays for the neuromorphic computing core, as shown in the schematic and SEM image. On a 128 × 128 pixel image, three different types of 3 x 3 kernel operations (edge detection, sharpen, and soften) are performed.

Hetero-integrated chip modules

To prepare assembly-ready freestanding layers for reconfigurable hetero-integrated chips, we first developed a light sensing layer (eye layer) and a processor layer separately. At the eye layer, sensor arrays are connected to LEDs where the sensors communicate external stimuli while LEDs transfer information to the next layer. The processor layer is composed of a communication section and an AI section. The communication section contains photodetectors for receiving light signals from the sensor layer and LEDs to send information to the next process layers, while the AI section has arrays of memristor crossbars for computing. In this work, we utilized an eye layer as a representative case for a layer that senses one of sensing modalities. The eye layer is purposely designed to capture visual information, as shown in Fig. 3-2a. In the eye layer, photodiodes (PDs) receive light input (letter images) from the bottom through holes and transfer the information to the next layer using LED devices. Inset of Fig. 3-2a illustrates that three different light inputs are generated by patterned photomasks and sensed by photodiodes in an eye layer. The letter patterns were generated in our confocal optical setup (See Fig. 3-6 for details). In this work, we used three letter patterns, namely ‘M’, ‘I’, and ‘T’. As shown in a stack of two chips (Fig. 2b), we have demonstrated chip-to-chip light communication enabled by stacks of optoelectronic devices (i.e., LEDs and photodiodes) attached to the layers. Each layer has 6 × 6 pixels of the LED/Photodiode stacks and their I-V characteristics are presented in Fig. 3-2b. I-V characteristics of 6 × 6 photodiode array and 6 × 6 LED array in the eye layer are presented in Fig. 3-7. While GaAs photodiodes are positioned on the bottom of each stack receiving light input, InGaP LED devices are positioned on the top of each stack and send light information to the next layer. LEDs of the bottom layer (eye layer) can communicate with photodiodes of the top layer (processor layer). We presented the response of a photodiode to the LED light in chip-to-chip light communications as a function of light intensity. With varying LED operating voltages, the output currents of photodiodes

were measured at 0.1 V of reverse bias. At the maximum LED light intensity of ~ 1 mW/mm^2, the output current is around 1×10^{-6} A. We confirmed that there is no crosstalk between each LED/Photodiode stack in chip-to-chip light communications with carefully aligned frontside LED/Photodiode stacks and backside holes. In this way, photodiodes arrays successfully receive patterned light inputs through aligned holes and transfer the information to the next layer using LED devices.

Fig. 3-2c shows schematic illustrations and scanning electron microscope (SEM) images of a hetero-integrated chip. They are two main components: (1) optoelectronic device stack including GaAs photodiodes and InGaP LED and (2) neuromorphic computing core which is a 32 × 32 Ag-Cu alloy-based Si memristor crossbar array. As shown in the optoelectronic device stack, each LED/Photodiode stack has three contacts to operate the LED and photodiode separately. SEM images are taken for a single optoelectronic device stack and its 6 × 6 array. The front- and the back- sides of the array are presented. The square-shaped holes on the backside of the chip are precisely aligned with the LED/Photodiode stacks on the frontside so the light from the back can be sensed by photodiodes through a transparent SiO_2 layer. The details of the optoelectronic device fabrication processes on a silicon-on-insulator (SOI) wafer can be found in Fig. 3-8 and Methods Section. We provide the sideview of our hetero-integrated chip using a cross-section SEM image. Pseudo-colored SEM images show each component of heterogeneously integrated optoelectronic device stack on a silicon-oxide-insulator (SOI) wafer.

In addition, 'processor layer' was created to accept the light information from the bottom layer and to perform AI computer operations. To do so, the processor layer has another component called 'neuromorphic computing core'. The inset of 'neuromorphic computing core' shows an illustration and a SEM image of memristor crossbar arrays. We fabricated the Ag-Cu alloy-based Si memristor crossbar arrays using the method from ref [39]. After receiving light

information from the photodiodes, we have performed neuromorphic computing operations (e.g., kernel operation and fully connected operation) onto Ag-Cu alloy-based Si memristor crossbar arrays. For more information about our measurement setup for memristor crossbar arrays, see Fig. 3-9. We also performed 3×3 kernel operations on a 128×128-pixel image. The input pixel information is flattened and converted to voltage pulses before reaching the memristor devices that have programmable resistance values. Each kernel requires 9×2 memristors to represent both positive (G^+) and negative (G^-) weights which allow for both a sign and a magnitude. The current values are generated by the sum of the dot product of each row's input voltage (V) and each row's conductance level (G^+ or G^-), as defined by Kirchhoff's law, $I = \Sigma (V \cdot G)$. As a result, the current values are subtracted to express their weights ($G = G^+ - G^-$). We programmed linear kernel operations into crossbar arrays and confirmed the weights by reading conductance values from each device. We successfully performed the three 3×3 kernel operations of vertical edge detection, sharpen, and soften for image processing.

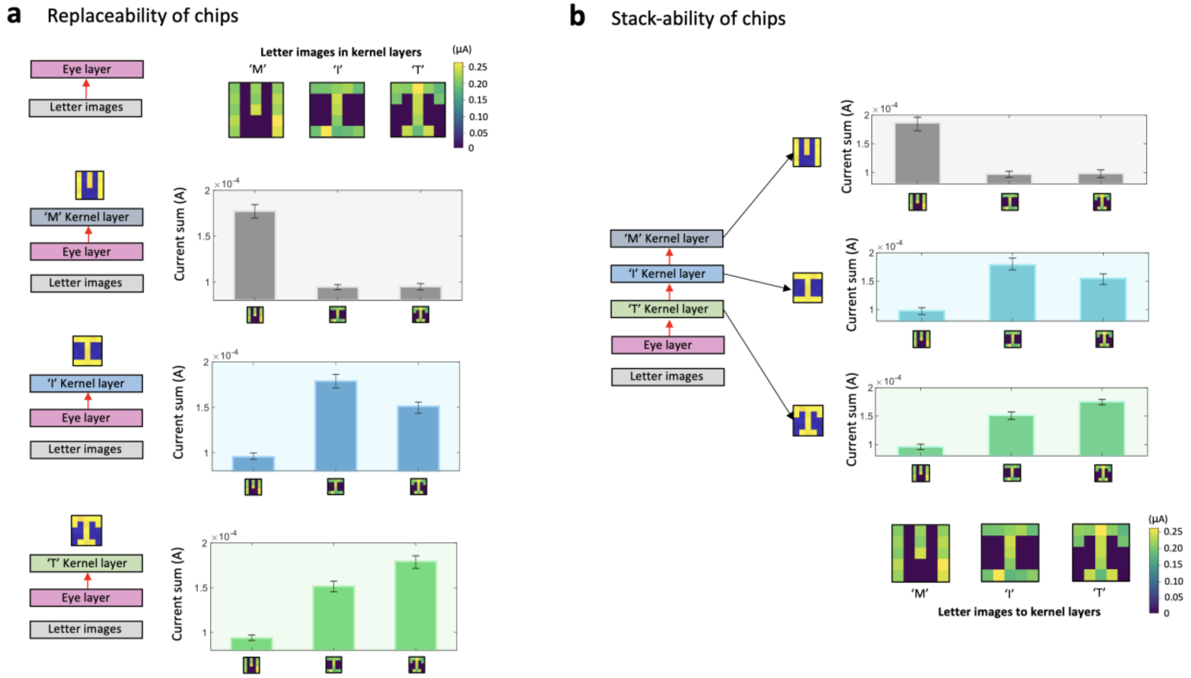

Figure 3-3: Hetero-integrated neuromorphic chips that are replaceable and stackable as well as their robust kernel operations. a, Replaceable neuromorphic chips for kernel operations. Patterned images produced by photodiodes in kernel layers with a 0.1 V reverse bias are shown (Top row). The stacking and replacement of hetero-integrated chips is depicted in block diagrams. Different patterned images have had kernel operations performed on them by replacing a kernel layer. Each kernel operation's current sums are shown in the graph. The maximum and minimum values of the current sums are indicated by error bars. b, The multi-layer neuromorphic chip stack is depicted in a block diagram. Pre-programmed memristor crossbar arrays process the input letter images, which are shared across three kernel layers. The current sums are the result of kernel operations on the letters in the input images below (Bottom row). Kernel operations were performed in parallel by stacking three different kernel layers. The maximum and minimum values of the current sums are indicated by error bars.

