Epitaxial SiGe Synapses for Neuromorphic Arrays

by

Scott H. Tan

B.A., Pomona College (2016)

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

© Massachusetts Institute of Technology 2018. All rights reserved.

Signature redacted

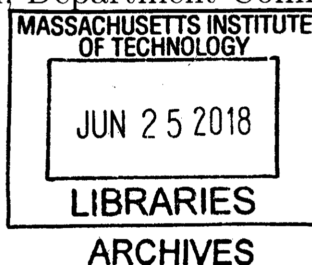
Author . . . .
Department of Mechanical Engineering
May 11, 2018

Signature redacted

Certified by . . . .
Jeehwan Kim
Associate Professor
Thesis Supervisor

Signature redacted

Accepted by . . . .
Rohan Abeyaratne
Chairman, Department Committee on Graduate Theses


MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUN 25 2018
LIBRARIES
ARCHIVES

Epitaxial SiGe Synapses for Neuromorphic Arrays

by

Scott H. Tan

Submitted to the Department of Mechanical Engineering
on May 11, 2018, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Intelligent machines could help to facilitate language translation, maximize attentive learning, and optimize medical care. However, hardware to train and deploy AI systems are power-hungry and too slow for many applications. Neuromorphic arrays could potentially offer better efficiency compared to conventional hardware by storing high-precision analog weights between digital processors. However, neuromorphic arrays have not experimentally demonstrated learning accuracy comparable to conventional hardware due to irreproducibility associated with existing artificial synapses. Large variation arises in conventional devices due to the stochastic nature of metal movement through an amorphous synapse. Hence, passive arrays have only been demonstrated as small-scale systems.

In this thesis, I developed single-crystalline Silicon-Germanium (SiGe) artificial synapses that have suitable properties for large-scale neuromorphic arrays. In contrast to amorphous films, epitaxially-grown SiGe can confine metal filaments within widened threading dislocations for uniform conductance update thresholds. Metal confinement reduces temporal variation to as low as 1%, which is the lowest variation reported to date, to the extent of the author's knowledge. Dislocations are selectively etched to allow for high ON/OFF ratio, good retention, many cycles of endurance, and linear conductance change. Simulations accounting for non-ideal device properties suggest that SiGe synapses in passive crossbar arrays could perform supervised learning for handwriting digit recognition with up to 95.1% accuracy. Hence, SiGe synapses demonstrate great promise for large-scale neuromorphic arrays.

Thesis Supervisor: Jeehwan Kim
Title: Associate Professor

Acknowledgments

I would like to acknowledge and thank Professor Jeehwan Kim, Dr. Shinhyun Choi, Dr. Hanwool Yeon, Dr. Peng Lin, Yunjo Kim, Zefan Ki, Chanyeol Choi, Pai-Yu Chen, and Professor Shimeng Yu for their contributions to this thesis. I also would like to thank Professor Joshua Yang, Professor Qiangfei Xia, Yunning Li, Mingyi Rao, and Ye Zhuo for fruitful discussions and valuable help. Thank you to the National Science Foundation, Alfred P. Sloan Foundation, Lemelson Foundation, and American Indian Science and Engineering Society for funding.

Contents

1 Introduction **15**

1.1 Motivation 15

1.2 Neuromorphic Arrays 16

1.3 Artificial Synapses 19

1.4 Overview of Thesis 20

2 Silicon-Germanium Epitaxy **21**

2.1 Background 21

2.1.1 Silicon Germanium 21

2.1.2 Low-Pressure Chemical Vapor Deposition 22

2.2 Threading Dislocations in Epitaxial SiGe 23

2.2.1 Metastable SiGe 23

2.2.2 Relaxed SiGe 24

3 Epitaxial Artificial Synapses **29**

3.1 Background 29

3.1.1 Device Architecture 29

3.1.2 Current-Voltage Analysis 30

3.1.3 Pulse Measurements 31

3.2 Filament Dynamics in Epitaxial Synapses 31

3.3 Widening Threading Dislocations 33

3.4 Variations 34

3.5 Bandgap Engineering 35

3.6 Nano-Scale Synapses 35
3.7 Analog Measurements 36
3.8 Retention 38
3.9 Artificial Neural Network Simulations 39
3.9.1 Weight Update 40
3.10 Handwriting Recognition 42

4 Conclusions **57**

List of Figures

1-1 Deep learning on Neuromorphic Arrays. Input data is classified by a trained network. In this example, a handwritten digit is input as voltage pulses and processed by the neuromorphic arrays. The maximum current output is from the column representing the correct number. Synaptic weights at each crosspoint (conductance values) are adjusted during training. Peripheral circuitry (not depicted) compute error and apply activation functions between each layer. 17

2-1 Critical Thickness Curves for SiGe epitaxy on Si. At 900^oC, growth above the equilibrium critical thickness permits strain relaxation. At 750^oC, a metastable region arises above the equilibrium critical thickness and below the thermodynamic critical thickness. In this regime, misfits accompanied by threading dislocation half-loops are generated, but dislocation glide for strain relaxation is kinetically limited. Curves adapted from Houghton *et al.*[38] 22

2-2 SiGe epitaxy at kinetically-limiting temperatures (750 ^oC). a) Before epitaxy, Silane and Germane decompose and adsorb onto the substrate. b) Pseudomorphic growth occurs as SiGe is compressively strained, which cause surface roughening. c) Metastable SiGe films are grown above the equilibrium critical thickness and contain many TDs. d) Relaxed SiGe films are thicker than the kinetically-limited critical thickness and have fewer TDs since threading dislocation glide is no longer suppressed. 25

2-3 Dislocation density analysis of heteroepitaxial SiGe on Si. The dislocation density is in the range of $10^{11} cm^{-2}$. This suggests that the epitaxial SiGe synapses can be scaled down to tens of nanometers. a) SEM image of decorated TDs in $1\mu m \times 1\mu m$ area. b) Magnified SEM with dislocation pinholes highlighted by red circles. In the $200nm \times 200nm$ area, 75 dislocations can be observed c) Color map highlighting the distribution of TDs across the entire $1\mu m \times 1\mu m$ area. d) Histogram showing the dislocation counts in 5 $\mu m \times 5\mu m$ area. 26

2-4 Threading Dislocations in SiGe with different film thickness and Ge concentrations. a) SEM of etched TDs in 60-nm $Si_{0.9}Ge_{0.1}$. b) SEM of etched TDs in 60-nm $Si_{0.7}Ge_{0.3}$ appears to have less dislocations than 60-nm $Si_{0.9}Ge_{0.1}$. c) Transmission electron microscopy (TEM) of $Si_{0.7}Ge_{0.3}$ reveals strain fields from TDs. d) SEM of etched threading dislocations in 300-nm $Si_{0.9}Ge_{0.1}$ appears to have fewer TDs than 60-nm $Si_{0.9}Ge_{0.1}$. 27

3-1 Device architecture of Epitaxial SiGe synapses. a) Illustration of the device cross-section. b) Plan-view optical microscopy image of a device. Scale bar: 40 μm. c) Cross-section SEM of an epitaxial SiGe synapse. Scale bar: 100 nm. 30

3-2 Resistive switching in SiGe. a) I-V measurements of epitaxial SiGe synapses (with unetched dislocations) and of Ag/crystalline i-Si/crystalline p-Si device, where no hysteresis is observed. The signature of filament rupture is highlighted in b). c) HRS and LRS states measured at 0.8 V for 100 cycles. d) Set voltage measured over 100 quasi-static I-V sweeps. Inset: histogram for set voltage distribution. e) Cu/SiGe/p+ Si and f) Ni/SiGe/p+ Si devices show very little hysteresis and high current, likely due to formation of stable conductive compounds. . . . 44

3-3 Ag/Amorphous-Si(a-Si)/p+ Si switching devices. a) I-V measurements for a-Si devices. b) Set voltage measured over 100 quasi-static I-V sweeps. Inset: histogram for set voltage distribution. Set threshold: 44 μA. c) Retention for a-Si device. 45

3-4 Effect of etching on epitaxial SiGe synapses. a,f,k,p) Plan-view SEM images reveal increasing etching time widens dislocations. Scale bar: 200 nm. b,g,l,q) I-V measurements for different etch times. c,h,m,r) HRS and LRS temporal evolution shows etching increases the on/off ratio, but over-etching increases variation of the HRS state. d,i,n,s) Set voltage temporal evolution shows increasing variation with etch time. e,j,o,t) Histograms of set voltage for different etch times. 45

3-5 Complete Ag filament rupture is enabled by defect-selective etching. a) Set voltage and variation plotted vs. etching time shows that as etch time increases, set voltage decreases and variation increases. b) Semi-logarithmic I-V characteristics of the reset process without etching (0s) and after 5s of etching plotted with i-Si shows larger decrease in current is enabled by widening TDs. c) Semi-logarithmic DC I-V characteristics without etching (0s) and after 5s of etching. d) Linear-scale DC I-V characteristics without etching (0s) and after 5s of etching. e) I-V characteristics of SiGe in the virgin state and after Ag filament rupture, showing nearly-complete reset to the HRS is allowed with widened TDs. f) Higher analog on/off ratio is observed for SiGe with widened TDs. 46

3-6 Characteristics of SiGe artificial synapses after widening TDs with defect-selective etching. a) I-V measurements of the 1st and 750th cycle showing devices initially have higher set voltage, lower current level, and less effective reset. b) The temporal evolution of set voltage for a device for over 700 quasi-static I-V sweeps. Insets: DC I-V plots at 250th cycle and 500th cycle, along with the histogram for set voltage distribution. c) Map of set voltages at devices overlapped with optical images of devices (50 devices from batch 1 and 50 devices from batch 2 are shown). d) Histogram for spatial set voltage distribution. 47

3-7 Bandgap engineering to tune characteristics of epitaxial SiGe synapses. a) I-V measurements with different doping concentrations of p+ Si below SiGe. b) Set voltage and read current (at 0.8 V) plotted as a function of doping concentration suggests that higher doping concentration results in lower set voltage and higher read current due to the lower Schottky barrier. c) Linear-scale and d) Logarithmic-scale I-V curve of p-i-p SiGe back-to-back epitaxial SiGe synapse. LRS-state current is rectified at negative bias. 48

3-8 I-V measurements of nano-scale epitaxial SiGe devices. a) $25nm \times 25nm$, b) $50nm \times 50nm$, c) $75nm \times 75nm$, d) $100nm \times 100nm$, and e) $125nm \times 125nm$ devices have similar hysteresis to f) $5\mu m \times 5\mu m$ devices. 49

3-9 Variation of nano-scale epitaxial SiGe synapses. a) Temporal variation of set voltage of a $25nm \times 25nm$ device cycled over 700 times. b) Histogram showing the spatial variation of set voltage for 100 $25nm \times 25nm$ devices. 49

3-10 Digital and analog access of different conductance levels. a) I-V measurements with different current compliance can set LRS to different levels. b) Read current of LRS increases with current compliance since a more conductive (stronger) channel can be formed. c) Pulsing scheme consisting of gradually increasing set pulses spaced with read pulses allows d) analog set voltage to be determined. e) Pulsing scheme consisting of gradually decreasing reset pulses allows f) analog reset voltage to be determined. 50

3-11 Potentiation-depression (P-D) of epitaxial SiGe synapses. a) Read current in response to 100 set pulses and 100 reset pulses. b) Pulsing scheme used to for P-D measurements. Voltage pulses with amplitude above V_{SET} increase the weight value ($+w_{ij}$) and amplitude below V_{RESET} decrease the weight value ($-w_{ij}$), while weaker pulses that do not change the conductance ($\delta w_{ij} = 0$) are applied to obtain the read current. 51

3-12 Analog characteristics of epitaxial SiGe synapses. a) Potentiation-depression (P-D) with and without widenened TDs showing widened TDs promote larger analog on/off ratio. Table: Analog set and reset voltages. b) P-D shows analog on/off ratio increase with number of applied pulses. The pulse train consists of 100/200/500 consecutive set pulses (5 V, 5 μs) followed by 100/200/500 consecutive reset pulses (-3 V, 5 μs), respectively. Current is measured by read pulses (2 V, 1 ms) after each set/reset pulse. c) The definition of non-linearity magnitude d) The non-linear magnitude depending of number of P-D pulses. Inset: table showing non-linearity magnitudes for 200, 400, and 1000 P-D pulses. e) Endurance measured for 10^9 set/reset pulses (10^6 cycles). The first and fifth of each order of magnitude is shown. Each P-D cycle consists of 500 consecutive potentiation pulses and 500 consecutive depression pulses with read pulses between each update. After 10^6 cycles, similar P-D can still be observed. 52

3-13 Retention of SiGe artificial synapses. a) Untreated TDs show poor retention. Inset: plan-view SEM of relatively flat SiGe. b) The two-day retention test at 85°C at LRS for a SiGe artificial synapse etched for 5s. The device performance remains unchanged after the test. Inset: plan-view SEM of 5s etched SiGe. c) Over-etched TDs also show poor retention. Inset: plan-view SEM of 10s etched SiGe. a-c are measured with 1.5 V read pulses with pulse width of 1 ms. d) Retention tests for 5s etched SiGe at elevated temperatures. The extrapolation of the plot to room temperature indicates 1.87 years of retention. Measurements are collected twice from 10 devices at each temperature between 398K to 443K. 53

3-14 Schematic for the image recognition simulation. a) A three-layer MLP neural network with a black-and-white input signal for each layer in the algorithm level. The inner product (summation) of the input neuron signal vector and the first synapse array vector is transferred after activation and binarization as the input vector for the second synapse array. b) Circuit block diagram of a neuromorphic crossbar array and the peripheral circuits. MUX, multiplexer; ADC, analog-to-digital converter. c) Diagram illustrating the process for determining outputs for prediction and delta-weight calculation. d) Schematic of potentiation (weight increase) and depression (weight decrease) phases during training. The number of write or erase pulses for each synapse is determined by the delta-weight calculation. Figures are adapted from [15] 54

3-15 Influence of non-ideal device parameters for MLP simulation. The influence of a) Device-to-Device variation, b) cycle-to-cycle variation, c) Read noise, and d) number of conductance levels, e) wire resistance, and f) on/off ratio on the final MNIST digit recognition accuracy. Measured values for epitaxial SiGe synapses are shown in red. e) Evolution of accuracy to training epochs for ideal software and for neuromorphic arrays with epitaxial SiGe synapses. 55

Chapter 1

Introduction

1.1 Motivation

Recent success of modern artificial intelligence (AI) can largely be attributed to the advancement of deep learning[25, 24, 53, 36]. Today, artificial neural networks are already used in many applications, including speech recognition[31, 98, 36], object recognition[53, 85, 92], robotics[67], and decision-making[32, 79, 88, 7, 84]. However, next-generation AI-enabled technologies such as real-time data analytics, natural language translation, automated transportation, and multimodal IoT (Internet of Things) sensor processing systems require advancements of AI algorithm and hardware to reduce large power requirements.

Today, AI is almost synonymous with multi-layer artificial neural networks, popularly known as deep learning[56, 81]. Training these artificial neural networks involves adjustment of connection strengths (synaptic weights) between layers of neurons to reduce dimensionality or minimize error functions. Out of the many existing learning algorithms[77, 34, 6, 35, 37], stochastic gradient decent is by-far the most widely used method for supervised learning [78, 61, 55, 54].

The main computation executed in deep learning is the weighted summation. Graphics Processing Units (GPUs) have become the most popular hardware platform for accelerating deep learning since they can handle many operations in parallel[2]. GPUs process data in single-instruction multiple threads (SIMT) that use central-

ized control of many parallel arithmetic logic units (ALUs) that fetch data from the memory hierarchy[91]. However, excessive power is lost in data movement between memory and ALUs. Application-specific integrated circuit (ASIC) accelerators are being developed to minimize power consumption[33, 75, 16, 18, 19, 17, 48]. Also, field-programmable gate array (FPGA) accelerators are used to optimize computing engines and minimize memory bandwidth usage[60, 21, 109, 11].

In addition to these approaches, crossbar arrays composed of two-terminal resistive switching devices have great potential to minimize power requirements and speed up operation since the same hardware can be used for both memory and processing. Artificial synapses learn optimal synaptic weight values as conductance[89, 58]. Crossbar arrays of artificial synapses physically implement weight storage between two fully-connected neuron layers. Hence, these arrays have been coined "neuromorphic," meaning they are inspired by neuronal networks. Neuromorphic arrays in deep learning hardware can potentially achieve up to 4 orders of magnitude lower power consumption[63, 64, 30, 28]. Fig. 1-1 illustrates the general crossbar structure for neuromorphic arrays.