Letter recognition with neuromorphic computing modules

We initially programmed three different 5 × 5 kernels (the ‘M' kernel, ‘I' kernel, and ‘T' kernel) in three different processor layers (the processor ‘M' layer, processor ‘I layer, and processor ‘T' layer), as shown in Fig. 3-3a, for the letter recognition task. We then tested the performance of the three different 5 × 5 kernels on the letter recognition task. In Fig. S3-10 and the Methods Section, the specifics of the 5 × 5 kernel operation on memristor crossbar arrays are described in greater detail. Fig. 3-3a shows the three patterned images that are sensed by photodiodes in the processor layers, which are shown in the top row. (See Fig. 3-11 for an illustration of the patterned images sensed by the eye layer and the filtering effect produced by the processor layers.) We were able to perform kernel operations by replacing three processor layers with the help of these images. The output current values of the three pattern images are converted to the number of read pulses of 0.5 V by multiplying them by the number of pattern images. The current sum is accumulated over the number of read pulses and is then divided by the number of read pulses to obtain the average current sum. For more information on measurements, please see the Methods Section and Fig. 3-12. As a result, we were able to obtain the current sums for each column line. Figure 3-3a shows that the processor 'M' layer outputs the highest current sum to the patterned image 'M' rather than the pattern images 'I' and 'T'. This is illustrated in the second row of Figure 3-3a. This indicates that the matching ‘M' pattern has been successfully recognized by the processor ‘M' layer of the processor. The other processor layers, ‘I' and ‘T,' produce similar results to the matching patterns, ‘I' and ‘T,' by producing the highest current sums to the patterns, ‘I' and ‘T,' respectively. We repeated each of these kernel operations 15 times for a total of 15 patterned image data sets. It is indicated by error bars that the maximum and minimum values of current sums, which can be considered to be the values of the multiply-accumulate (MAC) operation, respectively, have been reached.

After that, we attempted parallel kernel operation by taking full advantage of the stack-ability of multiple computing layers that we had previously trained in advance. Figure 3-3b depicts a schematic of a stack with four layers, each of which is composed of one eye layer and three processor layers (see Fig. 3-3a). After the eye layer, three patterned inputs are transferred across three processor layers, one after the other. The results of the current sums of each kernel to pattern images, as demonstrated in Fig. 3-3a, are shown in the first three rows of Fig. 3-3b. The letter recognition task was successfully completed for three different kernels, with the highest output current to the matching input patterned images being used for each of the three kernels. According to our findings, free-space light communication between chips allows us to quickly assemble and disassemble chips, allowing us to switch between diverse kernel operations and stack multiple layers for parallel computing with ease. The ability to build hardware that can be optimized for processing large amounts of complex sensory data in the edge computing era will be a further benefit of this research. When combined with advanced optical circuits, we believe this hardware will be able to facilitate multimodal data processing in robotics and edge computing, among other applications.

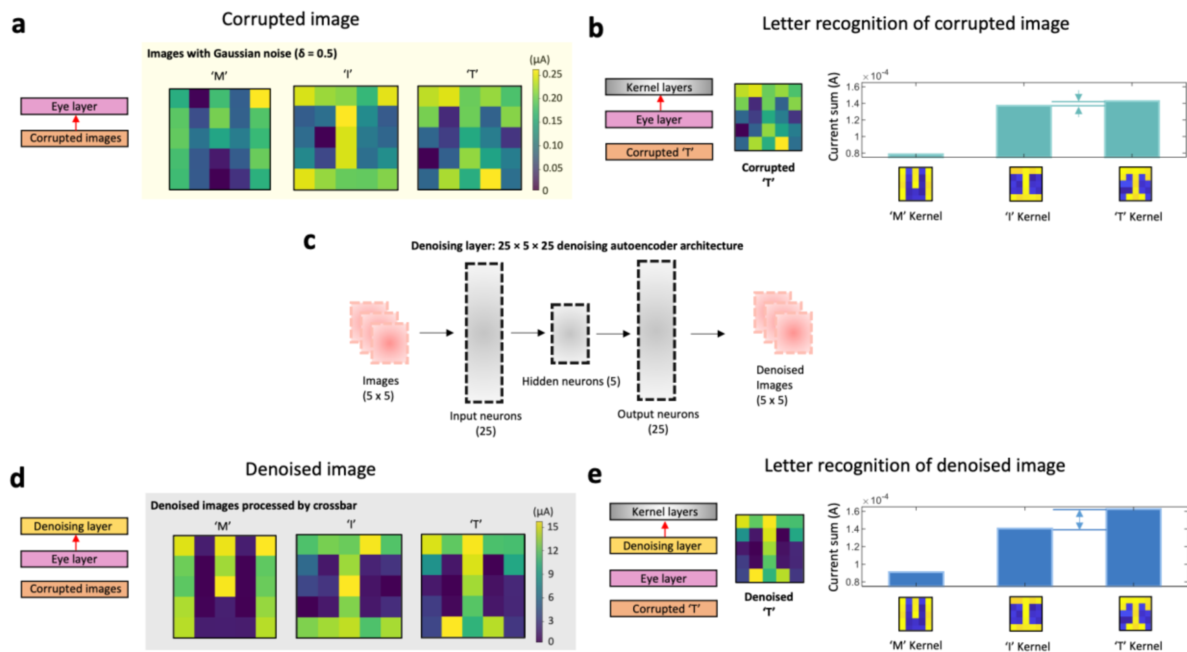

Figure 3-4: The use of stackable neuromorphic chips in a noisy environment: the insertion of a denoising functional layer in the midst of the chaos. a, the addition of Gaussian noise (0.5) to images from the eye layer results in the generation of corrupted letter images. b, block diagram of the letter recognition task that was performed on a corrupted 'T' letter image. On the corrupted letter image of the letter T, we performed three different kernel operations. The current sums are calculated for each kernel individually. The difference between the current sum and the previous sum is indicated by arrows. Because of the noise in the letter "T," the outputs of the current sum from the "I" kernel and the "T" kernel show only a marginal difference, indicating that letter recognition is difficult. c, Denoising functional layer neural network architecture with denoising autoencoder (25 × 5 × 25 neurons) and denoising autoencoder. The denoising layer, which includes memristor crossbar arrays, is added after the letter images have been denoised in step d, as shown in the black diagram. Example of letter recognition task performed on denoised letter image of the letter T, as described in the block diagram shown in Figure 3b. Following the denoising process, the current sum from the 'T' kernel yields a significantly higher value than the current sum from the 'I' kernel. This

implies that, following the completion of the denoising process, the task of letter recognition will be more successful.

System reconfiguration

Aside from chip stackability and replaceability, one of the most significant advantages of reconfigurability is the ability to add new functionalities to a hetero-integrated system. In this article, we successfully showed the processing of damaged pictures by adding a denoising layer to prefabricated stacks that were not previously prepared for noise processing. While the human brain still performs well in moderate noise, the accuracy of artificial neural networks declines dramatically with noise [85]. Although humans' robust representation of visual stimuli is not fully understood, recognition under noise is a long-term goal in computer vision because input data quality distortions from edge devices in insecure and unstable environments such as low light conditions and motion blurring are unavoidable [85]. Image denoising was created using various neural network designs, which may improve network performance on reasoning tasks [86]. First, we applied Gaussian noise to the patterned pictures in Fig. 3-4a, eye layer. For our experiment, we used a high Gaussian noise level ($\delta = 0.5$). It is worth noting in Fig. 3b that distinguishing letters ‘I' and ‘T' using kernel operations is difficult when the input is noisy. This is due to the fact that the pixel information for 'I' and 'T' is identical. In fact, prior to the denoising procedure, a damaged ‘T' picture (Fig. 3-4b) was difficult to distinguish by kernel operations ‘I' and ‘T' due to the noise, resulting in identical current sum values. To address this issue, we developed a denoising autoencoder layer, which is an extension of a traditional autoencoder that employs an unsupervised learning criteria between the eye layer and the kernel layer of the previous stacking architecture [87], [88]. As shown in Fig. 3-4c, the neural network architecture, $25 \times 5 \times 25$ denoising autoencoder, is used to denoise the damaged pictures. 5×5 patterned pictures are flattened and processed by two fully-connected layers with 5 neurons (the first fully-connected layer) and 25 neurons (the second fully-connected layer) (the second fully-connected layer). The output neurons' values are adjusted to produce 5×5 denoised patterned pictures. After introducing a denoising layer, severely damaged

patterned pictures are denoised, as illustrated in Fig. 3-4d. As a result, a denoised ‘T' picture (Fig. 3-4e) seems to output the greatest current sum value to the ‘T' kernel with a higher output current differential. Figures 3-13, 3-14, and the Methods Section provide the software simulation results and training details for the denoising autoencoder. Details of Figs. 3-4b and 3-4e may also be found in Fig. 3-15 for distributions of multiply-accumulate (MAC) operations in software and distributions of output currents in memristor crossbar arrays. We anticipate that our work will have an impact on the field in three ways. For starters, the ability to swap out and reuse chip components would decrease waste and set an example for more ecologically responsible alternatives. Second, our study may be utilized to develop high-yield multi-modal sensor-computing systems, which would have a wide range of applications in the edge computing age. Finally, our research will move the field closer to analog computing AI hardware.