1.2 Neuromorphic Arrays

Neuromorphic arrays are capable of two important functions: 1) vector-matrix multiplication and 2) analog weight update. For a fully-connected array with n input neurons and m output neurons, each artificial synapse weight value is represented by the electrical conductance between two neurons, mathematically represented by a matrix with m rows and n columns, $G_{m\times n}$. Input voltage pulses, V_n, enter top crossbar rows. According to Kirchoff's law, the total current at bottom crossbar columns, I_m, are weighted summations according to the synapse conductance values and the input voltage pulses. The vector-matrix (inputs-weights) multiplication is expressed in Eqn. 1.1:

$$I_m = G_{m\times n}V_n \tag{1.1}$$

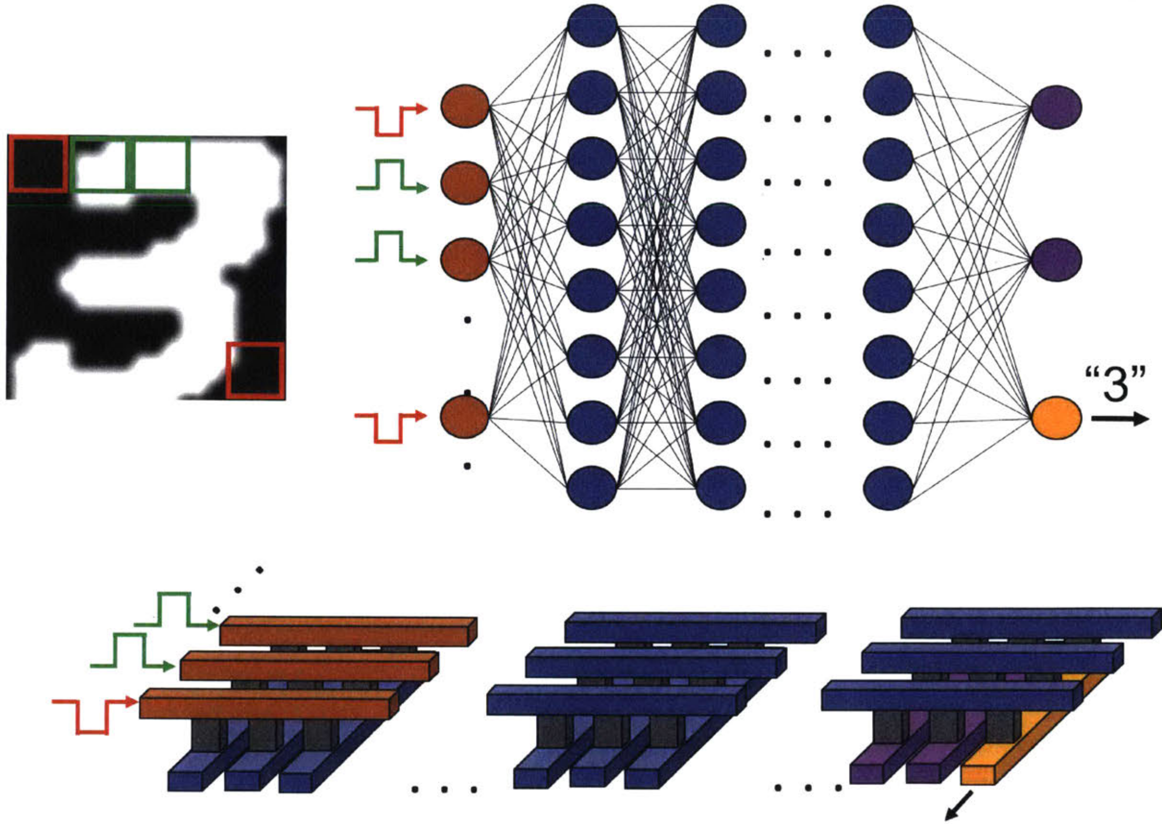


Figure 1-1: Deep learning on Neuromorphic Arrays. Input data is classified by a trained network. In this example, a handwritten digit is input as voltage pulses and processed by the neuromorphic arrays. The maximum current output is from the column representing the correct number. Synaptic weights at each crosspoint (conductance values) are adjusted during training. Peripheral circuitry (not depicted) compute error and apply activation functions between each layer.

The integrated current output is converted to a digital signal to be processed in peripheral circuitry. An activation function, $y = f(I_m)$, is used in deep learning to introduce non-linearity. A common activation function is the logistic sigmoid, which is expressed as

$$y = f(I_m) = \frac{1}{1 + e^{-I_m}} \tag{1.2}$$

The result of this activation function is converted to voltage pulse inputs for the next (hidden) layer of synaptic weights, G_L. The output y is the result of the activation function f_L acting on the vector-matrix product of G_L and the outputs of

the first array, expressed as

$$y = f_L(G_L f(I_m)) \tag{1.3}$$

A deep neural network is composed of multiple hidden layers with activation functions between each layer. Known as a multi-layer perceptrion (MLP), this neural network can be mathematically expressed as[56]

$$y = f_\theta(V_n) = f_{L+1}(G_{L+1} f_L(G_L ... f(I_n)...) \tag{1.4}$$

Using deep learning, neuromorphic arrays can approximate any function $y = f_\theta(V_n)$. To perform classification, artificial synapses (weights) are trained to minimize the error when the approximate output function is compared to values of a label.

The second critical function of artificial synapses in a neuromorphic array is analog reconfigurability of conductance states. In order for the neural network to be trained, artificial synapses must be able to access conductance levels representing different synaptic weight values. During training, the conductance of an artificial synapse relaxes towards a value that converges the error of the training set towards a global minimum. In crossbar arrays, conductance level increase (potentiation) occurs when voltage across the artificial synapse exceeds a SET threshold. On the other hand, oppositely-polarized voltage with amplitude exceeding a RESET threshold lowers the artificial synapse conductance (called depression). After partial derivatives of error for each output are propagated backwards and calculated for each neuron, weights can can be updated by the delta rule[78]

$$w_{ij,new} = w_{ij,old} + \eta x_i \delta_j \tag{1.5}$$

where w_{ij} is the synaptic weight value for the ith row and j^{th} column, η is the learning rate, x_i is the activity at the neuron input, and δ_j is the partial derivative of error computed for neuron j. This weight update scheme can be accomplished with post-processing circuitry including comparators and integrators[10, 72, 107, 9, 59, 99, 82, 86, 15]. Weight update increments can be converted to voltage pulse trains to

modify conductance values of artificial synapses[29, 30].

1.3 Artificial Synapses

Various types of analog switching devices have been demonstrated as synapses for neuromorphic computing[45, 8, 10, 95, 51, 57, 30]. Most rely on filamentary switching mechanisms, such as oxide-based resistive random access memory (oxide-based RRAM) and conductive-bridging RAM (CBRAM). Oxide-based RRAM operation is based on alignment of anion vacancies inherent in amorphous-phase binary oxides to form conductive filaments[57, 83, 68, 108, 90]. While these devices exhibit reasonably good retention and endurance, they suffer from small on/off ratio and unavoidable temporal (cycle-to-cycle) and spatial (device-to-device) variation due to uncontrollable filament dynamics in an amorphous solid[51, 57, 83, 68, 108, 90]. Resistive switching using single crystalline-based ternary oxide films have been attempted, where dislocations become active filaments due to self-doping effect of crystalline defects in $SrTiO_3$[93]. However, analog weight update using these devices has not been reported to the extent of the authors' knowledge. CBRAM operation is based on metal conductive bridging through an amorphous solid electrolyte [95, 50, 102, 46, 47, 96]. Because metal cations are more mobile than oxygen vacancies, the ON/OFF ratio for CBRAM are substantially higher than that of the oxide-based RRAM[95, 101, 43, 100]. However, uncontrollable ion transport through defects in amorphous films results in three-dimensional stochastic filament formation. This results in large temporal switching threshold variations[50, 102, 101, 52]. These variations make large-scale analog neural computing impractical without transistors at each artificial synapse. Thus, securing a strategy to better control metal movement in artificial synapses is an essential step towards achieving deep learning on passive neuromorphic arrays [4].

1.4 Overview of Thesis

This thesis focuses on epitaxial SiGe synapses for neuromorphic arrays. Analog conductance change is achieved by utilizing enhanced ion transport and one-dimensional conduction channel confinement in engineered dislocations. Threading dislocation density is maximized in metastable SiGe films[87], which allows for scaling as small as $25nm \times 25nm$ for 60 nm-thick $Si_{0.9}Ge_{0.1}$ grown on p+ Si substrates.

High on/off ratio, minimal spatial and temporal variations, long retention, and good endurance suggests that epitaxial SiGe synapses could be suitable for transistor-free neuromorphic computing arrays.

In addition, the epitaxy of p-i-p back-to-back diodes in epitaxial SiGe synapses permits self-selection behavior that can suppress sneak path during large-scale array operation. Precise doping modulation during epitaxy allows for modulation of set voltage and read current by varying the Schottky barrier height at Ag/Si interface. Simulations based on characteristics of epitaxial SiGe synapses shows 95.1% accurate supervised learning with the MNIST handwritten recognition dataset, which is comparable to software training baseline of 97%.

When nano-scale devices are imaged via cross-sectional transmission electron microscopy (TEM) after conduction channel formation, Ag conduction channels confined in engineered dislocations can be visually observed.

The next chapter describes heteroepitaxy to grow metastable Silicon-Germanium films. Chapter 3 covers fabrication and characterization of epitaxial SiGe synapses. Finally, conclusions and future work are discussed in Chapter 4.

Chapter 2

Silicon-Germanium Epitaxy

Epitaxially-grown films with threading dislocations are suitable materials for the switching layer in artificial synapses. This section will describe epitaxial growth of Silicon-Germanium (SiGe) films on p+ Silicon (p+ Si) substrates. First, the growth mechanics of heteroepitaxial SiGe will be discussed. Then, experimental results for metastable SiGe films will be presented.

2.1 Background

2.1.1 Silicon Germanium

Silicon (Si) and Germanium (Ge) both crystallize in diamond cubic structures with lattice constants of 5.431 Å and 5.658 Å, respectively. Because Ge is slightly larger than Si, Silicon-Germanium (SiGe) alloys have a larger lattice constant than pure Si. The lattice constant for SiGe, $a_{Si_{1-x}Ge_x}$, can be approximated by Vengard's law[94]

$$a_{Si_{1-x}Ge_x} = a_{Si}(1-x) + a_{Ge}x \tag{2.1}$$

where a_{Si}, and a_{Ge} are the lattice constants of Si and Ge, respectively, and x is the atomic percentage of Ge. Since Ge is slightly larger, there is some lattice mismatch between SiGe and Si. As a result thin SiGe films are compressively strained at the onset of heteroepitaxial growth.

2.1.2 Low-Pressure Chemical Vapor Deposition

Single-crystalline SiGe can be epitaxially grown on Si substrates by low-pressure chemical vapor deposition (LPCVD). During LPCVD, Silane and Germane gas precursors, carried by Hydrogen carrier gas, enter a heated ($\sim 750\ ^oC$) low-pressure ($\sim$ 100 Torr) close-coupled showerhead reactor. Low pressure allows for uniform film growth. At high temperatures, Silane and Germane decompose into Si, Ge, and H_2. Si and Ge adatoms adsorb onto the surface, forming a SiGe film. Si and Ge adatoms tend to align with the diamond cubic structure of the Si substrate and mimic the single-crystalline structure of the underlying substrate.

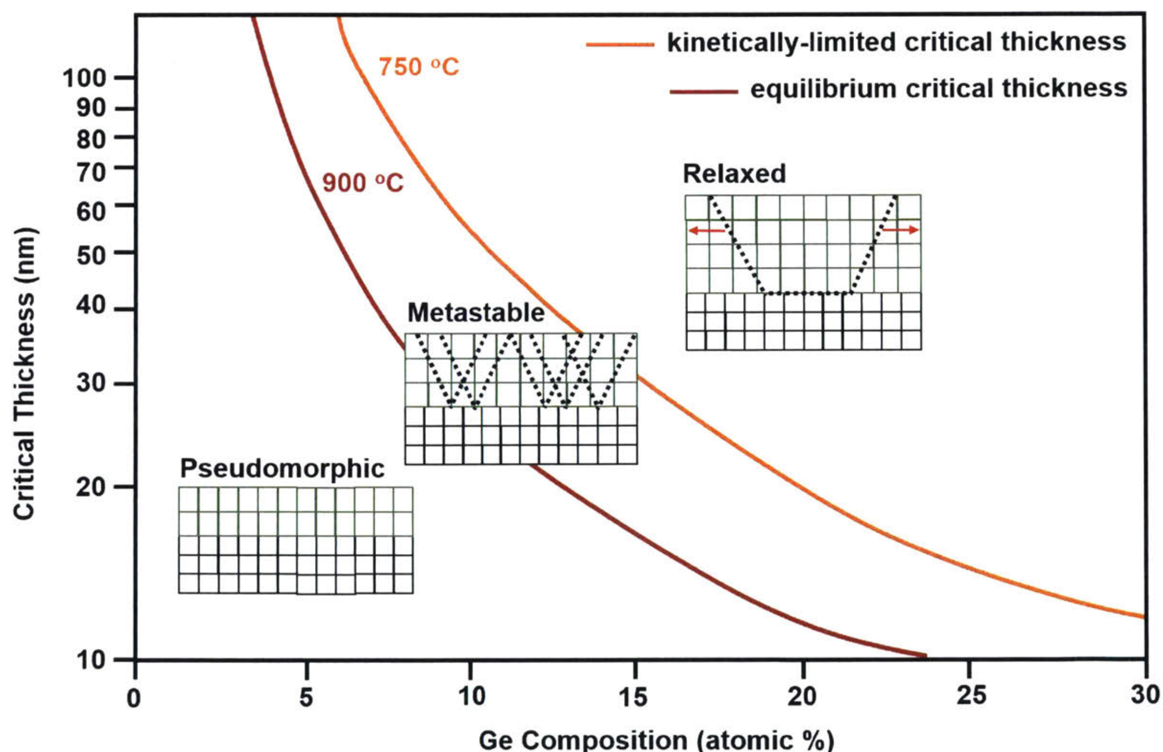


Figure 2-1: Critical Thickness Curves for SiGe epitaxy on Si. At 900^oC, growth above the equilibrium critical thickness permits strain relaxation. At 750^oC, a metastable region arises above the equilibrium critical thickness and below the thermodynamic critical thickness. In this regime, misfits accompanied by threading dislocation half-loops are generated, but dislocation glide for strain relaxation is kinetically limited. Curves adapted from Houghton *et al.*[38]

As the SiGe film grows thicker, strain energy increases. When SiGe film thickness reaches the thermodynamic equilibrium critical thickness, h_c, the outward force due

to film stress overcomes the tension of a dislocation. The thermodynamic equilibrium critical thickness, first reported by Matthews and Blakeslee, can be calculated by [65, 66]

$$h_c \approx \frac{b(1-\nu\cos\theta^2)}{4\pi(1+\nu)\epsilon} ln(\frac{h_c}{b}) \quad (2.2)$$

where b is the Burger's vector magnitude ($b_{SiGe} = 3.9\,\text{Å}$), ν is Poisson's ratio ($\nu_{SiGe} = 0.28$[42]), θ is the angle of a threading dislocation in the 111 glide plane ($\theta_{SiGe} = 60^o$), and ϵ is the strain mismatch ($\epsilon = \frac{a_{Si}-a_{SiGe}}{a_{Si}}$). Above h_c, coherency of pseudomorphic films breaks down, leading to the nucleation of misfit dislocation cores.

Misfit generation must be accompanied by threading dislocations (TDs) that terminate at a free surface. Solely based on intrinsic material properties, homogeneous nucleation of dislocation TD half-loops is not possible for low-mismatch systems [23, 20, 44]. However, it is known that surface roughening occurs during heteroepitaxial growth, which nucleates dislocation half-loops by material rearrangement (without any gliding)[69, 26]. This heterogeneous nucleation in $Si_{1-x}Ge_x$ on Si ($x < 0.5$) is effectively barrier-less [44].

Under low temperature growth conditions, kinetic limitations to TD glide are imposed. This results in a kinetic critical thickness above the thermodynamic critical thickness [39], which must be exceeded for TDs to glide and relax strain. Epitaxial SiGe grown below the kinetic critical thickness and above the thermodynamic critical thickness is metastable since TDs nucleate, but do not glide outward. As a result, metastable films are expected to contain more TDs than pseudomorphic films or relaxed films[39, 87].

2.2 Threading Dislocations in Epitaxial SiGe

2.2.1 Metastable SiGe

In our experiments, we found 60-nm-thick metastable $Si_{0.9}Ge_{0.1}$ has high TD density on the order of $10^{11} cm^{-2}$, as shown in Fig. 2-3. Defect-selective etching was used to

decorate TDs so that they are visible using scanning electron microscopy (SEM).

2.2.2 Relaxed SiGe

Increasing the atomic percentage of Germanium increases strain due to lattice mismatch. Because threading dislocations gliding in opposite directions along the same atomic plane can annihilate, $Si_{0.7}Ge_{0.3}$ has a lower dislocation density compared to $Si_{0.9}Ge_{0.1}$, as shown in Fig. 2-4 a and b, respectively. In $Si_{0.7}Ge_{0.3}$, the strain field from TDs can be observed as diagonal extensions through the epitaxial layer, as shown in Fig. 2-4 c. Increasing the thickness of the heteroepitaxial layer also increases strain. Hence, increasing thickness also lowers dislocation density, as observed in Fig. 2-4 d.