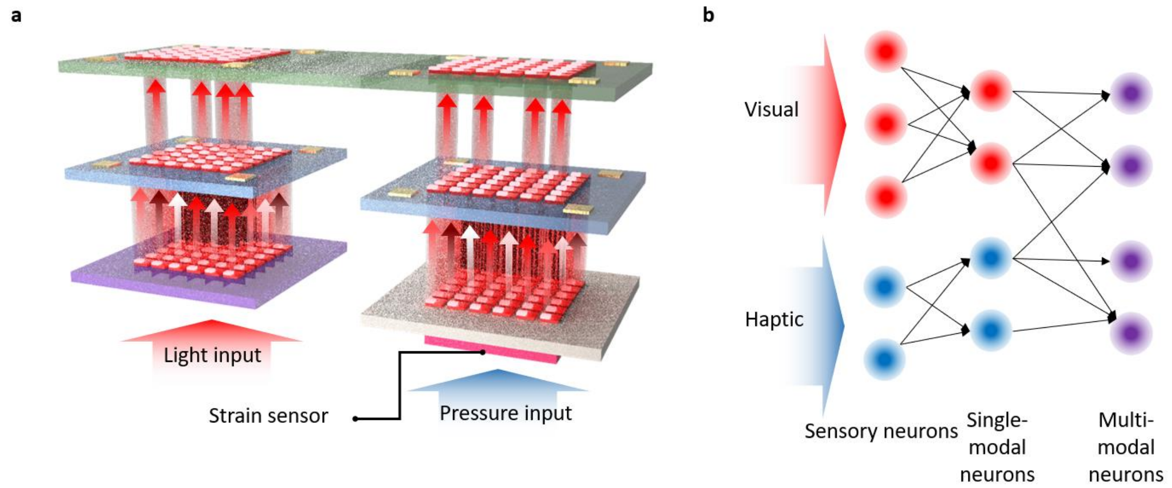

Figure 3-5: Schematic illustration of three-dimensionally stacked neuromorphic chips for multi-modal sensor fusion. a, Photodetectors and strain sensors can provide several sensory inputs. The depth and width of networks can be changed by simply stacking or changing chips. Furthermore, the size and kind of memristor crossbar array can be changed based on the size and function of the sensor array. b, an example of a basic neural network architecture for sensor fusion. The top left (red) neurons represent early visual processing. Bottom left (blue) neurons represent early haptic processing. Right (purple) neurons exhibit multi-modal sensor fusion with sparse connections.

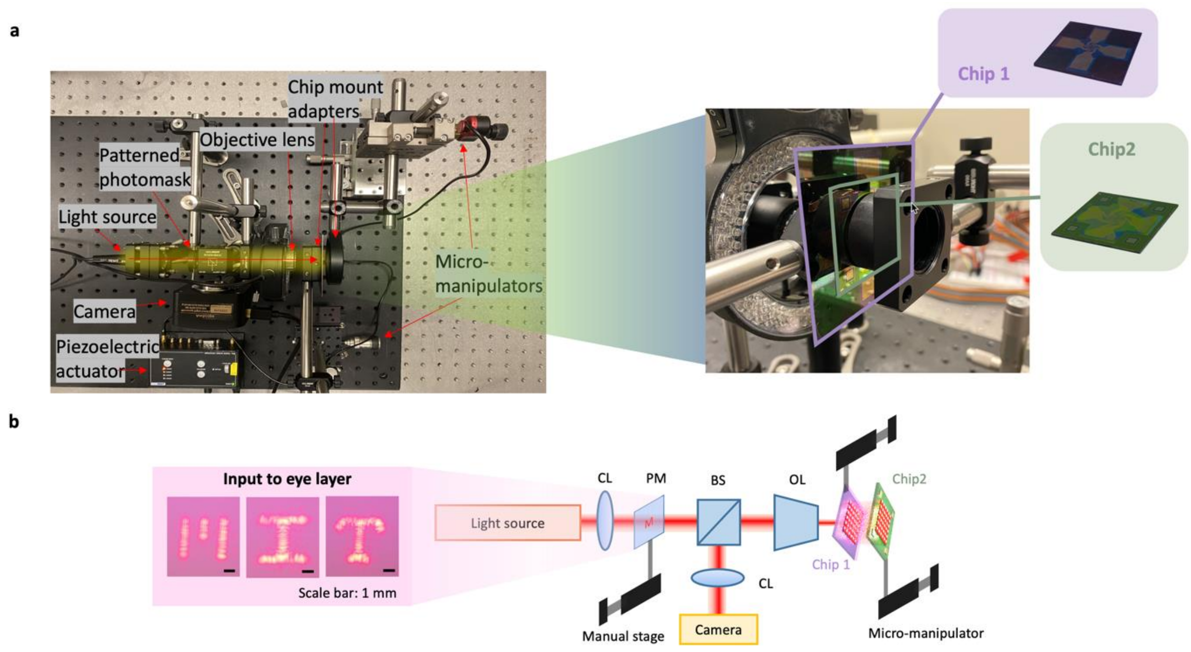

Figure 3-6: Confocal optical measurement setup for chip-to-chip communication. a, An optical image of a confocal imaging setup. The beam path of light input from a 635 nm laser diode is indicated by a yellow arrow (light source). Hetero-integrated chips are mounted to adapters as shown in the photo in the right-hand side. Purple lines represent the chip for the eye layer, whereas green lines represent the chip for the process layer. Micro-manipulators fine-tune two chips. When more than two chips are communicated, they are replaced and measured sequentially. b, A confocal imaging setup diagram. The light is formed into three letters ('M,' 'I,' and 'T') using a patterned photomask. PM stands for patterned mask; CL stands for free-space collimator lens; BS stands for beam splitter; OL stands for objective lens; and CHIP stands for hetero-integrated chip. To adjust the position of light pattern input and chips, a confocal setup was employed. We were able to accomplish minimal optical crosstalk between chips by carefully tweaking micro-manipulators and shifting the positions of chip mount adapters with proximity.

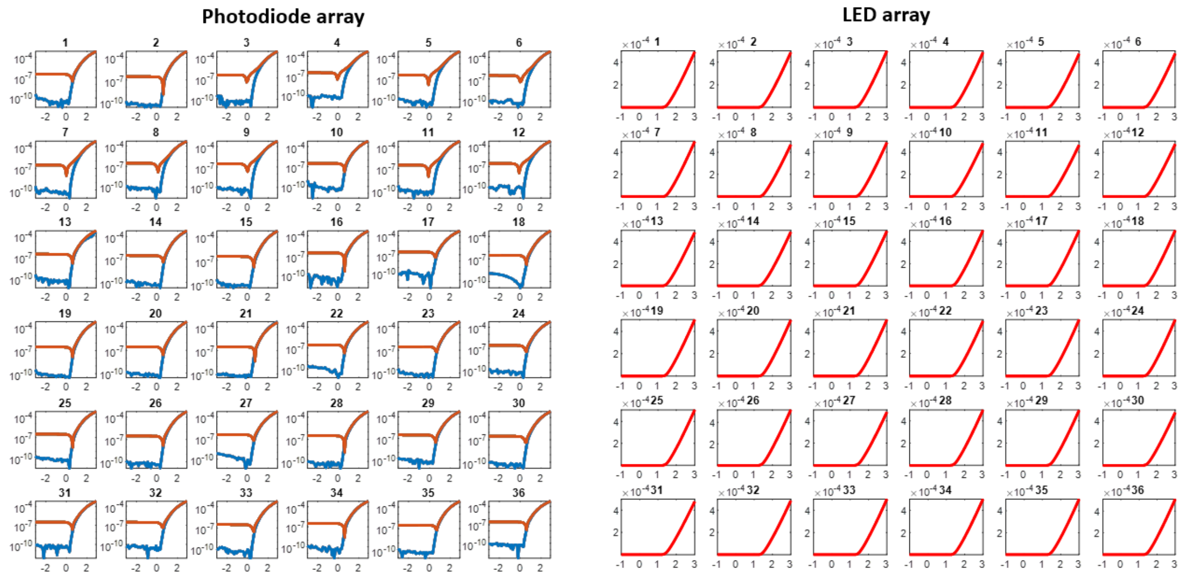

Figure 3-7: I-V characteristics of 6 × 6 photodiode array and 6 × 6 LED array in an eye layer. a, I-V curves of photodiodes. Orange curves show the response of photodiodes to the light (0.1 mW/mm^2) while blue curves show the I-V characteristics with no light. b, I-V curves of LED devices in 6 × 6 array. Optoelectronic devices (photodiodes and LEDs) in other layers (e.g. denoising layer and classification layer) show the similar levels of performance. For the demonstration, only 25 devices were used for 5 × 5 pixels images.

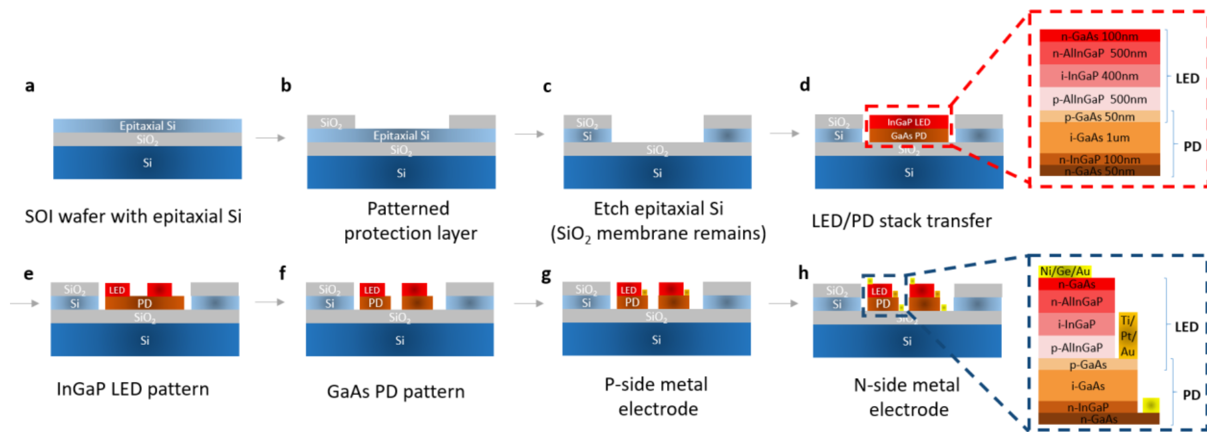

Figure 3-8: Schematic illustrations of photodiodes/LEDs array fabrication for a stackable heterogeneously integrated neuromorphic chip. a, Silicon on insulator (SOI) wafer with epitaxially grown Si was prepared. b, SiO_2 layer was deposited on the SOI wafer by plasma enhanced chemical vapor deposition (PECVD). The SiO2 layer was patterned using a photolithography tool and wet etch (BOE 7:1). c, epitaxial Si was patterned and etched by wet etch (KOH solution). d, After epitaxial lift-off (ELO) process, LED/PD stack was transferred to the SiO_2 membrane treated with APTES and polyimide. The detail of LED/PD stack structure is shown in red box. e, LED was patterned by wet etch (Cr etchant, HCl + H_3PO_4). f, PD was patterned by wet etch. g, Ti/Pt/Au was deposited on the p-doped side. h, Ni/Ge/Au was deposited on the n-doped side. The final structure of LED/PD stack is shown in navy box.