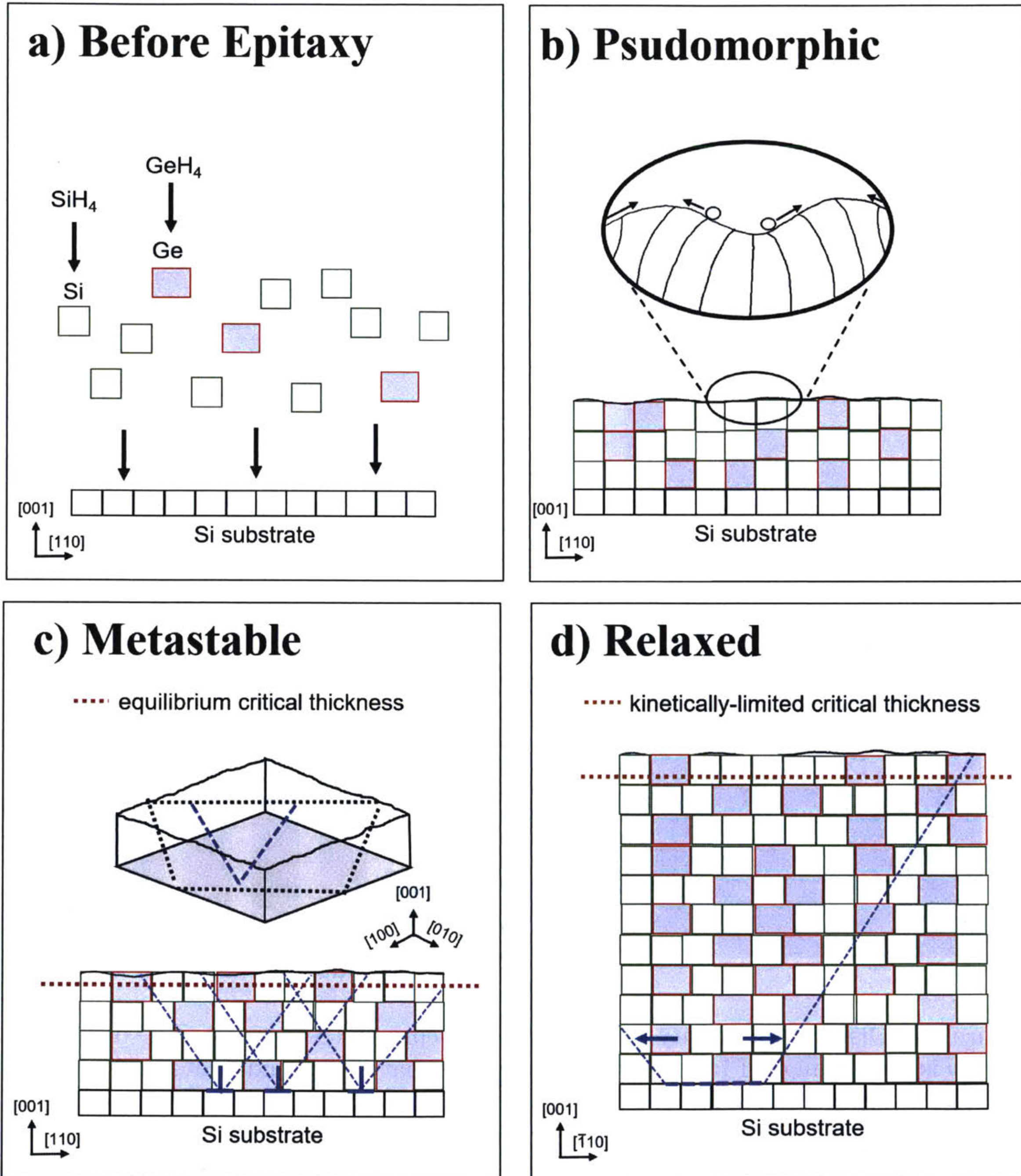


Figure 2-2: SiGe epitaxy at kinetically-limiting temperatures (750 ^oC). a) Before epitaxy, Silane and Germane decompose and adsorb onto the substrate. b) Pseudomorphic growth occurs as SiGe is compressively strained, which cause surface roughening. c) Metastable SiGe films are grown above the equilibrium critical thickness and contain many TDs. d) Relaxed SiGe films are thicker than the kinetically-limited critical thickness and have fewer TDs since threading dislocation glide is no longer suppressed.

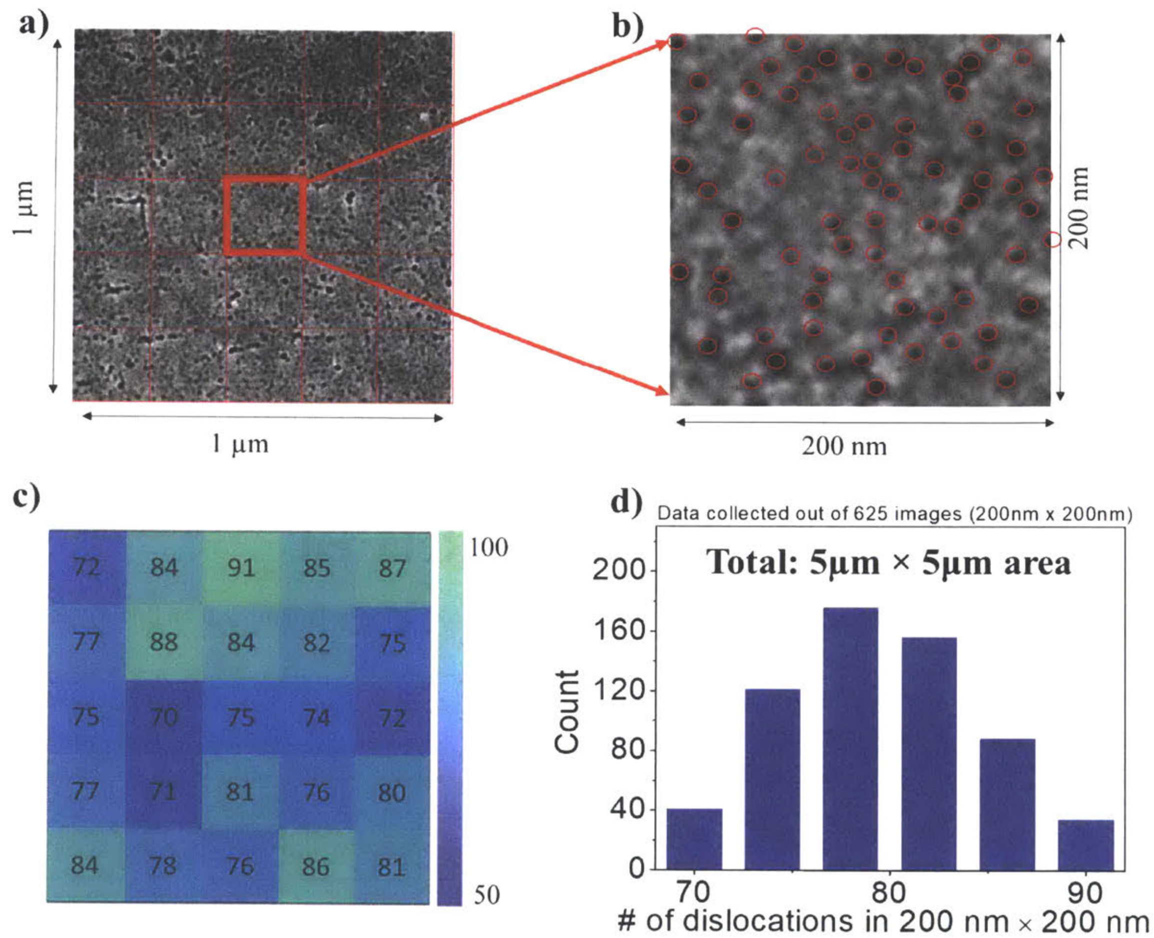


Figure 2-3: Dislocation density analysis of heteroepitaxial SiGe on Si. The dislocation density is in the range of $10^{11} cm^{-2}$. This suggests that the epitaxial SiGe synapses can be scaled down to tens of nanometers. a) SEM image of decorated TDs in $1\mu m \times 1\mu m$ area. b) Magnified SEM with dislocation pinholes highlighted by red circles. In the $200nm \times 200nm$ area, 75 dislocations can be observed c) Color map highlighting the distribution of TDs across the entire $1\mu m \times 1\mu m$ area. d) Histogram showing the dislocation counts in 5 $\mu m \times 5\mu m$ area.

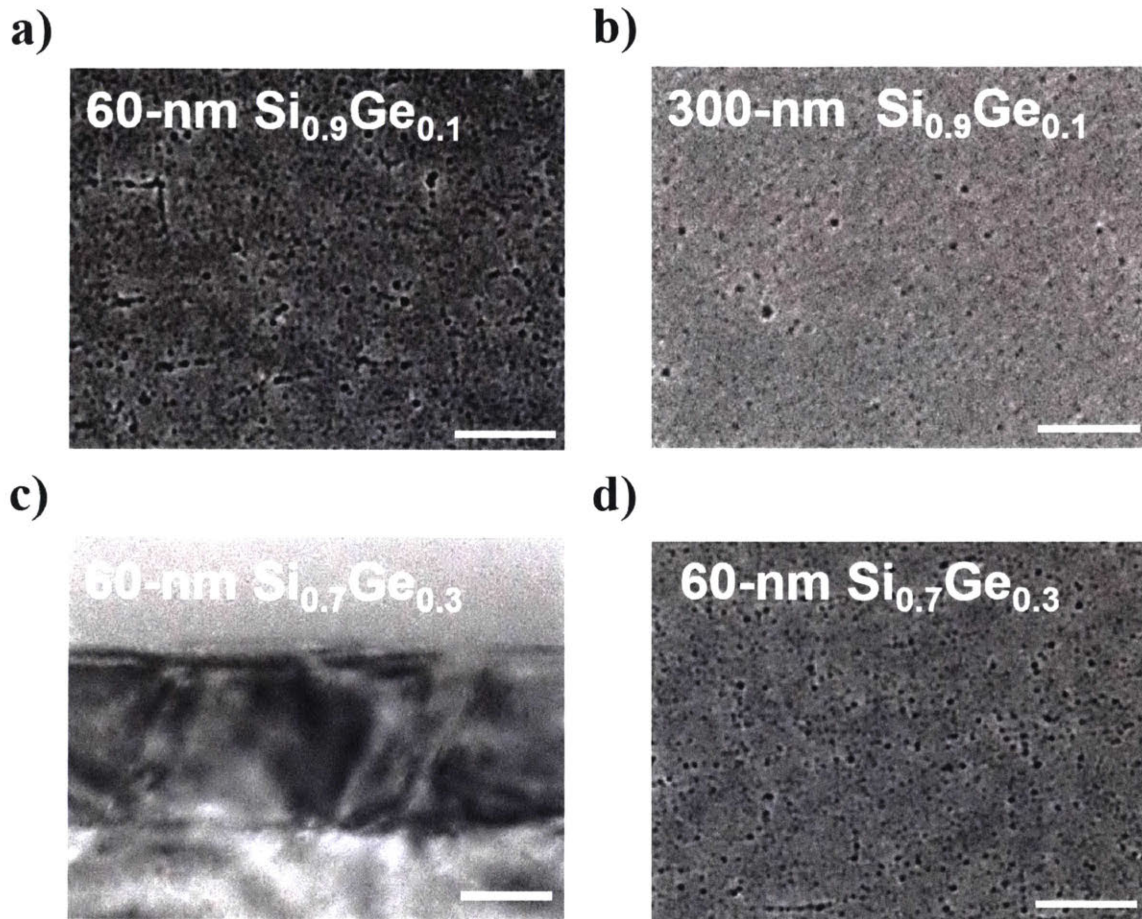


Figure 2-4: Threading Dislocations in SiGe with different film thickness and Ge concentrations. a) SEM of etched TDs in 60-nm $Si_{0.9}Ge_{0.1}$. b) SEM of etched TDs in 60-nm $Si_{0.7}Ge_{0.3}$ appears to have less dislocations than 60-nm $Si_{0.9}Ge_{0.1}$. c) Transmission electron microscopy (TEM) of $Si_{0.7}Ge_{0.3}$ reveals strain fields from TDs. d) SEM of etched threading dislocations in 300-nm $Si_{0.9}Ge_{0.1}$ appears to have fewer TDs than 60-nm $Si_{0.9}Ge_{0.1}$.

Chapter 3

Epitaxial Artificial Synapses

Resistive switching devices are promising candidates for artificial synapses in neuromorphic hardware. This chapter describes fabrication and performance of artificial synapses using heteroepitaxial Silicon-Germanium (SiGe) as the switching layer. The confinement of the conducting filament into widened dislocations in SiGe offers superior spatial and temporal uniformity, long retention, excellent endurance, high on/off ratio, and linear weight update. In addition, bandgap engineering of layers during epitaxial growth could be used to customize device characteristics. According to simulations, epitaxial SiGe-based neuromorphic arrays can achieve 95.1% learning accuracy. The development of epitaxial SiGe artificial synapses is a step towards realizing fully-functioning passive large-scale neuromorphic arrays.

3.1 Background

3.1.1 Device Architecture

Metastable 60-nm SiGe is epitaxially grown on p+ Si substrates, as described in Ch. 2. On the backside of substrates, 100-nm of Al is evaporated and annealed at 450 ^{o}C to form ohmic contact. To isolate devices, 100-nm SiO_2 is deposited on top of SiGe by plasma-enhanced chemical vapor deposition (PECVD) and etched back by buffered oxide etch (BOE 5:1) solution after photolithography patterning. To create

nano-scale devices, patterning is done by electron-beam lithography and reactive-ion etching (RIE) is used to etch back SiO_2. 100-nm of Ag and 20-nm of Pd are evaporated as top electrodes. 5-nm of Ti and 100-nm of Au are evaporated as contact pads. Cross-sectional SEM of the completed device and plan-view optical microscopy images are shown in Fig. 3-1.

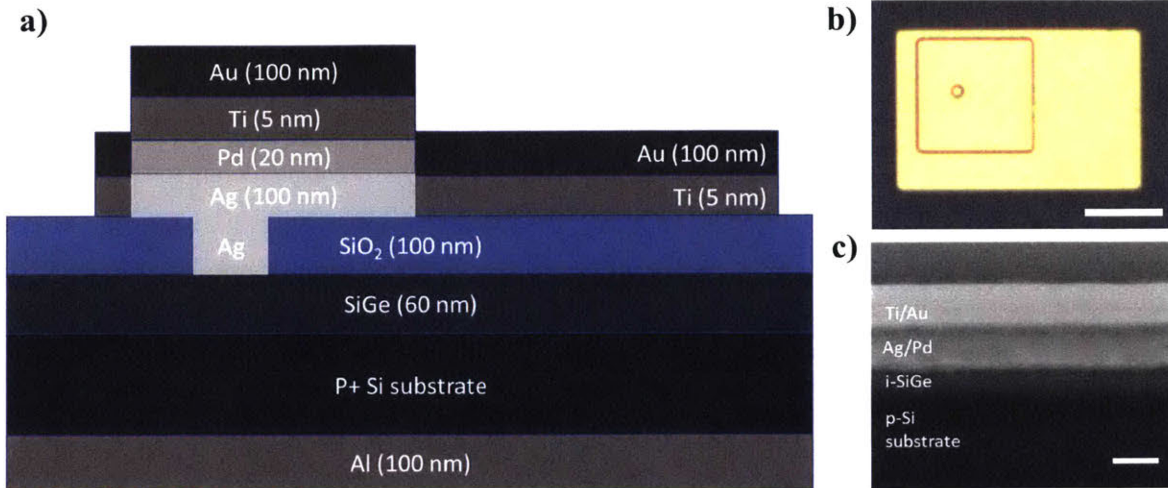


Figure 3-1: Device architecture of Epitaxial SiGe synapses. a) Illustration of the device cross-section. b) Plan-view optical microscopy image of a device. Scale bar: 40 μm. c) Cross-section SEM of an epitaxial SiGe synapse. Scale bar: 100 nm.

3.1.2 Current-Voltage Analysis

To characterize device performance, Quasi-static DC Current-Voltage (I-V) measurements are executed with a B1500A Semiconductor Device Parameter Analyzer with a B1517A High Resolution Source/Measure Unit (HRSMU). Devices are tested with bidirectional I-V sweep measurements with current compliance of 500 μA, unless otherwise stated. From I-V measurements, set voltage is defined as the voltage where the current first exceeds 300 μA. The low-resistance state (LRS) and high-resistance state (HRS) are defined as the current levels at 0.8 V before the device is set on the upward sweep and after the device is set on the downward sweep, respectively.