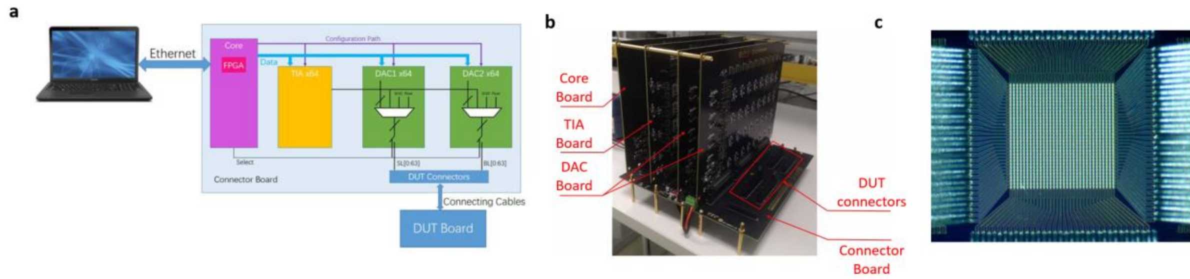

Figure 3-9: Field-programmable gate array (FPGA) system for memristor crossbar array measurement. a, Schematic of FPGA system. Parallel programming and reading were controlled by software on computer. b, Image of the FPGA system. The system consists of a core board, a 64 channels-TIA board, two 64 channels-DAC boards, and a connector board with DUT connectors. DUT connectors were connected to the probe card for programming and reading. We used 5K ohm resistors to initiate the crossbar array. c, Image of the crossbar array mounted to the probe card.

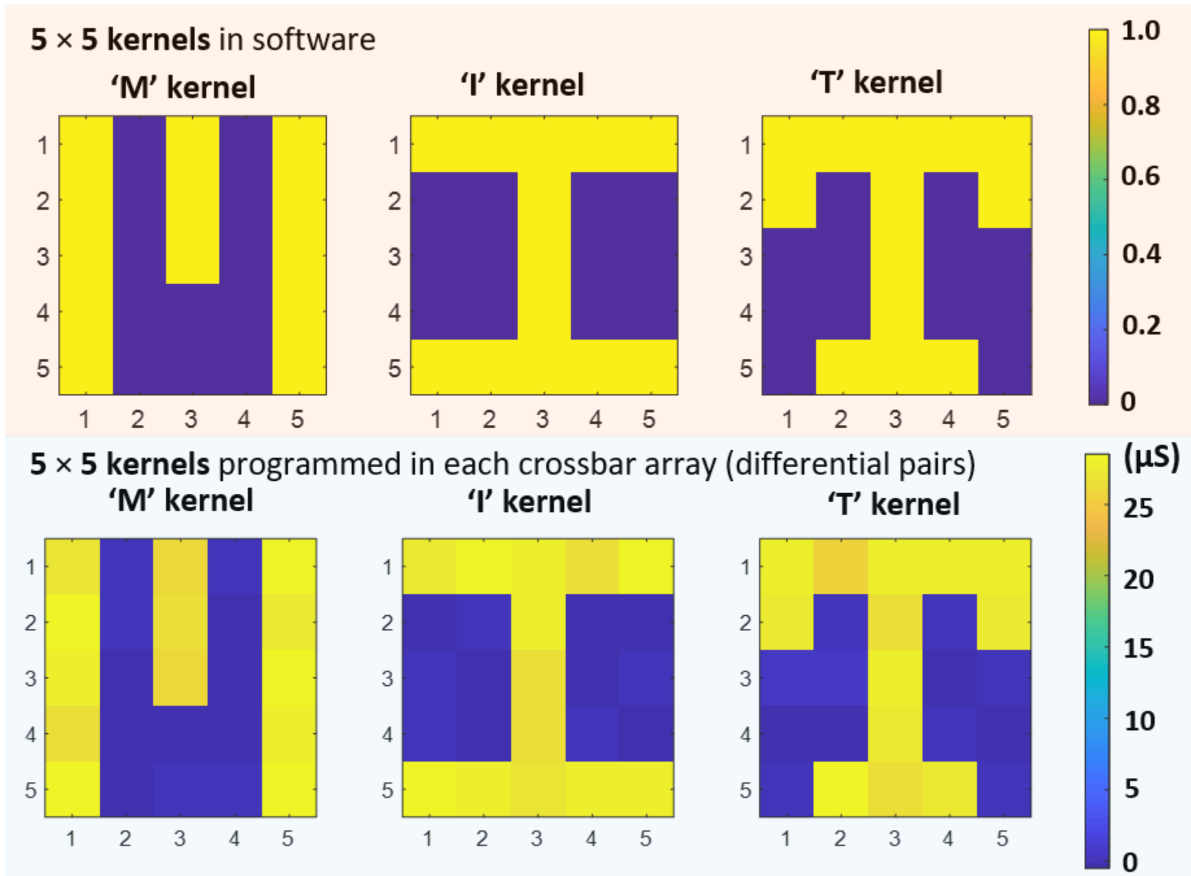

Figure 3-10: Schematic of 5 x 5 kernel operation on 5 x 5-pixel images. To perform a recognition task of three 5 × 5 letter patterns, we implemented three 5 × 5 kernels into Si memristor crossbar arrays. Fig. 3-10 presents three 5 × 5 kernels represented in software (Top) and programmed in memristor crossbar arrays (Bottom), respectively.

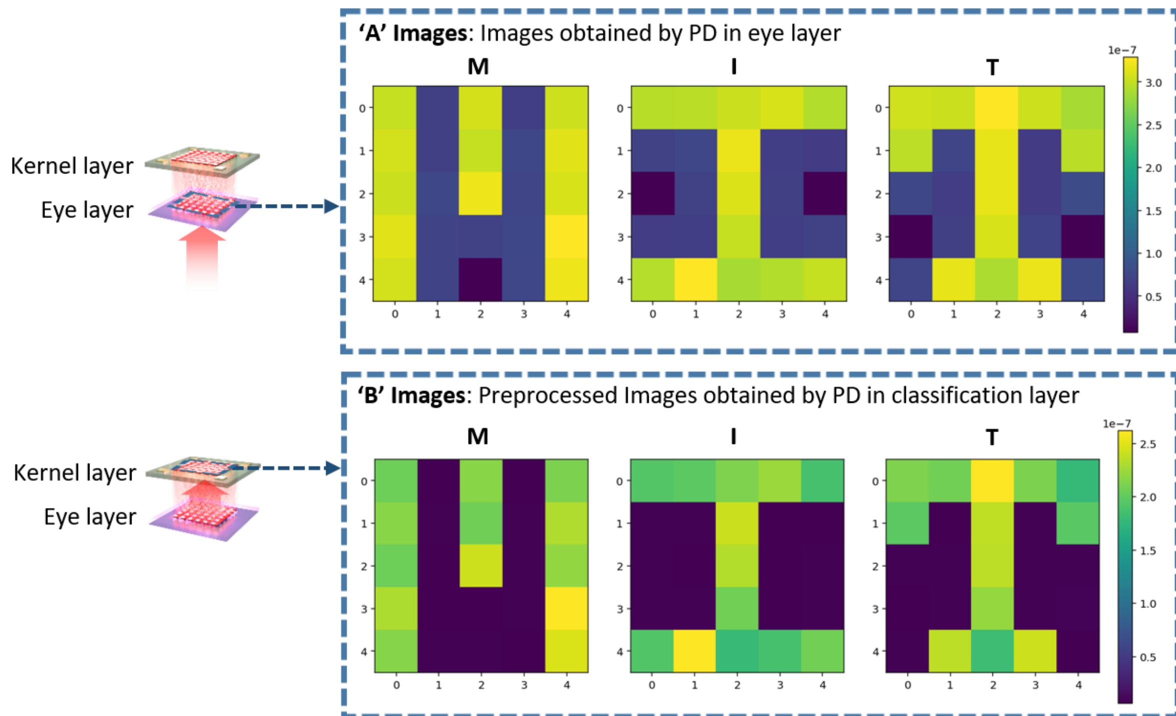

Figure 3-11: Image preprocessing in the layer-to-layer light communication. Patterned Images obtained by photodiode arrays with 0.1 V reverse bias both in eye layer ('A' images) and in classification layer ('B' images). Due to the light diffraction, there are some noises around the letter patterns (Top, 'A' images). These noises are filtered when the light information is processed by a photodiode array in classification layer (Bottom, 'B' images).

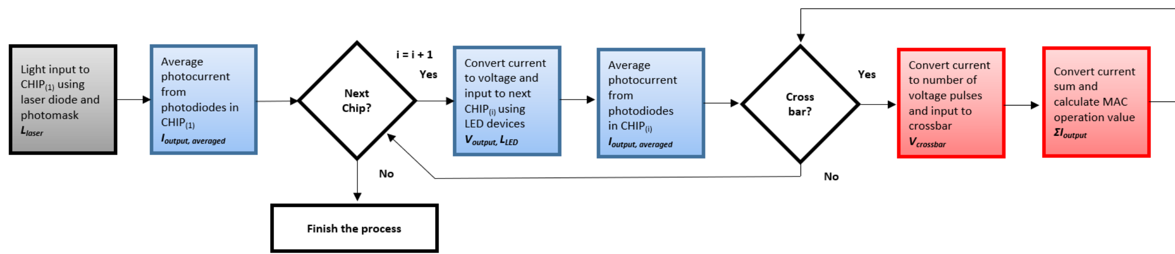

Figure 3-12: Flow chart of the optical and electrical measurements in chips. Black boxes indicate the interactions between a laser diode and photodiodes. Blue boxes indicate the interactions between LED/PD stacks. Red boxes indicate the interactions between a memristor crossbar and the FPGA system. In the process of a crossbar, the conductance programming has been performed using a closed-loop scheme before Multiply-Accumulate (MAC) operations.

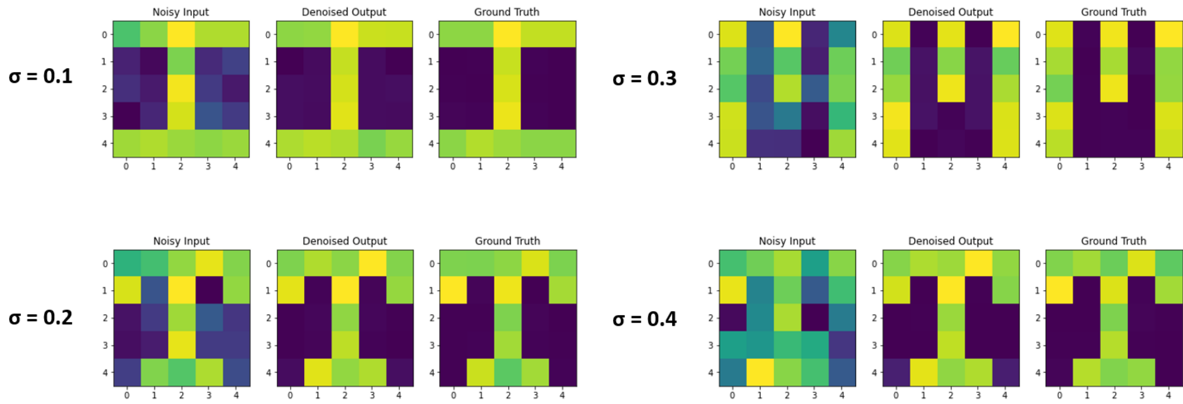

Figure 3-13: Software simulation results of denoising autoencoder to noisy input as a function of Gaussian noise level. Denoising autoencoder has been implemented to improve the pattern recognition. Gaussian noise has been added to ground truth with different noise levels. Noisy inputs are inserted to the $25 \times 5 \times 25$ denoising autoencoder. Denoised output show the simulation result of noisy input. Learning parameters are as follows. Adam optimizer has been adopted with 0.0003 of learning rate and 100 of epochs. Loss has been calculated by the mean squared error (MSE) and weights have been updated by back-propagation. Different level of gaussian noise has been added to the original data.

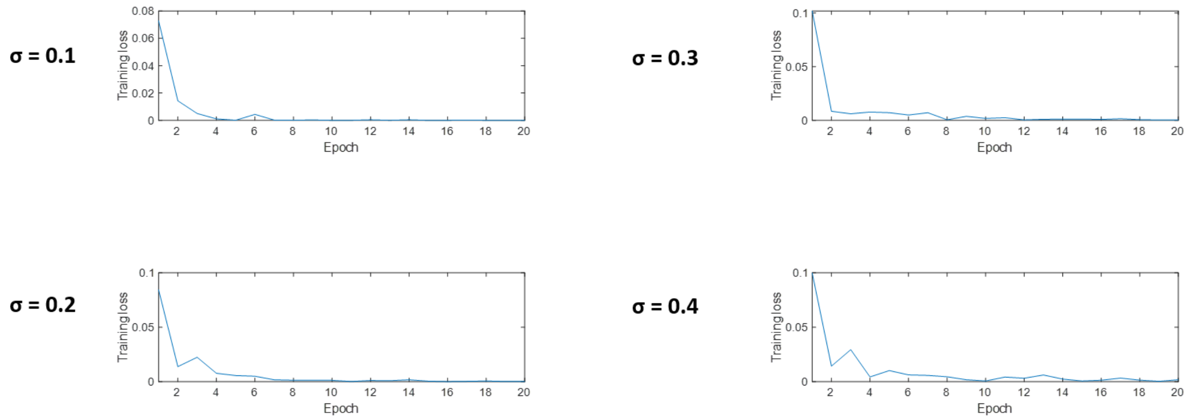

Figure 3-14: Denoising autoencoder training loss. Training loss from denoising autoencoder. Adam optimizer has been adopted with 0.0003 of learning rate. Loss has been calculated by the mean squared error (MSE) and weights have been updated by back-propagation. Different level of gaussian noise has been added to the original data. Total number of dataset is 180 and 20 % of dataset is used for validation.

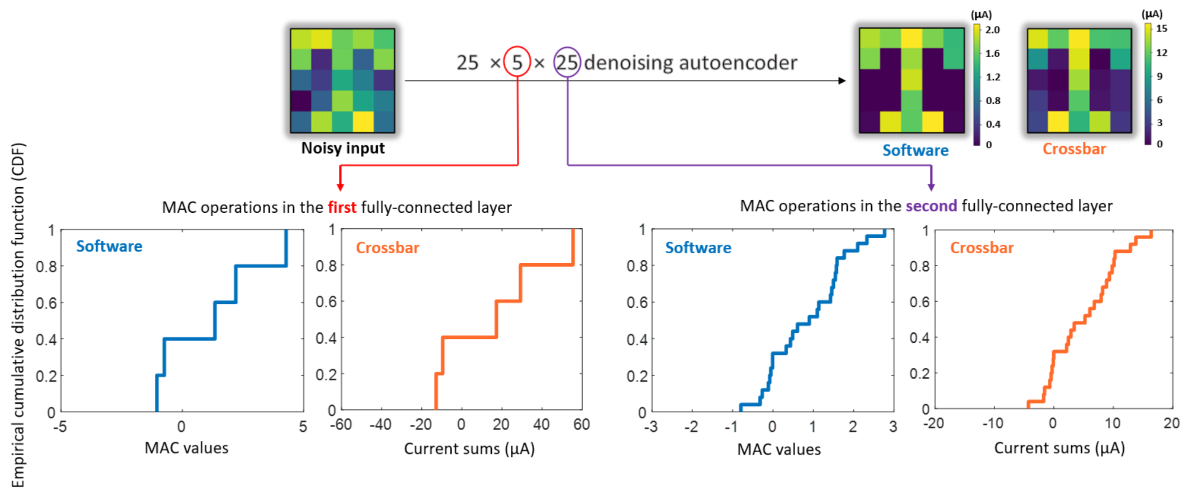

Figure 3-15: Distribution (cumulative distribution function) of Multiply-accumulate (MAC) operations and current sums of denoising autoencoder in software and memristor crossbar arrays. As described in Fig 3-4c, the 25 × 5 × 25 denoising autoencoder has been implemented into Ag-Cu alloy Si-based memristor crossbar arrays. No bias value has been used in the neural network. For both fully-connected layers, differential pairs of memristors are used to represent positive and negative MAC values. A rectified linear unit (ReLu) has been used to MAC values in the first fully-connected layer. Then, MAC values are normalized and converted to voltage pulses before the second fully-connected layer. Histograms show the MAC values performed by software (floating number) and Ag-Cu alloy Si-based memristor crossbar arrays (current sums averaged by the number of maximum number pulses = 100, unit - μA). Left shows the MAC values in the first fully-connected layer and right shows the MAC values in the second fully-connected layer. The output current from the crossbar is calculated by memristor differential pairs. The negative MAC values are zeroed for the next kernel layer.

This section is dedicated for Methods.

Stackable Optoelectronic Neuromorphic Chip Fabrication.