3.1.3 Pulse Measurements

Voltage pulses are used to measure endurance, retention, and analog conductance states. Data is collected by a custom data acquisition system and a DL 1211 current pre-amplifer from DL industries. Retention measurements are performed in vacuum at elevated temperatures with a LakeShore Model TTP4-1.5K Probe Station. For analog measurements, a stabilization stage before measurements consisted of repeating 30 sets of 200/400/1000 P-D pulses to partially stabilize Ag filaments.

3.2 Filament Dynamics in Epitaxial Synapses

Threading dislocations are preferential diffusion paths in crystalline solids [70]. When positive electrical bias is applied to the top electrode, Ag at the Ag/SiGe interface are oxidized to Ag^{+} cations and electrons[103]

$$Ag \rightarrow Ag^{+} + e^{-} \tag{3.1}$$

After oxidation, Ag^{+} ions drift into the SiGe film. Threading dislocations contain the majority of Ag^{+} ions, which are reduced to form Ag clusters extending along the defects that act as quasi-one-dimensional pathways through SiGe [103]

$$Ag^{+} + e^{-} \rightarrow Ag \tag{3.2}$$

Aligned Ag clusters through SiGe form electron conduction channels, also called conductive filaments. Low miscibility between Ag and SiGe and the absence of stable compounds[76, 74] localizes filament formation to threading dislocations rather than other interstitial or substitutional lattice sites. As shown in Fig. 3-2 a, conduction channel formation is suggested by hysteresis observed in I-V measurements.

In contrast, defect-free intrinsic Si epitaxially grown on p+ Si shows typical diode-like behavior without any hysteresis. Because there is no strain from lattice mismatch, threading dislocations are unlikely to form by half-loop nucleation through intrinsic Si. The absence of threading dislocations likely prevents formation of rupturable

metal conduction channels through the film.

Before conduction channels form, SiGe artificial synapses have low conductivity. Under high enough voltage, Ag oxidizes into ions, Ag^+ drifts into SiGe along dislocations, and ions are reduced within SiGe. This metal movement results in a net increase of the effective device conductivity. Formation of a conductive filament rapidly increases current through the device until the current compliance is reached. The current compliance limits current so that permanent shorting through the device (electrical breakdown) does not occur. As voltage is lowered, the higher current level indicates the conductivity of the device has increased due to the newly-formed conduction channel. Non-linear I-V characteristics after filament formation suggests that the conductance is governed by the effective barrier at the interface of the Ag filament and the p+ Si bottom electrode.

Oppositely-polarized voltage can retract the conductive filament and lower the device conductance, as shown in Fig. 3-2 b. As electrical current flows through the conduction channel, Joule heating occurs. More resistance occurs at narrowest filament regions, therefore these locations are likely to be where oxidation of Ag first occurs to initiate conduction channel rupture. Ag^+ ions diffuse away from the heated region and drift in the direction of the electric field. As a result, the current through the conductive filament is reduced. If Ag^+ ions are reduced again in the same location they are oxidized, the conductance state of the artificial synapse will remain the same. On the other hand, if reduction occurs in a different location, the conductance state will be different. When negative bias is applied to the Ag electrode, Ag+ ion drift will move towards the Ag electrode away from the p+ Si electrode. This reduces the effective Schottky barrier between SiGe and p+ Si. It is worth noting that conductance is believed to be exponentially-dependent on tunneling gap distance. Hence, further investigation of atomic configuration changes occurring during linear conductance change is of great interest.

The current through the device in the high resistance state (HRS) and low resistance state (LRS), as defined at 0.8 V, maintains temporal (cycle-to-cycle) uniformity, as shown in Fig. 3-2 c. This demonstrates that conductive filament formation

through SiGe is well-confined by threading dislocations. As displayed in Fig. 3-2 d, SiGe with threading dislocations (TDs) shows uniform resistive switching with only 1.7 % temporal set voltage variation (σ/μ) during 100 I-V cycles.

As shown in Fig. 3-2 e-f, other top electrode metals that form compounds with Si do not reset with up to 500 μA of reverse bias current. According to Cu-Si, Cu-Ge, Ni-Si, and Ni-Ge phase diagrams[3, 1], stable compounds with SiGe are energetically favorable. These conductive compounds are likely constituents of irreversible conductive pathways through the epitaxial film.

As a comparison, 60-nm-thick amorphous Si (a-Si) is grown by PECVD on p+ Si. I-V measurements for a-Si are shown in Fig. 3-3 a. As shown in Fig. 3-3 b, a-Si switching devices have large temporal set voltage variation (28%), which contrasts the uniform set voltage observed for single-crystalline SiGe in Fig. 3-2 d. Also, retention within a-Si is poor, as shown in Fig. 3-3 c. Retention for epitaxial SiGe synapses is discussed in Sec. 3.8.

3.3 Widening Threading Dislocations

Although epitaxial SiGe synapses are uniform, these devices have limited conductance range (on/off ratio). As observed in Fig. 3-2 b, the conductance change during reset is relatively small. This suggests that during reset, Ag^+ ions are prevented from diffusing away from the conduction channel to rupture the filament. Small range of conductance values limits the number of synaptic weights for training large-scale neuromorphic arrays [107]. Since this is suspected to be due to tight spatial accommodation [41, 97], widening TDs with defect-selective enchant[80] is predicted to allow for higher on/off ratio while maintaining confinement effects from the crystalline-SiGe lattice.

After epitaxial growth, defect selective etching with a mixture of 44 % 32 M Chromium trioxide (CrO_3) solution and 64 % hydrofluoric acid (HF) can be performed to widen threading dislocations. As shown in the SEM images in Fig. 3-4, threading dislocations (TDs) are widened since they are preferential reaction sites for oxidation in the presence of CrO_3 and subsequent oxide removal with HF. Consequences of

the etching process on I-V characteristics include 1) more effective reset at negative voltage bias, 2) higher on/off ratio by 3 orders of magnitude at positive voltage bias, and 3) lower set voltage by $\sim 0.7V$. Average set voltage and variation are plotted vs. etching time in Fig. 3-5 a.

Five seconds of etching results in sufficiently widened TDs for effective filament rupturing, as shown in Fig. 3-5 b. Fig. 3-5 c and d show I-V characteristics for etched and unetched SiGe in linear and semi-logerithmic scale, respectively. The current in the HRS state of a device with widened TDs is much lower than that of an unetched device. Widening TDs also increases LRS current and decreases set voltage, which both indicate etching helps to facilitate conduction channel formation. The negative bias I-V characteristics before forming a conduction channel and after rupturing a conduction channel are similar, as shown in Fig. 3-5, which further confirms effective Ag retraction. In addition to higher on/off ratio observed in I-V measurements ($\sim 10^4$), Widened TDs also exhibit a larger analog conductance range when subject to voltage pulses, as shown in Fig. 3-5 f. Analog performance with widened TDs is discussed further in Sec. 3.7.

3.4 Variations

Filament confinement in dislocations results in exceptionally low temporal variation while the uniform distribution of dislocations throughout the SiGe film allows for low spatial variation (measured for 500 devices). These low variations are essential for accurate pattern learning and recognition when implemented into neuromorphic hardware [10, 73, 105]. After an initial forming stage where Ag is first injected to SiGe (shown in Fig. 3-6 a), I-V cycling of SiGe is highly repeatable. As shown in Fig. 3-6 b, temporal set voltage variation (σ/μ) for over 700 switching cycles is as low as 1 %. This cycle-to-cycle uniformity makes a clear contrast to that of many amorphous-based device architectures even after modification to improve temporal/spatial uniformity by metal doping, field localization by nanoparticles, or confinement of cation transport by nanopore graphene[105, 62, 12, 104, 40].

In addition to temporal uniformity, epitaxial SiGe average set voltage is also spatially uniform since dislocations are well-distributed across the wafer (see Fig. 2-3). The average set voltage out of a hundred cycles was mapped for a hundred devices from two batches (Fig. 3-6 c). All measured devices show comparable average set voltage with spatial variation of only 4.9 % and uniform batch-to-batch performance (see Fig. 3-6 d).

3.5 Bandgap Engineering

Layer-by-layer controllability of films during epitaxial growth can be utilized for tuning device properties. For example, the Schottky barrier height between Ag and Si can be precisely controlled by specifying the doping concentration of the Si epilayer before SiGe epitaxy. As shown in Fig. 3-7 a-b, set voltage and read current can be modulated by varying the Schottky barrier height at Ag/Si interface. The ability to tune epitaxially-grown devices with Schottky barrier heights could allow optimization of recognition accuracy, power consumption, and prevention of sneak paths. For example, linear I-V is more robust to noise, while a non-linearity allows for much lower current at lower voltages, which minimizes sneak currents in crossbar arrays. In addition, the layer-by-layer growths of p-i-p back-to-back diodes in SiGe switching medium permits self-selection behavior as shown Fig. 3-7 c-d. This could be an effective route to further reduce sneak currents in neuromorphic arrays.

3.6 Nano-Scale Synapses

Scaling is an important consideration for portable electronics. Nano-scale epitaxial SiGe devices with active areas of $25nm \times 25nm$, $50nm \times 50nm$, $75nm \times 75nm$, $100nm \times 100nm$, and $125nm \times 125nm$ demonstrate similar I-V characteristics to $5\mu m \times 5\mu m$ devices, as shown in Fig. 3-8. This suggests conduction channel formation and rupture predominately occur at localized areas, where a limited number of dislocations are likely responsible for the majority of ionic movement among multiple TDs.

Nano-scale devices are able to operate with 100% yield due to the high threading dislocation density in epitaxial SiGe films ($\sim 10^{11} cm^{-2}$). As shown in Fig. 3-9 a, $25nm \times 25nm$ devices show good temporal uniformity. Spatial uniformity is also maintained, as shown in the histogram of set voltages for 100 devices plotted in Fig. 3-9 b. Slightly larger variation for nano-sized devices compared to micro-sized devices could be due to defects induced during RIE when defining the active region.

3.7 Analog Measurements

Analog artificial synapses must be capable of stabilizing at many different conductance states. During the training stage of supervised learning, weights are updated to minimize the error at the output. If there are too few accessible conductance levels, conductance values will not be able to precisely represent optimal weight values, which degrades the network accuracy. Analog on/off ratio is desired to be sufficiently high so that at least 64 distinct conductance states are accessible (6 bits)[107, 13].

Various conductance states can be digitally programmed by varying the current compliance under constant voltage bias. As shown in Fig. 3-10 a, the current compliance can be used to set the conductive state to different levels as plotted in Fig. 3-10 b. Without a transistor at each artificial synapse, passive arrays rely on programming by voltage pulses. The analog set threshold for an artificial synapses can be determined as the minimum voltage amplitude necessary to change the current level. When the pulsing scheme shown in Fig. 3-10 c is applied to an epitaxial SiGe synapse, this analog set threshold (V_{SET}) can be observed as the voltage that begins to change the readout current, as in Fig. 3-10 d. For reset, the pulsing scheme shown in Fig. 3-10 e can be applied to find the analog reset threshold (V_{RESET}), as shown in Fig. 3-10 f.

Analog artificial synapses can be characterized by measuring read currents in pulse trains consisting of repeated set-read and reset-read pulses. As shown in Fig. 3-11 a-b, the current during read pulses with amplitude below V_{SET} increases in response to set pulses (potentiation), and decreases in response to reset pulses (depression).

Potentation-depression (P-D) of SiGe artificial synapses exhibit analog switching with and without widening TDs, as shown in Fig. 3-12 a. Widening TDs allows for switching to occur with lower analog set and reset voltages. Also, the analog on/off ratio, measured as the current after potentiation divided by the current after depression, is larger for widened TDs since stronger filaments can be formed, and more complete filament rupture is possible.

Applying more P-D pulses to SiGe with widened TDs results in higher current levels and more condutance states. As shown in Fig. 3-12 b, applying 100, 200, and 500 P-D pulses results in analog on/off ratios of 100, 180, and 240, respectively.

As shown in the P-D plot in Fig. 3-12 b, remarkably high analog on/off ratio of 240 is measured for 1000 P-D pulses (500 potentiation/500 depression); the ratio decreases as the number of P-D pulses is reduced (180 for 400 P-D pulses and 100 for 200 P-D pulses).

While SiGe artificial synapses exhibit extremely high analog on/off ratio after 500 P-D pulses, conductance response upon P-D is non-linear. This is a typical characteristic of filamentary-type switching devices. Such non-linearity is more prominent when conductance saturates at its maximal value upon maximized potentiation pulses and abruptly decays upon depression. At maximum potentiation, filament conductivity can no longer increase in response to additional voltage pulses, which could be, in part, due to spatial limitations within widened TD channels.

Linearity can be quantified by a non-linearity magnitude assigned by the best-fit curve to the measured P-D read currents, as shown in Fig. 3-12 c. As shown in Fig. 3-12 d, increasing the number of pulses increases the magnitude of non-linearity. Linear conductance response with analog on/off conductance ratio 100 can be achieved using 100 P-D pulses, which is sufficient for training synaptic weights to accurately perform MNIST pattern recognition [13, 40]. The trade-off between linearity and the number of P-D pulses implies that widened TDs have limited spatial capacity to accommodate Ag. Linear conductance change can be obtained by avoiding saturated conductance limits.

In conventional CBRAM, the filament metal can diffuse into switching medium

and remain stuck in irreversible atomic configurations during repeated operation, which can limited endurance. However, epitaxial SiGe artificial synapses are capable of conductance update in response to more than 10^9 P-D pulses with stable current levels, as shown in Fig. 3-12 e. The immiscibility of Ag into SiGe likely contributes to allowing many repeated cycles.

3.8 Retention

In conventional CBRAM devices, metal filaments can pressurize an amorphous switching medium[5] and easily diffuse into amorphous phase, resulting in poor retention. Single-crystalline SiGe confines metal to predominantly occupy widened threading dislocations. However, without widening dislocations, a strongly-set filament eventually degrades and ruptures after removing electrical bias. As shown in Fig. 3-13 a, the current level at read voltage begins to decrease after about 1000 seconds. This is possibly caused by compressive stress on Ag filaments from the SiGe lattice, in addition to the tenancy for Ag to form stable clusters larger than TDs may spatially permit [101, 95].

Retention of epitaxial SiGe synapses is greatly improved by widening TDs with the optimal etching time. A strongly set Ag filament remained stable for over 48 hours at an elevated temperature of 85 ^oC, as shown in Fig. 3-13 b. However, over-etching TDs has a negative consequence on retention, as shown in Fig. 3-13 c. This could be due to motion if Ag in over-etched SiGe, which eventually results in filament rupture.

Assuming that diffusion is the main mechanism causing conduction channel rupture, the activation energy for Ag diffusion can be extracted from the Arrhenius plot of retention times at different elevated temperatures. Using the plot shown in Fig. 3-13 d for SiGe with widened dislocations, the activation energy of Ag is estimated to be 1.04 eV, which is similar to the value reported for Ag diffusion in single-crystalline Si with dislocations [22]. Extrapolating this plot to room temperature indicates that strongly-set conductive filaments can retain high conductivity for around 1.87 years at room temperature.

3.9 Artificial Neural Network Simulations

To demonstrate the suitability of SiGe artificial synapses for AI, an artificial neural network is simulated while taking into account measured device properties. The simulation is conducted on the basis of the platform "+NeuroSim." The source code is written with C++ programming language and is able to run on LINUX operation systems. A three-layer MLP neural network ($784 \times 300 \times 10$) is used with 28×28 MNIST images.

Training is iterated for one million patterns, randomly selected from the 60,000-image training set. Inference is performed using a 10,000-image testing set with non-ideal factors including finite on/off ratio, spatial/temporal variation, read noise, and wire resistance, and quantization of read currents.

The original patterns from the MNIST database are converted to black-and-white patterns with a threshold of 128 for pixel values ranging from 0 to 255. A logistic function, as described by Eqn. 1.2, is used as the activation function. The optimized learning rate for the first and second layer of the synapse is 0.4 and 0.2, respectively. The read voltage is 2 V and the read-out current is quantified to 8 bits by the analog-to-digital conversion circuit.

A behavioral model described by the following equations is used to capture non-linear conductance change [106]

$$G_{LTP} = B(1 - e^{\frac{-P}{A}}) + G_{min} \tag{3.3}$$

$$G_{LTD} = -B(1 - e^{\frac{P-P_{max}}{A}}) + G_{max} \tag{3.4}$$

$$B = \frac{G_{max} - G_{min}}{(1 - e^{\frac{-P}{A}})} \tag{3.5}$$

where P is the number of pulses, G_{LTP} and G_{LTD} are conductance states during LTP and LTD, respectively. G_{max}, G_{min}, and P_{max} represent the maximum conductance, minimum conductance, and the maximum pulse number required to switch the device

between the minimum and maximum conductance states, respectively. These values are extracted from the experimental data. A is determined by the nonlinearity of weight update and can be positive or negative. $A_{LTP} = 0.5032$, $A_{LTD} = -0.3868$ (normalized by P_{max}) is used to fit data shown in Fig. 3-11. B is a parameter determined by A within the range of G_{max}, G_{min}, and P_{max}.

The cycle-to-cycle variation describes the variation of the outcome conductance after applying each pulse. Assuming the conductance at each level obeys a normal distribution, the cycle-to-cycle variation is defined as the standard deviation divided by the maximum conductance[13, 27].

The device-to-device variation describes the variation of the parameter A. We assume that the fitting parameter A obeys a normal distribution. Device-to-device variation is defined as the standard deviation divided by the average value of A.

The precision is defined as the number of available conductance states during LTP and LTD.

Wire resistance between each crosspoint is 5Ω based on standard 14-nm CMOS technology. Read noise is chosen as 5%. The read-out current is quantified, normalized, and transferred to subsequent controlling logic circuits to calculate the delta weight.

The recognition accuracy is calculated every 8,000 images during each training process. Each data point for recognition accuracy is the average value of the last ten accuracy calculations.

The detail of 3-layer multi-layer perceptron (MLP) schematic is shown in Fig. 3-14 a. The inner product (summation) of input neuron signal vector and the first layer of synapse matrix is transferred after activation and binarization for input vectors of the next layer. Schematic in Fig. 3-14 b shows the circuit block diagram for a neuromorphic array composed of epitaxial SiGe synapses with the peripheral circuit.

3.9.1 Weight Update

The conductance update is implemented with half-voltage operation and the entire array is written line-by-line. The peripheral circuit and most of the neuron circuit

are verified by HSPICE simulations, and the delta weight calculation is performed by software.

In neuromorphic array hardware, conductance values, $\mathbf{G}$, are only positive (0 to 1), while synaptic weights in artificial neural networks, $\mathbf{W}$ are both positive and negative (-1 to 1). Hence, using a single artificial synapse for per synaptic weight requires a two-step read operation. First, vector-matrix multiplication, as described by Eqn. 1.1, is performed by the neuromorphic array when read pulses according to the input vector, $\vec{V}$, are applied. The outputs are doubled using a 1-bit left-shift, and the input vector V is subtracted to construct the weighted summation $\mathbf{W}\vec{V}$:

$$\mathbf{W}\vec{V} = 2\mathbf{G}\vec{V} - \mathbf{J}\vec{V} \tag{3.6}$$

where $\mathbf{J}$ is a matrix of all ones with the same dimensionality as $\mathbf{W}$ and $\mathbf{G}$. The MSB (positive or negative sign bit) of the adder output is the 1-bit output of the low-precision activation function. This output from the first neuromorphic array is stored for calculating delta-weight and used as the input to the hidden layer. The MSB output of the second array is the prediction, which is also is used for calculating delta-weight. A schematic for this process is illustrated in Fig. 3-14 c.

Using backpropagation, the amount to change each synapse, or the delta-weight Δw, is the product of the learning rate, η, the array inputs x, and the partial derivatives of the error calculated at the outputs δ. The derivative of the logistic function (Eqn. 1.3) can be expressed as

$$f'(x) = f(x) \cdot (1 - f(x)) \tag{3.7}$$

Hence, the partial derivatives of the error for k output neurons from the second array can be determined as

$$\delta_k = (y_{target} - y_k)y_k(1 - y_k) \tag{3.8}$$

where y_{target} are the target outputs and y_k are the actual outputs. The delta-weight for the second array between j input neurons and k output neurons is calculated as

$$\Delta w_{jk} = \eta x_2 \delta_k \tag{3.9}$$

where x_2 are the outputs of first array and inputs to the second array. Similarly, the partial derivatives of the error for j output neurons of the first array are calculated as

$$\delta_j = (\sum_k w_{jk} \cdot \delta_k) x_2 (1 - x_2) \tag{3.10}$$

and the delta-weight for the first array between i input neurons and j output neurons is calculated as

$$\Delta w_{ij} = \eta x_1 \delta_j \tag{3.11}$$

where x_1 are the inputs to the first array. Calculated delta-weights determine the number of write or erase pulses to be applied for each synapse.

To change the conductance of artificial synapses in a neuromorphic array, potentiation and depression require different voltage pulse schemes. As shown in Fig. 3-14 d, half-voltage bias is used to protect unselected devices. During potentiation, the selected word line (WL) is held at the write voltage, V_w, while 0 V write pules are applied to bit lines (BL) to increase the conductance of selected synapses. During depression, the selected WL is held at the erase voltage, V_e, while 0 V erase pulses are applied to BL to decrease conductance.

3.10 Handwriting Recognition

Based on measured characteristics of epitaxial SiGe synapses, an artificial neural network is simulated to perform supervised learning with the MNIST handwritten recognition dataset[55]. A three-layer neural network with 28×28 pre-neurons, 300 hidden neurons, and 10 output neurons is utilized[15]. The multilayer perception (MLP) algorithm with stochastic gradient descent weight update is used. Non-ideal factors such as finite on-off ratio, finite number of conductance levels, device-to-device

variation, cycle-to-cycle variation, wire resistance, and read noise are accounted for. The 784 neurons of the input layer correspond to 28×28 MNIST image, and 10 neurons of the output layer correspond to 10 classes of digits (0-9)[49]. The impact of various device parameters on recognition accuracy considered for our simulation and specific values for epiRAM are displayed in Fig. 3-15 a-e.

After training with one million patterns randomly selected from 60,000 images from training set, recognition accuracy is tested with a separated 10,000 images from the testing set[15, 14]. Employing simple circuitry to compensate for the nonlinear conductance change with pulse number[49], the simulation suggests that the neural networks formed with epitaxial SiGe synapses can achieve 95.1% on average (96.5% as maximum) recognition accuracy, which is comparable to the accuracy of 97% obtained by the software baseline using binary input signal for the first two signal layers, as shown in Fig. 3-15 g[49, 71] Using gray-scale instead of binary input, this accuracy of this algorithm using software with similar network size is 98% [49, 55]

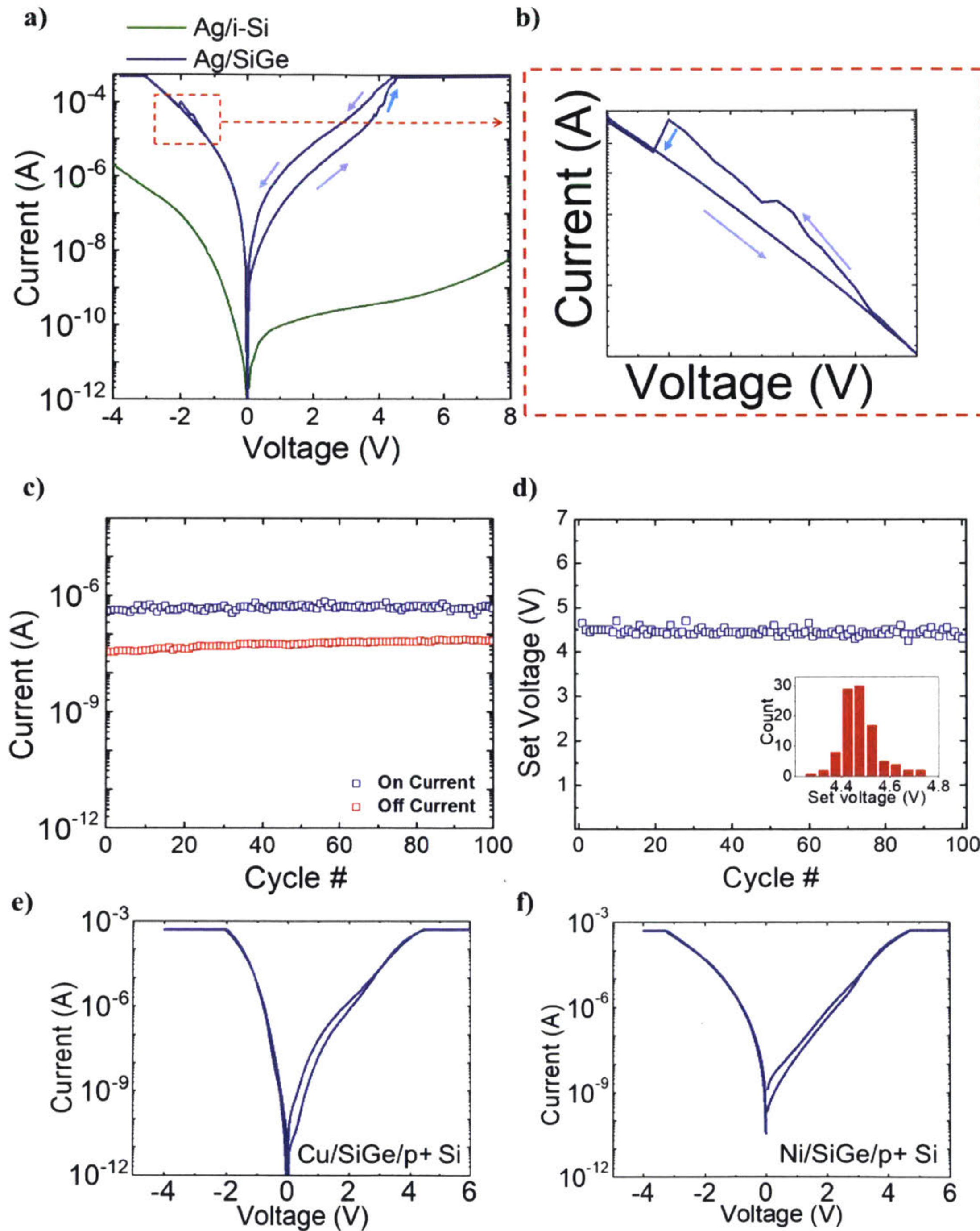


Figure 3-2: Resistive switching in SiGe. a) I-V measurements of epitaxial SiGe synapses (with unetched dislocations) and of Ag/crystalline i-Si/crystalline p-Si device, where no hysteresis is observed. The signature of filament rupture is highlighted in b). c) HRS and LRS states measured at 0.8 V for 100 cycles. d) Set voltage measured over 100 quasi-static I-V sweeps. Inset: histogram for set voltage distribution. e) Cu/SiGe/p+ Si and f) Ni/SiGe/p+ Si devices show very little hysteresis and high current, likely due to formation of stable conductive compounds.

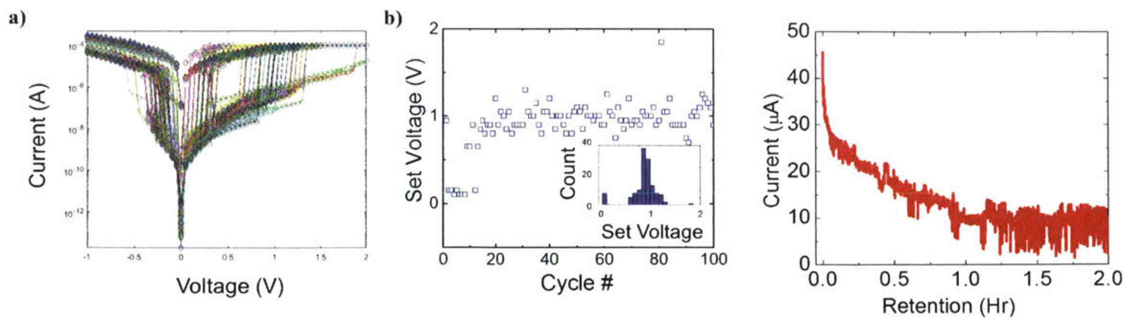


Figure 3-3: Ag/Amorphous-Si(a-Si)/p+ Si switching devices. a) I-V measurements for a-Si devices. b) Set voltage measured over 100 quasi-static I-V sweeps. Inset: histogram for set voltage distribution. Set threshold: 44 μA. c) Retention for a-Si device.

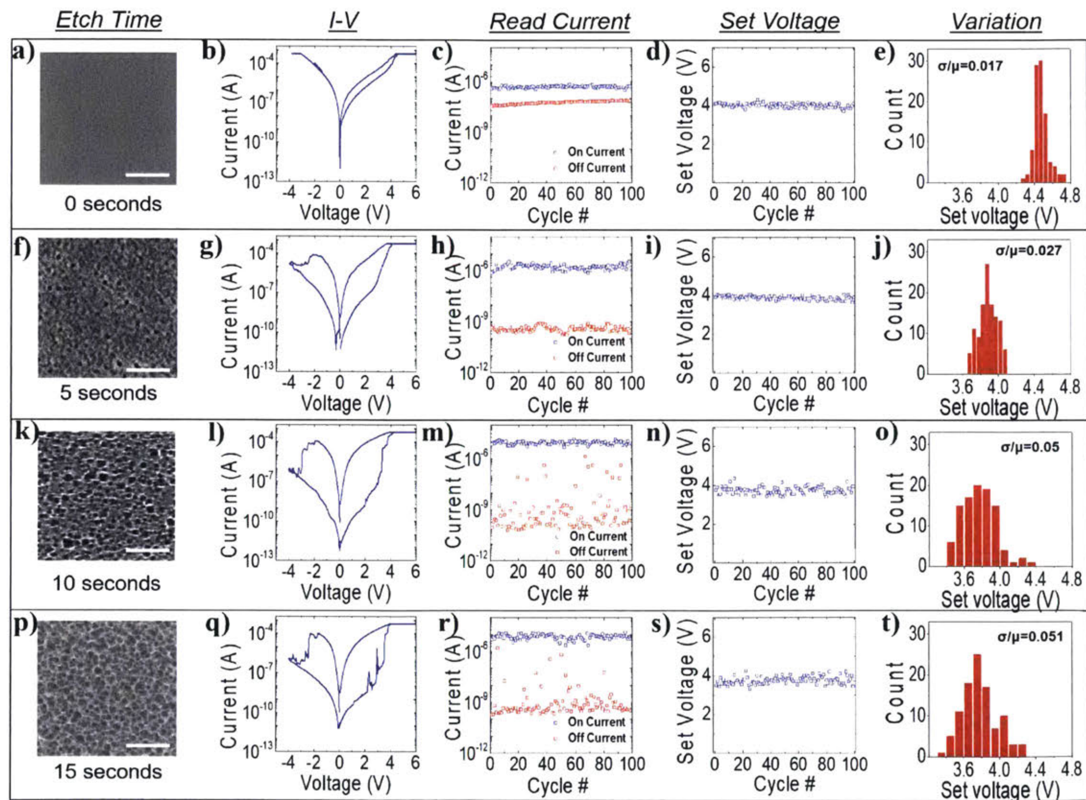


Figure 3-4: Effect of etching on epitaxial SiGe synapses. a,f,k,p) Plan-view SEM images reveal increasing etching time widens dislocations. Scale bar: 200 nm. b,g,l,q) I-V measurements for different etch times. c,h,m,r) HRS and LRS temporal evolution shows etching increases the on/off ratio, but over-etching increases variation of the HRS state. d,i,n,s) Set voltage temporal evolution shows increasing variation with etch time. e,j,o,t) Histograms of set voltage for different etch times.