We employed a silicon on insulator (SOI) wafer with 200 nm p^+ Si (0.01 Ω cm, 8E18) and 500 nm p^{++} Si (0.002 Ω cm, 7E19) layers epitaxially formed. First, a SiO_2 layer was formed as a protective layer (PECVD-Samco-PD220, high frequency plasma at 300° C) and patterned with photolithography and a 7:1 buffered oxide etchant (BOE). Second, we etched Si at 40 degrees Celsius with a 50 % KOH solution for GaAs LED/PD stack integration. Meanwhile, we used 49 percent HF to execute an epitaxial lift-off (ELO) technique on an LED/PD stack. Finally, the LED/PD stack was taken up by PDMS and released onto the SiO_2 layer, where it was treated with oxygen plasma (50 sccm, 100W, 5 mins) and coated with (3-Aminopropyl)triethoxysilane (APTES) and polyimide. The LED/PD stack was then patterned using a photolithography tool (MLA-150) and wet etching (Cr etchant for GaAs etch, HCl/H_3PO_4 for InGaP etch). While Ti/Pt/Au (10/20/100 nm) was used for the p-type side contact, Ge/Ni/Au (20/20/200 nm) was used for the n-type side contact. Wet etch was used to isolate each device after metal contact deposition. The LED/PD arrays were encased in SiO_2 (100 nm) and patterned with reactive ion etching (RIE). Then, using photolithography and electron beam evaporation, a top metal layer of Ti/Au (10/400 nm) was deposited. The manufacturing of the Ag-Cu alloy-based Si memristor crossbar array was the next stage, as detailed in ref [39]. After removing the SiO_2 protective layer placed in the first phase, the crossbar array was created. Finally, backside holes aligned with the LED/PD arrays were made using photolithography and the deep reactive ion etch (DRIE) procedure for chip-to-chip communications. Using scanning electron microscopy and optical microscopy, we demonstrated that Si was totally etched and that the SiO_2 layer persisted.

Optical measurement.

Fig. 3-6 depicts an image as well as an optical measuring schematic. The original chip's light source was a laser diode (635 nm wavelength), which was powered by a pulse generator unit. The laser diode's light was collimated using a mounted achromatic doublet and lenses. We created the confocal microscopy setup using a beam splitter and an objective lens to guarantee that the light was focused on the chip. Two micro-manipulators and a piezoelectric actuator were used to precisely position two chips for chip-to-chip communication. Every light-emitting diode and photodiode on the chips was fanned out of the device area, allowing for external I/O configuration. The chips' top metal layer was in contact with an anisotropic conductive film (3M) before being connected to a pulse generator unit (33600A) and a digital storage oscilloscope (DSOX3024T) via flat flex connectors (FFC).

Electrical Measurement Setup.

We outfitted the field-programmable gate array (FPGA) system with 64 channel transimpedance amplifiers (TIA) and two 64 channel digital-analog converters for the memristor crossbar array measurement (DACs). The probe card was designed to accommodate a 32 × 32 crossbar array. The resolution of the voltage output DAC is 16 bits. The resolution of the voltage acquisition DAC is 12 bits. There are 128 programmable I/O channels and 64 current collection channels in the system. The maximum voltage output amplitude is 10V, with a resolution of 1.2 mV. The array measurement was carried out using a bespoke board-level peripheral system with parallel accessing and programming capabilities. (More information on resistors and current compliance will be provided later.) Fig. 3-9 shows the schematic and image of the FPGA system, as well as a 32 × 32 array image. The conductance values in the memristor were programmed utilizing a closed loop technique for

kernel operations and denoising autoencoder. The read voltage pulses of 0.5 V with a duration of 10 milliseconds were used. The amplitude of the input determined the number of read pulses. In the paper, the output current is accumulated over the number of read pulses and then averaged by the number of read pulses to calculate the current sum.

Kernel operation in kernel layers.

In distinct memristor crossbar arrays, we programmed three letter kernels, 'M' kernel layer, 'I' kernel layer, and 'T' kernel layer. In memristor crossbar arrays, 25 × 2 Si memristors were employed for each kernel operation, as explained in the manuscript. In memristor crossbar arrays, the kernel procedure consists of three steps: (1) weight programming, (2) differential pairs testing, and (3) voltage input for kernel operations are all possible. As previously stated, we examined 5 × 5 differential pairs of memristors, which represent 5 × 5 pixels, prior to performing kernel operation. The measured values for 5 × 5 kernels are shown in Fig. 3-10. There was no substantial retention decay detected during the writing and reading processes.

Neural network simulation and implementation.

Python and Pytorch were used to create the denoising autoencoder. The denoising autoencoder's architecture is 25 × 5 × 25 without bias. (The number of neurons is represented by each number.) First, we measured the current output values of a photodiode array in response to patterned light input. The current output from photodiodes was amplified by 10^7 and used in the neural network simulation under the assumption that the gain of the transimpedance amplifier was set to 10^7. Following the first fully-connected layer of 25 × 5, the activation function was chosen as a Rectified linear unit (ReLU). During the training phase, we used MSE as the loss function, Adam as the optimizer, a learning rate of 0.003, and an epoch of 100. The dataset is 180 by 180 ('M': 60, 'I': 60, 'T': 60). We used 144 letter data for the training set and

36 letter data for the test set. Following the completion of the training, we obtained weight values in order to program memristor crossbar arrays. The specifically built probe card communicated between the FPGA system and the memristor crossbar arrays. The Python-MATLAB interface transforms write and read requests into FPGA commands for programming and accessing memristor conductance values. The denoising autoencoder output (denoised pictures) is normalized and transformed to pixel values. The pixel data are translated to voltage and delivered to LED devices in the next layer to conduct kernel actions. Fig. 3-12 depicts the workflow of light communications and memristor crossbar array activities.

3.3 Conclusion

In summary, we demonstrated the stability of light communication by performing kernel operations, which are core activities in AI, on multi-stacked neuromorphic circuits with three separate pre-programmed processor layers. Furthermore, by incorporating a denoising layer between the eye and kernel layers, our chip stack demonstrates good tolerance to high noise levels. Pattern identification by kernel operations improved dramatically after denoising damaged images. These Lego-like, heterogeneously connected neuromorphic processors give designers a lot of leeway when it comes to designing near-sensor computing systems that can be seamlessly optimized for edge AI applications. Our framework, which is based on the co-design of sensory/neuromorphic devices and chip architecture, enables a wide range of applications for reconfigurable edge neuromorphic computing. Our chips will also bring tremendous adaptability to neuromorphic edge computing by leveraging critical machine learning techniques like transfer learning, which simply require fine-tuning of the last layer.

Chapter 4

Conclusion

This chapter is devoted to a summary of my thesis, which includes a conclusion and recommendations for further research.

4.1 Conclusion

The traditional RRAM devices for neuromorphic computing, as described in Chapter 1, were suffering from poor device performance metrics such as endurance, retention, multi-level representation, temporal and spatial fluctuations. However, while each of the device performance metrics might be improved separately, there was no RRAM device that could satisfy all of the specifications at the same time. The capacity of multi-level representation is particularly important for the next generation of AI hardware, which is required to improve device density while also avoiding the need for analog-to-digital conversions. As a result, neuromorphic devices may be used to achieve completely analog computing. Because of the stochastic nature of Ag atoms, which may result in more slow ion motions in a switching medium, Ag-based RRAM devices have been under investigation for a long time. The stochastic behavior of Ag, on the other hand, is very difficult to regulate because of the weak

thermal interaction between Ag and the Si medium, which results in computing errors in neuromorphic computing.

In Chapter 2, I introduced an alloying technique for the typical Ag-based RRAM devices, which was followed by a discussion of the results. It was discovered that using this alloying method, device performance improved in all metrics, including long endurance (> 10^7), long retention (> 1 hr at high temperature), and improved variation (when compared to the performance of Ni-Ag and Cu-Ag alloys) at multi-level conductance values. Second, I demonstrated stackable hardware-wise reconfigurable chips that were equipped with a variety of sensors and processors. One of the chips' features is an artificial intelligence module, which is enabled via crossbar arrays of alloyed RRAM devices. It offered hardware-based stack-ability as well as the ability to reconfigure neural network architectures. I showed basic image processing, letter recognition, and denoising utilizing kernel operations in deep neural networks, as well as simple image processing, letter recognition, and denoising. Analog neuromorphic computing and edge computing will be made possible by the use of the optimized RRAM devices and reconfigurable architecture in a cost-effective and energy-efficient manner.

4.2 Future work

The primary objectives of neuromorphic computing as the next generation of artificial intelligence hardware are (1) acceleration of matrix multiplication, (2) reduced system power consumption, and (3) completely analog computing, all of which are achievable. Despite the fact that my work in this thesis, which includes materials science innovation and hardware architectural flexibility, may lead to advances in three objectives, there are still difficulties in (1) device, (2) circuit, and (3) architecture design and implementation of neuromorphic computing.

The following are the difficulties that neuromorphic devices must overcome. In the first place, it is difficult to integrate a neuromorphic device with traditional material systems. Second, only a small number of conductance ranges have been obtained thus far. Third, there are significant spatiotemporal device variations found. In order to address these problems, two-dimensional materials-based memristors, piezoelectric material-based memristors, and three-terminal memristors are presented as potential candidates. When compared to two-terminal memristive devices, gate-controlled three-terminal memristive devices may be able to regulate the ion movement with more precision than their counterparts.

There are additional difficulties in implementing neuromorphic computing at the circuit level, which must be overcome in order to be effective. First, there is a sneak path problem with crossbar arrays. Second, backpropagation is difficult to include into the circuit because calculating partial derivatives is unfriendly to the existing neuromorphic circuits, making it difficult to incorporate. Third, transfers between read mode and write mode, as well as digital-to-analog conversions, have the potential to waste a significant amount of energy. Suppressing a sneak route using a three-dimensional crossbar array with high resistance may be a good solution for dealing with the first problem mentioned above. Backpropagation may be accomplished by embedding analog circuits within the chips, which approximate the partial derivative values and update devices in the process. Selector devices have the potential to reduce the energy consumption associated with digital-to-analog conversions.

Because of a lack of weight update methods, system-level neuromorphic computing has not been completely explored to its potential. Recently, memristor crossbar arrays for convolutional neural networks have been developed and implemented [89]. Despite the fact that this study demonstrated successful convolutional neural network implementations with distinct programmed weights using RRAM devices, it was unable to attain high precision via the use of a weight transfer technique. With the introduction of hybrid training, which consider

device variability as well as incomplete training data, it achieves accuracy that is similar to that of software. Because the update algorithms for neuromorphic systems are still in the early stages of research, it is necessary to create hybrid techniques that are tailored to certain networks and applications. As neuromorphic devices, circuits, and architecture continue to improve, it is expected that fully hardware-implemented neuromorphic chips will be available at one point in the future. However, the requirements for device-, circuit-, and architectural design, on the other hand, are heavily influenced by the applications. Therefore, rather than focusing only on increasing performance at each level, I believe that future work will need to include co-designing hardware and software in order to optimize the neuromorphic system depending on the applications.

Chapter 5

Bibliography

[1] S. Choi *et al.*, “SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations,” *Nature Materials*, vol. 17, no. 4, pp. 335–340, 2018.

[2] C. Li *et al.*, “Analogue signal and image processing with large memristor crossbars,” *Nature Electronics*, vol. 1, no. January, pp. 52–59, 2018.

[3] C. Li *et al.*, “Efficient and self-adaptive in-situ learning in multilayer memristor neural networks,” *Nature Communications*, vol. 9, no. 1, pp. 7–14, 2018.

[4] M.-J. Lee *et al.*, “A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O5−x/TaO2−x bilayer structures,” *Nature Materials*, vol. 10, no. 8, pp. 625–630, 2011.

[5] H. Jiang *et al.*, “Sub-10 nm Ta Channel Responsible for Superior Performance of a HfO2 Memristor,” *Scientific Reports*, vol. 6, pp. 1–8, 2016.

[6] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, “Nanoscale memristor device as synapse in neuromorphic systems,” *Nano Letters*, vol. 10, no. 4, pp. 1297–1301, 2010.

[7] R. W. Olesinski, A. B. Gokhale, and G. J. Abbaschian, “The Ag-Si (Silver-Silicon)

system," *Bulletin of Alloy Phase Diagrams*, vol. 10, no. 6, pp. 635–640, Dec. 1989.

[8] S. SABOONI, F. KARIMZADEH, and M. H. ABBASI, "Thermodynamic aspects of nanostructured Ti5Si3 formation during mechanical alloying and its characterization," *Bulletin of Materials Science*, vol. 35, no. 3, pp. 439–447, Jun. 2012.

[9] A. B. Gokhale and G. J. Abbaschian, "The Cr−Si (Chromium-Silicon) system," *Journal of Phase Equilibria*, vol. 8, no. 5, p. 474, Oct. 1987.

[10] P. Nash and A. Nash, "The Ni−Si (Nickel-Silicon) system," *Bulletin of Alloy Phase Diagrams*, vol. 8, no. 1, pp. 6–14, Feb. 1987.

[11] H. Okamoto, "Cu-Si (Copper-Silicon)," *Journal of Phase Equilibria and Diffusion*, vol. 33, no. 5, pp. 415–416, Oct. 2012.

[12] J. L. Murray and K. J. Bhansali, "The Ag−Ti (Silver-Titanium) system," *Bulletin of Alloy Phase Diagrams*, vol. 4, no. 2, pp. 178–183, Sep. 1983.

[13] M. Venkatraman and J. P. Neumann, "The Ag-Cr (Silver-Chromium) system," *Bulletin of Alloy Phase Diagrams*, vol. 11, no. 3, pp. 263–265, Jun. 1990.

[14] M. Singleton and P. Nash, "The Ag–Ni (Silver-Nickel) system," *Journal of Phase Equilibria*, vol. 8, no. 2, pp. 119–121, Apr. 1987.

[15] P. R. Subramanian and J. H. Perepezko, "The ag-cu (silver-copper) system," *Journal of Phase Equilibria*, vol. 14, no. 1, pp. 62–75, Feb. 1993.

[16] S. Kuge and H. Nakashima, "Solubility and Diffusion Coefficient of Electrically Active Titanium in Silicon," *Japanese Journal of Applied Physics*, vol. 30, no. Part 1, No. 11A, pp. 2659–2663, Nov. 1991.

[17] W. Würker, K. Roy, and J. Hesse, "Diffusion and solid solubility of chromium in

silicon," *Materials Research Bulletin*, vol. 9, no. 7, pp. 971–977, 1974.

[18] E. R. Weber, "Transition metals in silicon," *Applied Physics A*, vol. 30, no. 1, pp. 1–22, Jan. 1983.

[19] A. A. Istratov and E. R. Weber, "Physics of Copper in Silicon," *Journal of The Electrochemical Society*, vol. 149, no. 1, pp. G21–G30, 2002.

[20] F. Rollert, N. A. Stolwijk, and H. Mehrer, "Solubility, diffusion and thermodynamic properties of silver in silicon," *Journal of Physics D: Applied Physics*, vol. 20, no. 9, pp. 1148–1155, Sep. 1987.

[21] S. H. Jo and W. Lu, "CMOS Compatible Nanoscale Nonvolatile Resistance Switching Memory," *Nano Letters*, vol. 8, no. 2, pp. 392–397, Feb. 2008.

[22] K. M. Kim, D. S. Jeong, and C. S. Hwang, "Nanofilamentary resistive switching in binary oxide system$\mathsemicolon$ a review on the present status and outlook," *Nanotechnology*, vol. 22, no. 25, p. 254002, May 2011.

[23] T. Tsuruoka, K. Terabe, T. Hasegawa, I. Valov, R. Waser, and M. Aono, "Effects of Moisture on the Switching Characteristics of Oxide-Based, Gapless-Type Atomic Switches," *Advanced Functional Materials*, vol. 22, no. 1, pp. 70–77, 2012.

[24] I. Valov *et al.*, "Nanobatteries in redox-based resistive switches require extension of memristor theory," *Nature Communications*, vol. 4, no. 1, p. 1771, 2013.

[25] I. Valov and T. Tsuruoka, "Effects of moisture and redox reactions in {VCM} and {ECM} resistive switching memories," *Journal of Physics D: Applied Physics*, vol. 51, no. 41, p. 413001, Aug. 2018.

[26] F. P. Fehlner, *Low-Temperature Oxidation: The Role of Vitreous Oxides*. New York: John Willey & Sons, Inc., 1986.

[27] L. Zhang, X. Xiong, Y. Yan, K. Gao, L. Qiao, and Y. Su, "Atomic modeling for the initial stage of chromium passivation," *International Journal of Minerals, Metallurgy, and Materials*, vol. 26, no. 6, pp. 732–739, Jun. 2019.

[28] S. P. Murarka, "Silicon Dioxide, Nitride, and Oxynitride," in *Encyclopedia of Materials: Science and Technology*, K. H. J. Buschow, R. W. Cahn, M. C. Flemings, B. Ilschner, E. J. Kramer, S. Mahajan, and P. Veyssière, Eds. Oxford: Elsevier, 2003, pp. 1–14.

[29] X. Yang, B. J. Choi, A. B. K. Chen, and I.-W. Chen, "Cause and Prevention of Moisture-Induced Degradation of Resistance Random Access Memory Nanodevices," *ACS Nano*, vol. 7, no. 3, pp. 2302–2311, Mar. 2013.

[30] S.-K. Kang *et al.*, "Dissolution Behaviors and Applications of Silicon Oxides and Nitrides in Transient Electronics," *Advanced Functional Materials*, vol. 24, no. 28, pp. 4427–4434, 2014.

[31] H. Lv *et al.*, "Evolution of conductive filament and its impact on reliability issues in oxide-electrolyte based resistive random access memory," *Scientific Reports*, vol. 5, pp. 1–6, 2015.

[32] J. B. Roldan *et al.*, "Recommended Methods to Study Resistive Switching Devices," *Advanced Electronic Materials*, vol. 5, no. 1, p. 1800143, 2018.

[33] J. G. Simmons and R. R. Verderber, "New conduction and reversible memory phenomena in thin insulating films," *Proceedings of the Royal Society of London Series A Mathematical and Physical Sciences*, vol. 301, no. 1464, pp. 77–102, 1967.

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

[35] Y. Shi *et al.*, “Electronic synapses made of layered two- dimensional materials,” *Nature Electronics*, vol. 1, no. August, pp. 458–465, 2018.

[36] Y. Yang, P. Gao, S. Gaba, T. Chang, X. Pan, and W. Lu, “Observation of conducting filament growth in nanoscale resistive memories,” *Nature Communications*, vol. 3, pp. 732–738, 2012.

[37] Y. H. Ting, J. Y. Chen, C. W. Huang, T. K. Huang, C. Y. Hsieh, and W. W. Wu, “Observation of Resistive Switching Behavior in Crossbar Core–Shell Ni/NiO Nanowires Memristor,” *Small*, vol. 14, no. 6, pp. 1–8, 2018.

[38] Y. Hirose and H. Hirose, “Polarity-dependent memory switching and behavior of Ag dendrite in Ag-photodoped amorphous As2S3 films,” *Journal of Applied Physics*, vol. 47, no. 6, pp. 2767–2772, 1976.

[39] H. Yeon *et al.*, “Alloying conducting channels for reliable neuromorphic computing,” *Nature Nanotechnology*, vol. 15, no. 7, pp. 574–579, 2020.