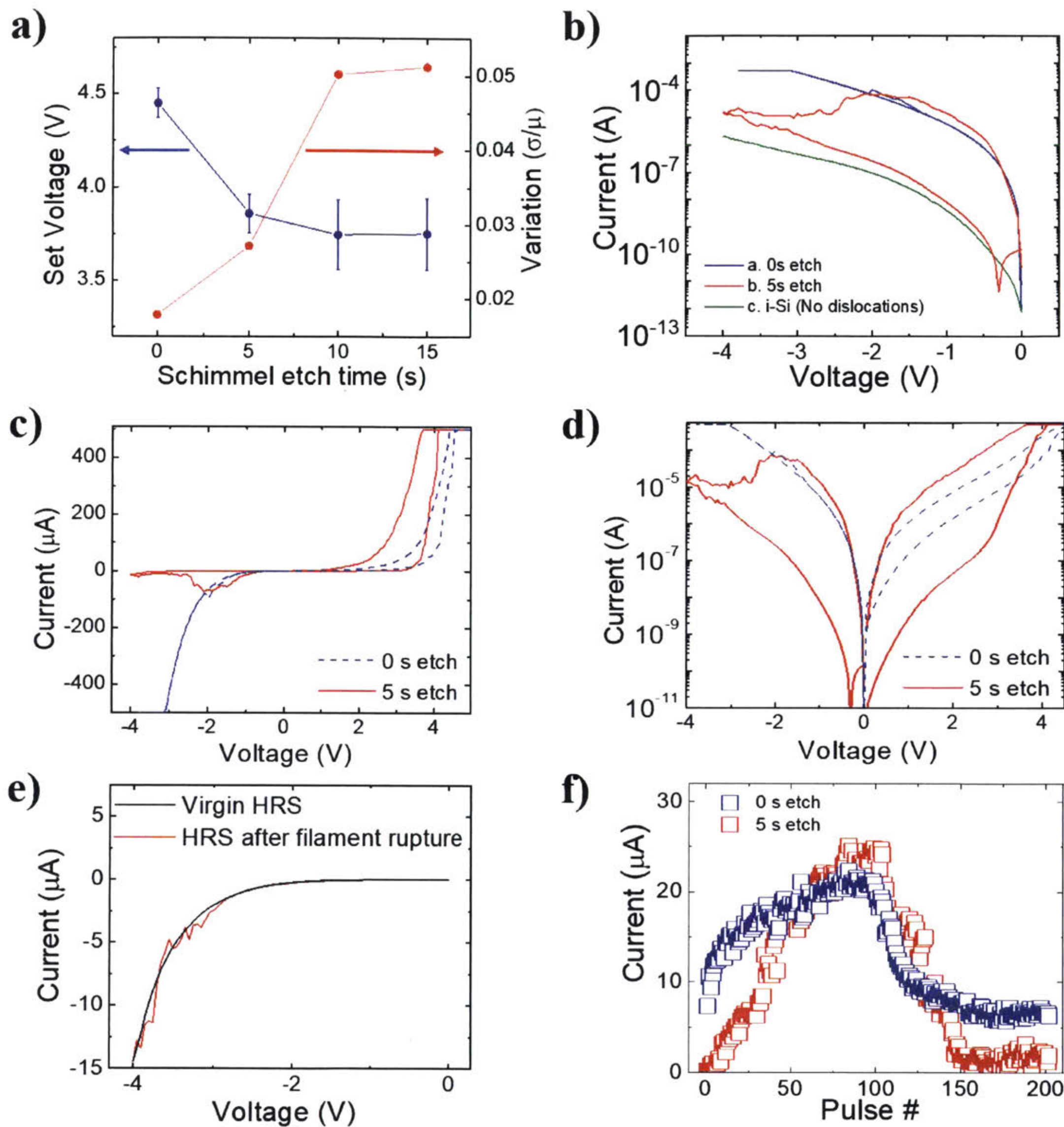


Figure 3-5: Complete Ag filament rupture is enabled by defect-selective etching. a) Set voltage and variation plotted vs. etching time shows that as etch time increases, set voltage decreases and variation increases. b) Semi-logarithmic I-V characteristics of the reset process without etching (0s) and after 5s of etching plotted with i-Si shows larger decrease in current is enabled by widening TDs. c) Semi-logarithmic DC I-V characteristics without etching (0s) and after 5s of etching. d) Linear-scale DC I-V characteristics without etching (0s) and after 5s of etching. e) I-V characteristics of SiGe in the virgin state and after Ag filament rupture, showing nearly-complete reset to the HRS is allowed with widened TDs. f) Higher analog on/off ratio is observed for SiGe with widened TDs.

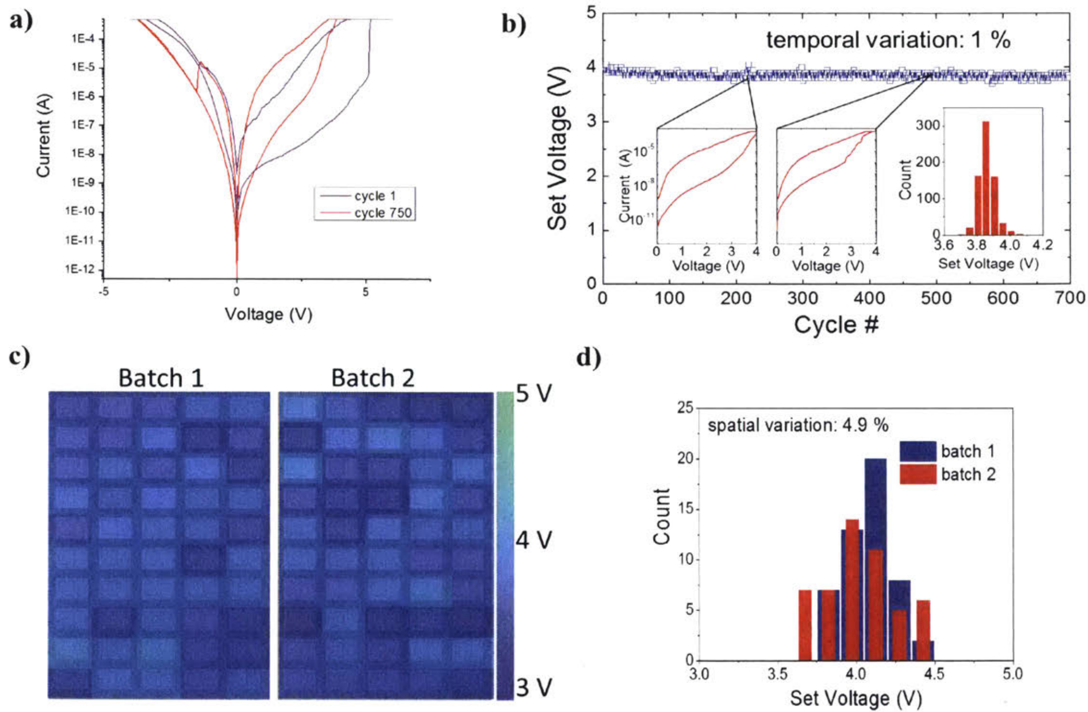


Figure 3-6: Characteristics of SiGe artificial synapses after widening TDs with defect-selective etching. a) I-V measurements of the 1st and 750th cycle showing devices initially have higher set voltage, lower current level, and less effective reset. b) The temporal evolution of set voltage for a device for over 700 quasi-static I-V sweeps. Insets: DC I-V plots at 250th cycle and 500th cycle, along with the histogram for set voltage distribution. c) Map of set voltages at devices overlapped with optical images of devices (50 devices from batch 1 and 50 devices from batch 2 are shown). d) Histogram for spatial set voltage distribution.

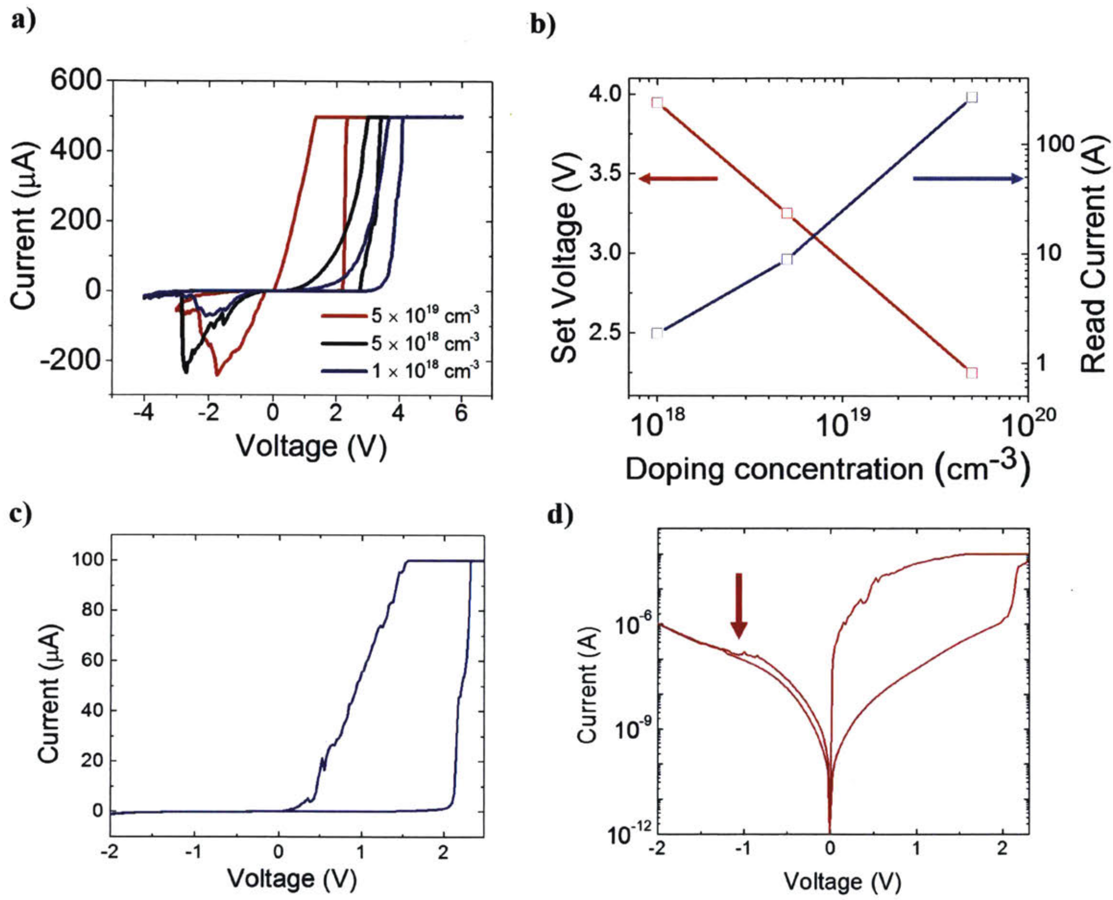


Figure 3-7: Bandgap engineering to tune characteristics of epitaxial SiGe synapses. a) I-V measurements with different doping concentrations of p+ Si below SiGe. b) Set voltage and read current (at 0.8 V) plotted as a function of doping concentration suggests that higher doping concentration results in lower set voltage and higher read current due to the lower Schottky barrier. c) Linear-scale and d) Logarithmic-scale I-V curve of p-i-p SiGe back-to-back epitaxial SiGe synapse. LRS-state current is rectified at negative bias.

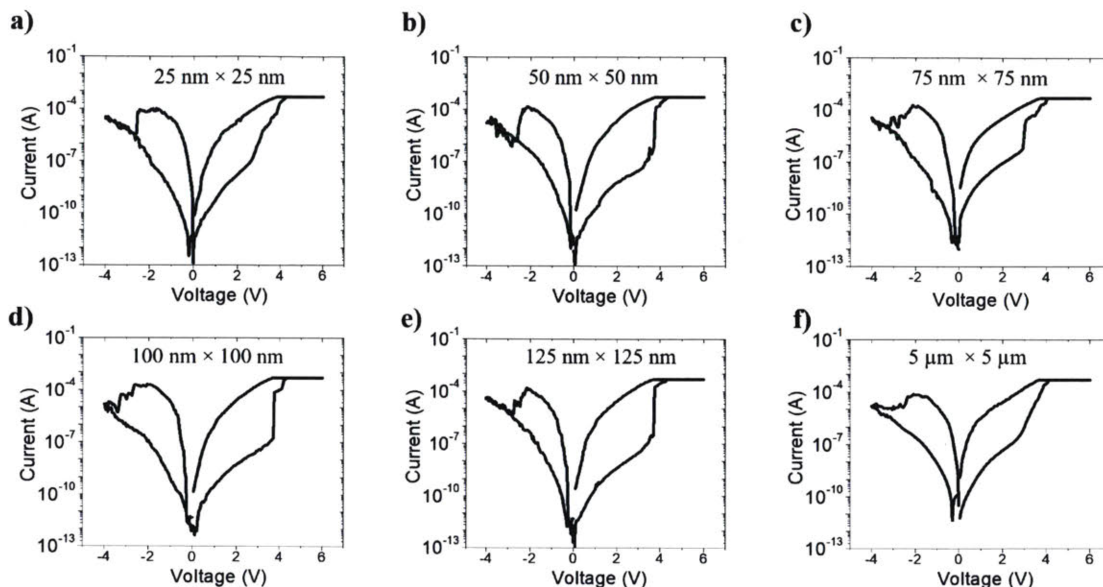


Figure 3-8: I-V measurements of nano-scale epitaxial SiGe devices. a) $25nm \times 25nm$, b) $50nm \times 50nm$, c) $75nm \times 75nm$, d) $100nm \times 100nm$, and e) $125nm \times 125nm$ devices have similar hysteresis to f) $5\mu m \times 5\mu m$ devices.

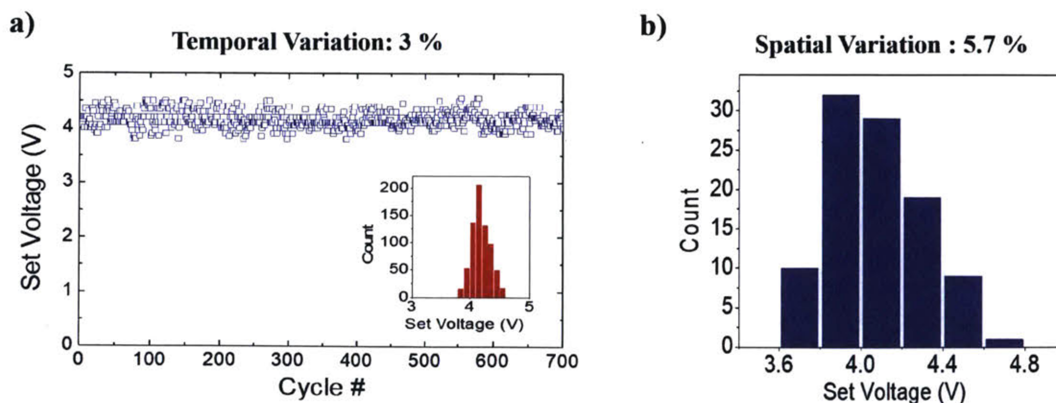


Figure 3-9: Variation of nano-scale epitaxial SiGe synapses. a) Temporal variation of set voltage of a $25nm \times 25nm$ device cycled over 700 times. b) Histogram showing the spatial variation of set voltage for 100 $25nm \times 25nm$ devices.

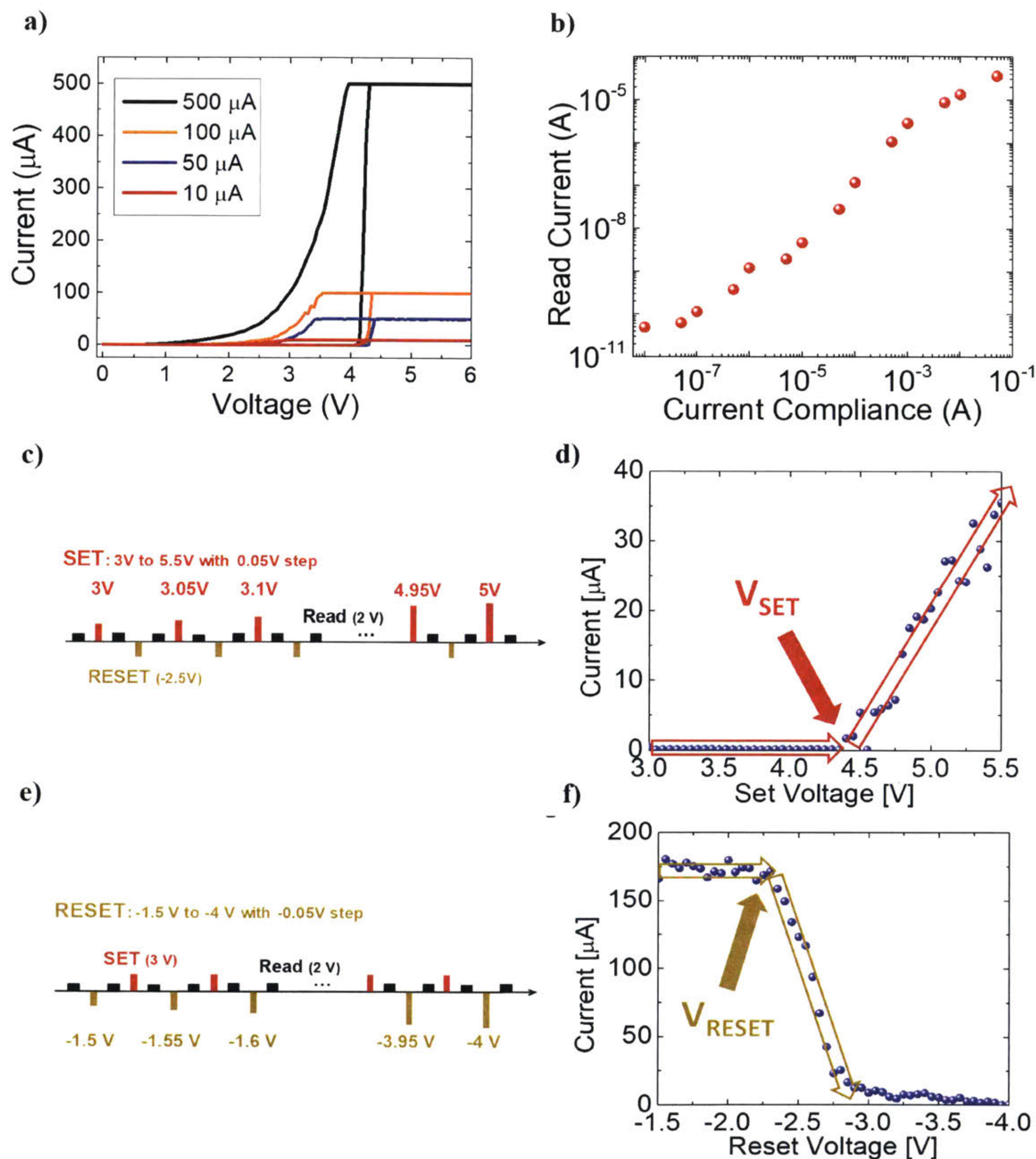


Figure 3-10: Digital and analog access of different conductance levels. a) I-V measurements with different current compliance can set LRS to different levels. b) Read current of LRS increases with current compliance since a more conductive (stronger) channel can be formed. c) Pulsing scheme consisting of gradually increasing set pulses spaced with read pulses allows d) analog set voltage to be determined. e) Pulsing scheme consisting of gradually decreasing reset pulses allows f) analog reset voltage to be determined.

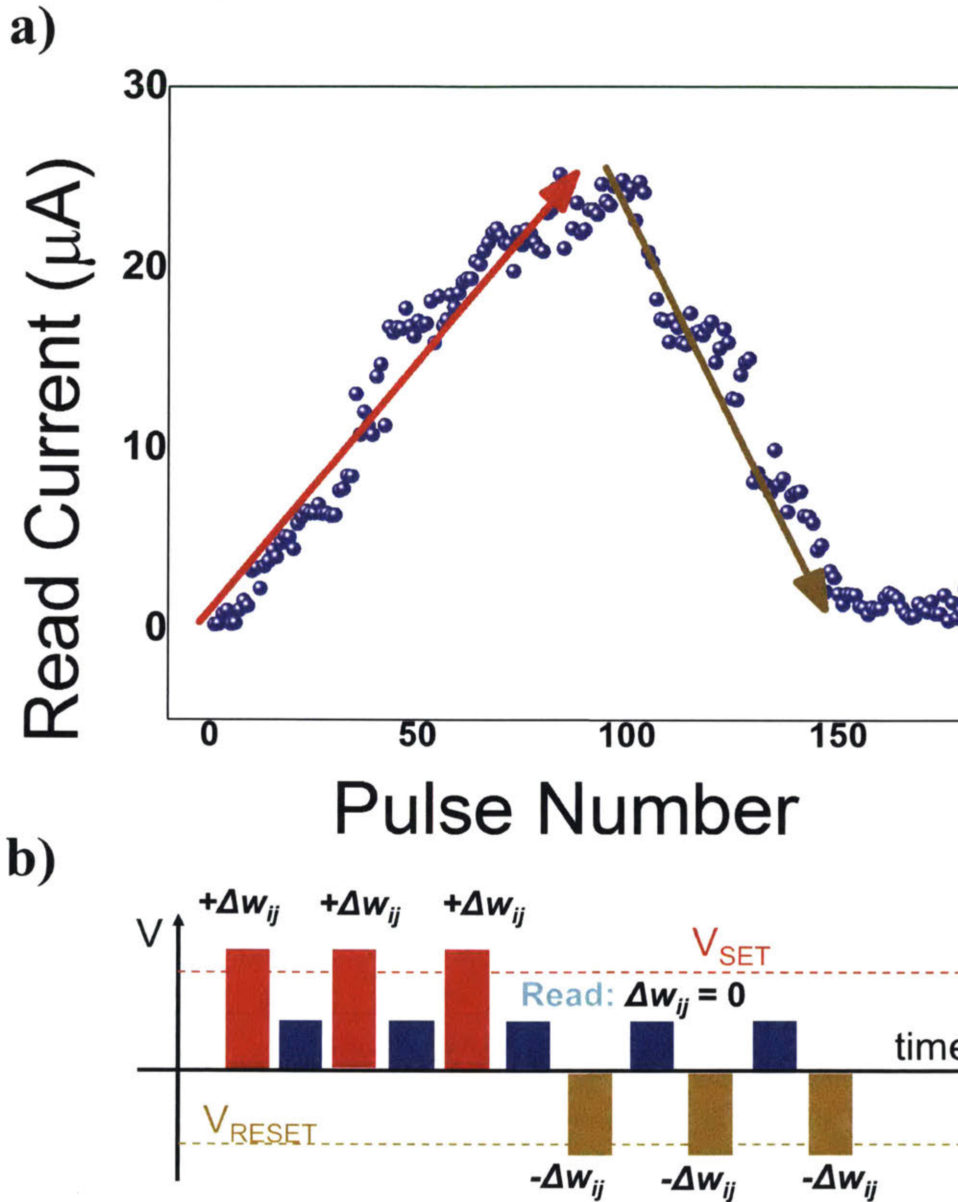
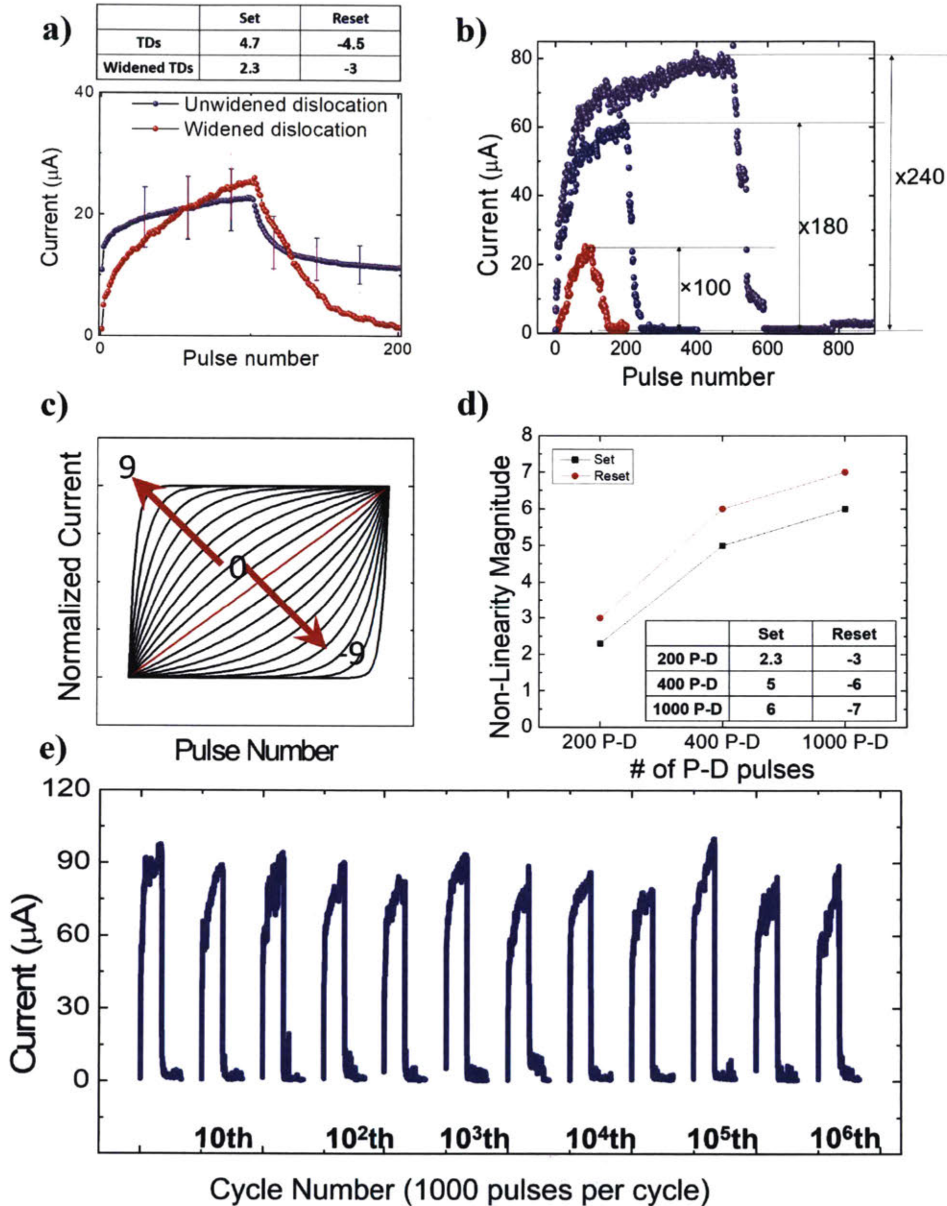


Figure 3-11: Potentiation-depression (P-D) of epitaxial SiGe synapses. a) Read current in response to 100 set pulses and 100 reset pulses. b) Pulsing scheme used to for P-D measurements. Voltage pulses with amplitude above V_{SET} increase the weight value $(+w_{ij})$ and amplitude below V_{RESET} decrease the weight value $(-w_{ij})$, while weaker pulses that do not change the conductance $(\delta w_{ij} = 0)$ are applied to obtain the read current.



	Set	Reset
TDs	4.7	-4.5
Widened TDs	2.3	-3

	Set	Reset
200 P-D	2.3	-3
400 P-D	5	-6
1000 P-D	6	-7

Figure 3-12: Analog characteristics of epitaxial SiGe synapses. a) Potentiation-depression (P-D) with and without widenened TDs showing widened TDs promote larger analog on/off ratio. Table: Analog set and reset voltages. b) P-D shows analog on/off ratio increase with number of applied pulses. The pulse train consists of 100/200/500 consecutive set pulses (5 V, 5 μs) followed by 100/200/500 consecutive reset pulses (-3 V, 5 μs), respectively. Current is measured by read pulses (2 V, 1 ms) after each set/reset pulse. c) The definition of non-linearity magnitude d) The non-linear magnitude depending of number of P-D pulses. Inset: table showing non-linearity magnitudes for 200, 400, and 1000 P-D pulses. e) Endurance measured for 10^9 set/reset pulses (10^6 cycles). The first and fifth of each order of magnitude is shown. Each P-D cycle consists of 500 consecutive potentiation pulses and 500 consecutive depression pulses with read pulses between each update. After 10^6 cycles, similar P-D can still be observed.

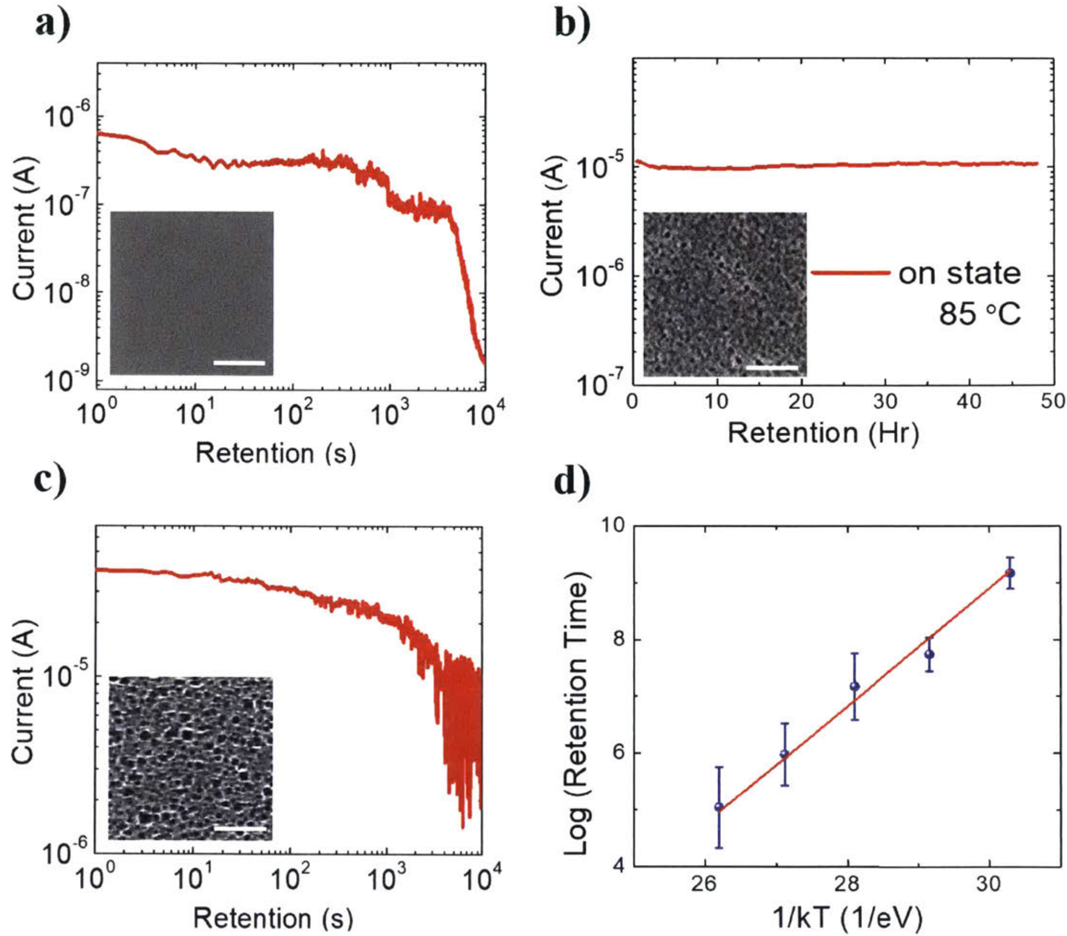


Figure 3-13: Retention of SiGe artificial synapses. a) Untreated TDs show poor retention. Inset: plan-view SEM of relatively flat SiGe. b) The two-day retention test at 85°C at LRS for a SiGe artificial synapse etched for 5s. The device performance remains unchanged after the test. Inset: plan-view SEM of 5s etched SiGe. c) Over-etched TDs also show poor retention. Inset: plan-view SEM of 10s etched SiGe. a-c are measured with 1.5 V read pulses with pulse width of 1 ms. d) Retention tests for 5s etched SiGe at elevated temperatures. The extrapolation of the plot to room temperature indicates 1.87 years of retention. Measurements are collected twice from 10 devices at each temperature between 398K to 443K.

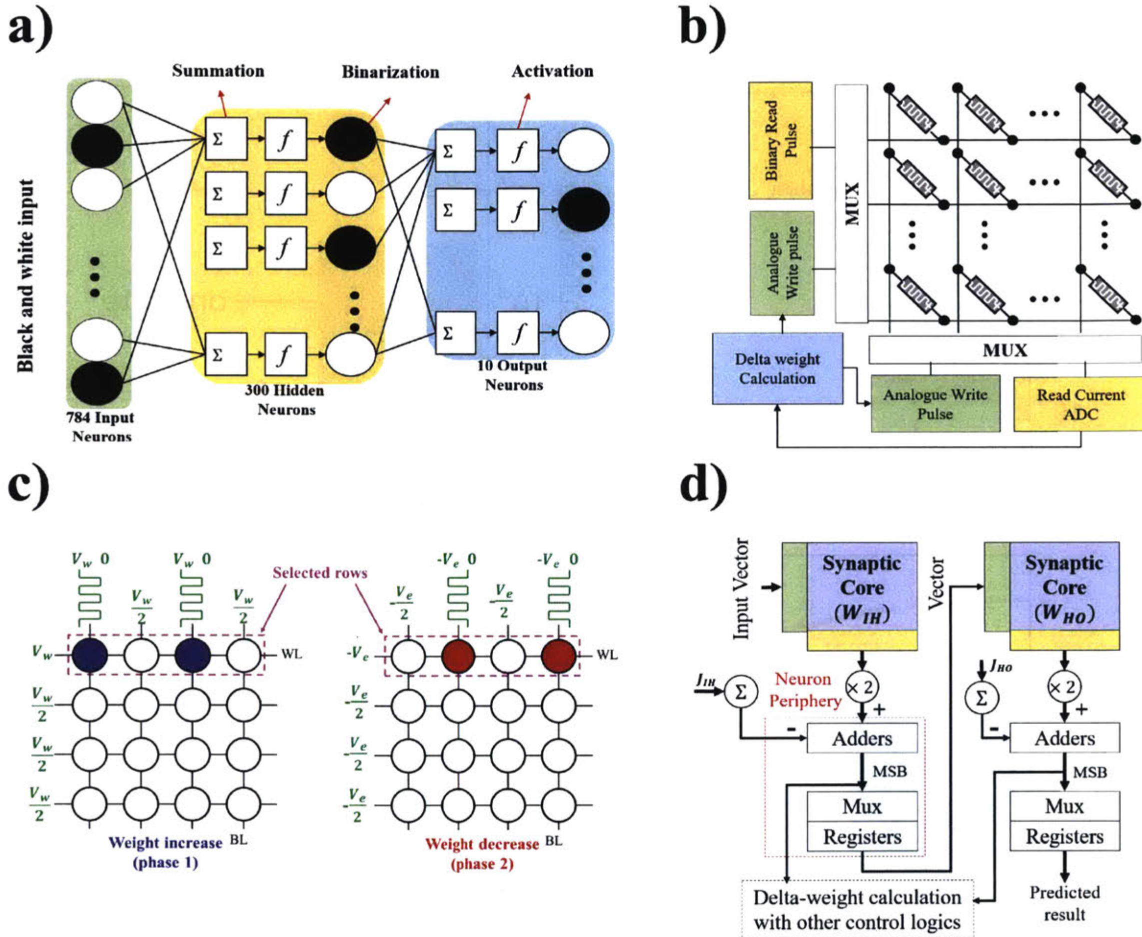


Figure 3-14: Schematic for the image recognition simulation. a) A three-layer MLP neural network with a black-and-white input signal for each layer in the algorithm level. The inner product (summation) of the input neuron signal vector and the first synapse array vector is transferred after activation and binarization as the input vector for the second synapse array. b) Circuit block diagram of a neuromorphic crossbar array and the peripheral circuits. MUX, multiplexer; ADC, analog-to-digital converter. c) Diagram illustrating the process for determining outputs for prediction and delta-weight calculation. d) Schematic of potentiation (weight increase) and depression (weight decrease) phases during training. The number of write or erase pulses for each synapse is determined by the delta-weight calculation. Figures are adapted from [15]