[40] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, “The missing memristor found,” *Nature*, vol. 453, no. 7191, pp. 80–83, 2008.

[41] Q. Xia and J. J. Yang, “Memristive crossbar arrays for brain-inspired computing,” *Nature Materials*, vol. 18, no. 4, pp. 309–323, 2019.

[42] G. W. Burr *et al.*, “Neuromorphic computing using non-volatile memory,” *Advances in Physics: X*, vol. 2, no. 1, pp. 89–124, 2017.

[43] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, “Training and operation of an integrated neuromorphic network based on metal-oxide memristors,” *Nature*, vol. 521, no. 7550, pp. 61–64, 2015.

[44] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, “Electrochemical metallization

memories - Fundamentals, applications, prospects,” *Nanotechnology*, vol. 22, no. 25. {IOP} Publishing, p. 254003, May-2011.

[45] M. Lübben and I. Valov, “Active Electrode Redox Reactions and Device Behavior in ECM Type Resistive Switching Memories,” *Advanced Electronic Materials*, vol. 5, no. 9, p. 1800933, 2019.

[46] P. Yao *et al.*, “Face classification using electronic synapses,” *Nature Communications*, vol. 8, no. May, pp. 1–8, 2017.

[47] Z. Wang *et al.*, “Fully memristive neural networks for pattern classification with unsupervised learning,” *Nature Electronics*, vol. 1, no. 2, pp. 137–145, 2018.

[48] F. Cai *et al.*, “A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations,” *Nature Electronics*, vol. 2, no. 7, pp. 290–299, 2019.

[49] Y. Shi *et al.*, “Neuroinspired unsupervised learning and pruning with subquantum CBRAM arrays,” *Nature Communications*, vol. 9, no. 1, p. 5312, 2018.

[50] S. Ambrogio *et al.*, “Neuromorphic learning and recognition with one-transistor-one-resistor synapses and bistable metal oxide RRAM,” *IEEE Transactions on Electron Devices*, vol. 63, no. 4, pp. 1508–1515, 2016.

[51] J. Woo and S. Yu, “Resistive Memory-Based Analog Synapse: The Pursuit for Linear and Symmetric Weight Update,” *IEEE Nanotechnology Magazine*, vol. 12, no. 3, pp. 36–44, 2018.

[52] S. Dietrich *et al.*, “A non-volatile 2Mbit CBRAM memory core featuring advanced read and program control,” in *IEEE Journal of Solid-State Circuits*, 2007, vol. 42, no. 4, pp. 839–845.

[53] W. Otsuka *et al.*, “A 4Mb conductive-bridge resistive memory with 2.3GB/s read-throughput and 216MB/s program-throughput,” *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, no. May 2010, pp. 210–211, 2011.

[54] J. van den Hurk *et al.*, “Physical origins and suppression of Ag dissolution in GeSx-based ECM cells,” *Physical chemistry chemical physics : PCCP*, vol. 16, 2014.

[55] J. Hajto, A. E. Owen, A. J. Snell, P. G. Le Comber, and M. J. Rose, “Analogue memory and ballistic electron effects in metal-amorphous silicon structures,” *Philosophical Magazine B*, vol. 63, no. 1, pp. 349–369, 1991.

[56] A. J. Snell, P. G. Lecomber, J. Hajto, M. J. Rose, A. E. Owen, and I. S. Osborne, “Analogue memory effects in metal/a-Si:H/Metal memory devices,” *Journal of Non-Crystalline Solids*, vol. 137–138, pp. 1257–1262, 1991.

[57] S. H. Jo, K.-H. Kim, and W. Lu, “High-Density Crossbar Arrays Based on a Si Memristive System,” *Nano Letters*, vol. 9, no. 2, pp. 870–874, Feb. 2009.

[58] H. Rickert, *Electrochemistry of Solids*. Berlin Heidelberg New York: Springer-Verlag, 1982.

[59] D. J. Fisher, *Diffusion in Silicon: 10 Years of Research*. Scitec Publications, 1998.

[60] M. Hu *et al.*, “Memristor-based analog computation and neural network classification with a dot product engine,” *Advanced Materials*, vol. 30, no. 9, p. 1705914, 2018.

[61] K. E. Porter, D. A.; Eastering, *Phase Transformations in Metals and Alloys*, 2nd ed. Chapman & Hall, 1992.

[62] F. Pan and V. Subramanian, “A Kinetic Monte Carlo study on the dynamic switching properties of electrochemical metallization RRAMs during the SET process,” *International Conference on Simulation of Semiconductor Processes and Devices,*

SISPAD, pp. 19–22, 2010.

[63] R. Becker and W. Doring, “Kinetic Treatment of Germ Formation in Supersaturated Vapour,” *Ann Phys*, vol. 24, pp. 719–752, 1935.

[64] J. Polte, “Fundamental growth principles of colloidal metal nanoparticles - a new perspective,” *CrystEngComm*, vol. 17, no. 36, pp. 6809–6830, 2015.

[65] H. Liu, M. Zhao, Y. Lei, C. Pan, and W. Xiao, “Formaldehyde on TiO2 anatase (1 0 1): A DFT study,” *Computational Materials Science*, vol. 51, no. 1, pp. 389–395, 2012.

[66] Z. Wang *et al.*, “Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing,” *Nature Materials*, vol. 16, no. 1, pp. 101–108, 2017.

[67] R. Waser, *Nanoelectronics and Information Technology: Advanced Electronic Materials and Novel Device*. Weinheim, Germany: Wiley-VCH, 2012.

[68] M. William., “Handbook of Chemistry and Physics, 93rd edition,” p. 2664, 2012.

[69] B. Parsons, “Standard Potentials in Aqueous Solutions,” vol. 0, no. 4, p. 2007, 1985.

[70] R. Waser, R. Dittmann, C. Staikov, and K. Szot, “Redox-based resistive switching memories nanoionic mechanisms, prospects, and challenges,” *Advanced Materials*, vol. 21, no. 25–26, pp. 2632–2663, 2009.

[71] P. Huang, B. Gao, B. Chen, F. Zhang, L. Liu, and G. Du, “Stochastic simulation of forming, SET and RESET process for transition metal oxide-based resistive switching memory,” *Sispad 2012*, no. 2011, pp. 312–315, 2012.

[72] B. Gao *et al.*, “Modeling disorder effect of the oxygen vacancy distribution in filamentary analog RRAM for neuromorphic computing,” *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 4.4.1-4.4.4, 2018.

[73] M. Wang *et al.*, “Gesture recognition using a bioinspired learning architecture that integrates visual data with somatosensory data from stretchable sensors,” *Nature Electronics*, vol. 3, no. 9, pp. 563–570, 2020.

[74] F. Zhou and Y. Chai, “Near-sensor and in-sensor computing,” *Nature Electronics*, vol. 3, no. 11, pp. 664–671, 2020.

[75] M. M. Shulaker *et al.*, “Three-dimensional integration of nanotechnologies for computing and data storage on a single chip,” *Nature*, vol. 547, no. 7661, pp. 74–78, 2017.

[76] W. Zhang *et al.*, “Neuro-inspired computing chips,” *Nature Electronics*, vol. 3, no. 7, pp. 371–382, 2020.

[77] P. Lin *et al.*, “Three-dimensional memristor circuits as complex neural networks,” *Nature Electronics*, vol. 3, no. 4, pp. 225–232, 2020.

[78] W.-H. Chen *et al.*, “CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors,” *Nature Electronics*, 2019.

[79] G. Hills *et al.*, “Modern microprocessor built from complementary carbon nanotube transistors,” *Nature*, vol. 572, no. 7771, pp. 595–602, 2019.

[80] M. D. Bishop *et al.*, “Fabrication of carbon nanotube field-effect transistors in commercial silicon manufacturing facilities,” *Nature Electronics*, vol. 3, no. 8, pp. 492–501, 2020.

[81] B. Mudassar and S. Yalamanchili, “Heterogeneous integration for artificial intelligence: Challenges and opportunities,” *IBM Journal of Research and Development*, vol. 63, no. 6, pp. 1–23, 2019.

[82] H. S. Kum *et al.*, “Heterogeneous integration of single-crystalline complex-oxide

membranes," *Nature*, vol. 578, no. 7793, pp. 75–81, 2020.

[83] G. Hills *et al.*, "Modern microprocessor built from complementary carbon nanotube transistors," *Nature*, vol. 572, no. 7771, pp. 595–602, 2019.

[84] M. D. Bishop *et al.*, "Fabrication of carbon nanotube field-effect transistors in commercial silicon manufacturing facilities," *Nature Electronics*, vol. 3, no. 8, pp. 492–501, 2020.

[85] S. Dodge and L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," *2017 26th International Conference on Computer Communications and Networks, ICCCN 2017*, 2017.

[86] T. Brooks, B. Mildenhall, T. Xue, J. Chen, Di. Sharlet, and J. T. Barron, "Unprocessing images for learned raw denoising," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 11028–11037, 2019.

[87] D. Charte, F. Charte, S. García, M. J. del Jesus, and F. Herrera, "A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines," *Information Fusion*, vol. 44, pp. 78–96, 2018.

[88] L. Xu, M. Cao, B. Song, J. Zhang, Y. Liu, and F. E. Alsaadi, "Extracting and Composing Robust Features with Denoising Autoencoders," *Proceedings of the 25th international conference on Machine learning (ICML)*, vol. July, pp. 1096–1103, 2008.

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