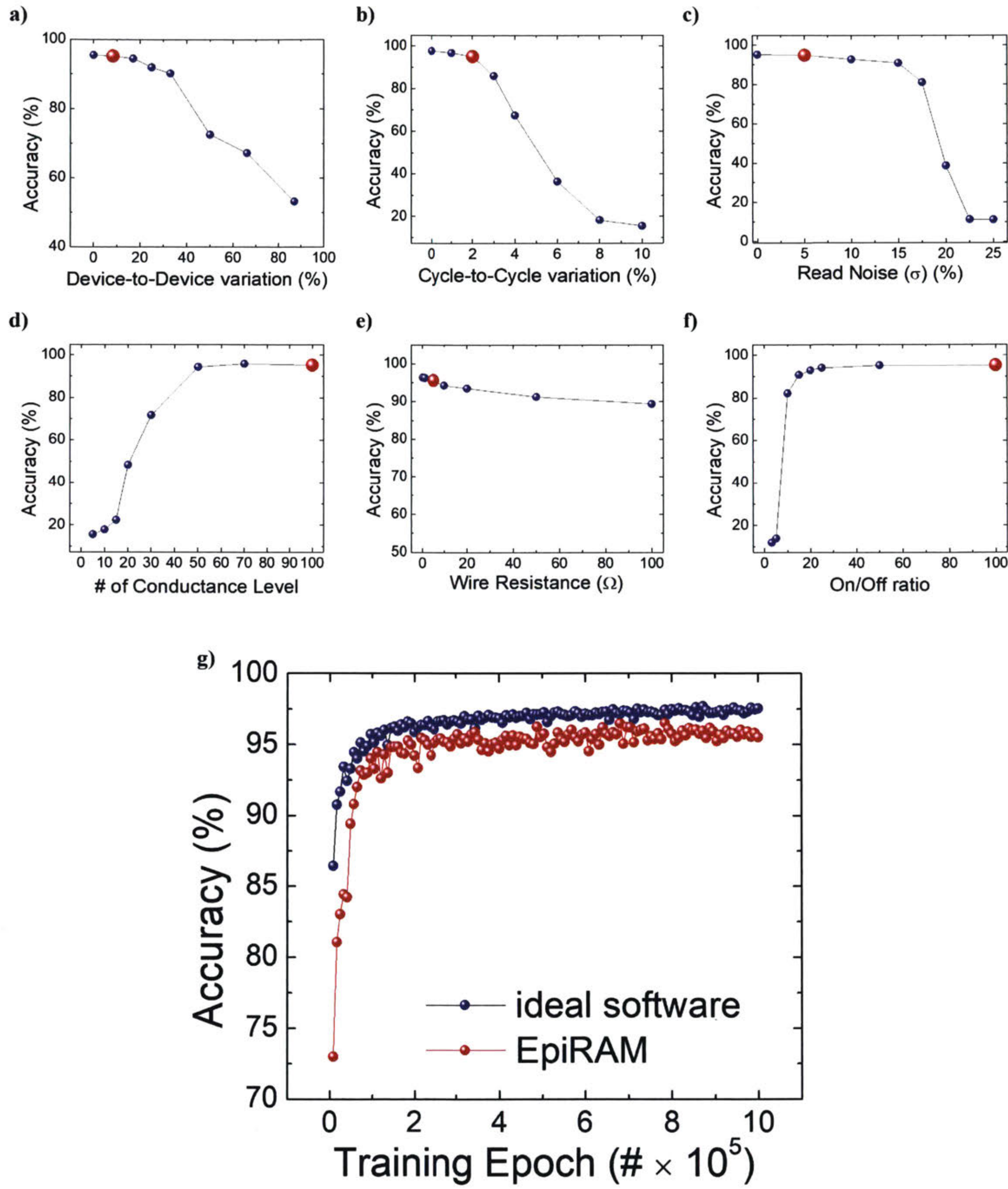


Figure 3-15: Influence of non-ideal device parameters for MLP simulation. The influence of a) Device-to-Device variation, b) cycle-to-cycle variation, c) Read noise, and d) number of conductance levels, e) wire resistance, and f) on/off ratio on the final MNIST digit recognition accuracy. Measured values for epitaxial SiGe synapses are shown in red. e) Evolution of accuracy to training epochs for ideal software and for neuromorphic arrays with epitaxial SiGe synapses.

Chapter 4

Conclusions

Several new computing system designs combine memory cells and transistors for efficient synaptic weight training and storage compared to conventional hardware. Neuromorphic arrays with two-terminal conductive bridging devices are promising, yet they typically rely on formation of filaments in an amorphous medium, which is stochastic and unreliable. Spatial and temporal variation of conductance response has therefore limited conventional devices to small-scale demonstrations.

For this thesis, I worked on epitaxial SiGe artificial synapses that display unprecedented uniformity and demonstrated all characteristics that suggest suitability for large-scale arrays. Metastable SiGe films containing many threading dislocations are grown by low-pressure chemical vapor deposition, and threading dislocations are widened using defect-selective etching. Widened threading dislocations in the single-crystalline SiGe switching layer can confine Ag filaments in *quasi*-one-dimensional channels. This confinement results in enhanced set voltage uniformity, long retention, high endurance, and high analog on/off ratio. Simulations using the MNIST dataset prove that epitaxial SiGe synapses could operate with online learning accuracy of up to 95.1%.

Future research at devices-level, circuits-level, and systems-level could help to realize large-scale passive neuromorphic arrays. At the devices-level, it remains a significant challenge to consistently achieve linear conductance change in response to identical voltage pulses. *In-situ* characterization during weight update could help

to reveal underlying mechanisms of filament formation and rupture guided by defect pipelines. Also, better retention is desired to maintain weight values at elevated temperatures. There is also incentive to reduce voltage threshold amplitudes to minimize power consumption during training. At the circuits-level, bandgap engineering can be used to optimize crosspoint architecture for large-scale operation in parallel. For example, sneak currents could be further minimized by integrating selectors. At the systems level, a significant challenge is to develop peripheral circuitry and tuning algorithms for operation of many array rows and columns simultaneously. Real demonstration of MNIST handwriting recognition using 784 x 300 and 300 x 10 neuromorphic arrays with performance advantages over conventional computing systems has yet to be realized. Epitaxial SiGe synapses have demonstrated uniformity, high on/off ratio, good endurance, stable retention, linear conductance update, and suppression of sneak currents. Hence, the development of epitaxial SiGe synapses is a step towards creating new computing hardware for AI to transcend language barriers, teach contextualized information, and enhance quality of life for all people.

Bibliography

[1] BINARY (SGTE) Alloy Phase Diagrams.

[2] Deep Learning Inference Accelerators | NVIDIA Tesla|NVIDIA.

[3] Ahmed A. Al-Joubori and C. Suryanarayana. Synthesis of metastable NiGe2 by mechanical alloying. *Materials & Design*, 87:520–526, 12 2015.

[4] Fabien Alibart, Elham Zamanidoost, and Dmitri B Strukov. Pattern classification by memristive crossbar circuits using ex situ and in situ training. *Nature communications*, 4(May):2072, 1 2013.

[5] Stefano Ambrogio, Simone Balatti, Seol Choi, and Daniele Ielmini. Impact of the Mechanical Stress on Switching Characteristics of Electrochemical Resistive Memory. *Advanced Materials*, 26(23):3885–3892, 6 2014.

[6] E L Bienenstock, L N Cooper, and P W Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 2(1):32–48, 1 1982.

[7] Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science (New York, N.Y.)*, page eaao1733, 12 2017.

[8] Yoeri van de Burgt, Ewout Lubberman, Elliot J. Fuller, Scott T. Keene, GrÃĺgorio C. Faria, Sapan Agarwal, Matthew J. Marinella, A. Alec Talin, and Alberto Salleo. A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing. *Nature Materials*, (February):1–6, 2017.

[9] G. W. Burr, P. Narayanan, R. M. Shelby, S. Sidler, I. Boybat, C. di Nolfo, and Y. Leblebici. Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: Comparative performance analysis (accuracy, speed, and power). In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 1–4. IEEE, 12 2015.

[10] G.W. Burr, R.M. Shelby, C. di Nolfo, J.W. Jang, R.S. Shenoy, P. Narayanan, K. Virwani, E.U. Giacometti, B. Kurdi, and H. Hwang. Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses),

using phase-change memory as the synaptic weight element. In *2014 IEEE International Electron Devices Meeting*, pages 1–29. IEEE, 12 2014.

[11] Srimat Chakradhar, Murugan Sankaradas, Venkata Jakkula, Srihari Cadambi, Srimat Chakradhar, Murugan Sankaradas, Venkata Jakkula, and Srihari Cadambi. A dynamically configurable coprocessor for convolutional neural networks. In *Proceedings of the 37th annual international symposium on Computer architecture - ISCA '10*, volume 38, page 247, New York, New York, USA, 2010. ACM Press.

[12] W.Y. Chang, C.A. Lin, J.H. He, and T.B. Wu. Resistive switching behaviors of ZnO nanorod layers. *Applied Physics Letters*, 96(24):242109, 6 2010.

[13] Pai-Yu Chen, Ligang Gao, and Shimeng Yu. Design of Resistive Synaptic Array for Implementing On-Chip Sparse Learning. *IEEE Transactions on Multi-Scale Computing Systems*, 2(4):257–264, 10 2016.

[14] P.Y. Chen, X. Peng, and S. Yu. User Manual of MLP Simulator (+NeuroSim).

[15] P.Y. Chen, X. Peng, and S. Yu. NeuroSim+: An integrated device-to-device algorithm framework for benchmarking synaptic devices and array architectures. In *IEEE International Electron Devices Meeting (IEDM)*, San Francisco, USA, 2017.

[16] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, Olivier Temam, Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, Olivier Temam, Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning. *ACM SIGPLAN Notices*, 49(4):269–284, 2014.

[17] Yu-Hsin Chen, Tushar Krishna, Joel S. Emer, and Vivienne Sze. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE Journal of Solid-State Circuits*, 52(1):127–138, 1 2017.

[18] Yunji Chen, Tao Luo, Shaoli Liu, Shijin Zhang, Liqiang He, Jia Wang, Ling Li, Tianshi Chen, Zhiwei Xu, Ninghui Sun, and Olivier Temam. DaDianNao: A Machine-Learning Supercomputer.

[19] Zidong Du, Robert Fasthuber, Tianshi Chen, Paolo Ienne, Ling Li, Tao Luo, Xiaobing Feng, Yunji Chen, Olivier Temam, Zidong Du, Robert Fasthuber, Tianshi Chen, Paolo Ienne, Ling Li, Tao Luo, Xiaobing Feng, Yunji Chen, and Olivier Temam. ShiDianNao. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture - ISCA '15*, volume 43, pages 92–104, New York, New York, USA, 2015. ACM Press.

[20] D. J. Eaglesham, E. P. Kvam, D. M. Maher, C. J. Humphreys, and J. C. Bean. Dislocation nucleation near the critical thickness in GeSi/Si strained layers. *Philosophical Magazine A*, 59(5):1059–1073, 5 1989.

[21] Clement Farabet, Cyril Poulet, Jefferson Y. Han, and Yann LeCun. CNP: An FPGA-based processor for Convolutional Networks. In *2009 International Conference on Field Programmable Logic and Applications*, pages 32–37. IEEE, 8 2009.

[22] D J Fisher, T C Nason, G R Yang, K H Park, and T M Lu. Diffusion in Silicon. *Journal of Applied Physics*, 70(3):1392–6, 1991.

[23] E. A. Fitzgerald, G. P. Watson, R. E. Proano, D. G. Ast, P. D. Kirchner, G. D. Pettit, and J. M. Woodall. Nucleation mechanisms and the elimination of misfit dislocations at mismatched interfaces by reduction in growth area. *Journal of Applied Physics*, 65(6):2220–2237, 3 1989.

[24] Kunihiko Fukushima. Artificial vision by multi-layered neural networks: Neocognitron and its advances. *Neural Networks*, 37:103–119, 1 2013.

[25] Kunihiko Fukushima. Training multi-layered neural network neocognitron. *Neural Networks*, 40:18–31, 4 2013.

[26] Huajian Gao and William D. Nix. SURFACE ROUGHENING OF HETEROEPITAXIAL THIN FILMS. *Annual Review of Materials Science*, 29(1):173–209, 8 1999.

[27] Ligang Gao, I-Ting Wang, Pai-Yu Chen, Sarma Vrudhula, Jae-sun Seo, Yu Cao, Tuo-Hung Hou, and Shimeng Yu. Fully parallel write/read in resistive synaptic array for accelerating on-chip learning. *Nanotechnology*, 26(45):455204, 2015.

[28] Vinayak Gokhale, Jonghoon Jin, Aysegul Dundar, Berin Martini, and Eugenio Culurciello. A 240 G-ops/s Mobile Coprocessor for Deep Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 696–701. IEEE, 6 2014.

[29] Tayfun Gokmen, O Murat Onen, and Wilfried Haensch. Training Deep Convolutional Neural Networks with Resistive Cross-Point Devices.

[30] Tayfun Gokmen and Yurii Vlasov. Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations. *Frontiers in neuroscience*, 10:333, 2016.

[31] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech Recognition with Deep Recurrent Neural Networks. 3 2013.

[32] Xiaoxiao Guo, Satinder Singh, Honglak Lee, Richard L. Lewis, and Xiaoshi Wang. Deep Learning for Real-Time Atari Game Play Using Offline Monte-Carlo Tree Search Planning, 2014.

[33] D. Hammerstrom. A VLSI architecture for high-performance, low-cost, on-chip learning. In *1990 IJCNN International Joint Conference on Neural Networks*, pages 537–544. IEEE, 1990.

[34] D. O. (Donald Olding) Hebb. *The organization of behavior : a neuropsychological theory.* L. Erlbaum Associates, 2002.

[35] G E Hinton, P Dayan, B J Frey, and R M Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science (New York, N.Y.)*, 268(5214):1158–61, 5 1995.

[36] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition.

[37] Geoffrey E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, 8 2002.

[38] D. C. Houghton. Strain relaxation kinetics in Si1-xGex/Si heterostructures. *Journal of Applied Physics*, 70(4):2136–2151, 1991.

[39] D. C. Houghton, D. D. Perovic, J. M. Baribeau, G. C. Weatherly, J. Dieleman, G. D. Pettit, and J. M. Woodall. Misfit strain relaxation in GexSi1-x/Si heterostructures: The structural stability of buried strained layers and strained-layer superlattices. *Journal of Applied Physics*, 67(4):1850–1862, 2 1990.

[40] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized Neural Networks Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations. 2016.

[41] Robert Hull. *Properties of crystalline silicon.* Institution of Electrical Engineers, 2006.

[42] Robert Hull and John C. (John Condon) Bean. *Germanium silicon : physics and materials.* Academic Press, 1999.

[43] Daniele Ielmini and Rainer Waser. *Resistive switching : from fundamentals of nanoionic redox processes to memristive device applications.* 2016.

[44] Uma Jain, S. C. Jain, A. H. Harker, and R. Bullough. Nucleation of dislocation loops in strained epitaxial layers. *Journal of Applied Physics*, 77(1):103–109, 1 1995.

[45] Sung Hyun Jo, Ting Chang, Idongesit Ebong, Bhavitavya B. Bhadviya, Pinaki Mazumder, and Wei Lu. Nanoscale Memristor Device as Synapse in Neuromorphic Systems. *Nano Letters*, 10(4):1297–1301, 4 2010.

[46] Sung Hyun Jo, Kuk Hwan Kim, and Wei Lu. High-density crossbar arrays based on a Si memristive system. *Nano Letters*, 9(2):870–874, 2009.

[47] Sung Hyun Jo and Wei Lu. CMOS Compatible Nanoscale Nonvolatile Resistance Switching Memory. *Nano Letters*, 8(2):392–397, 2 2008.

[48] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon. In-Datacenter Performance Analysis of a Tensor Processing Unit. 4 2017.

[49] Irina Kataeva, Farnood Merrikh-Bayat, Elham Zamanidoost, and Dmitri Strukov. Efficient training algorithms for neural networks based on memristive crossbar circuits. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 7 2015.

[50] Kuk Hwan Kim, Siddharth Gaba, Dana Wheeler, Jose M. Cruz-Albrecht, Tahir Hussain, Narayan Srinivasa, and Wei Lu. A Functional Hybrid Memristor Crossbar-Array/CMOS System for Data Storage and Neuromorphic Applications. *Nano Letters*, 12(1):389–395, 1 2012.

[51] Sungho Kim, Chao Du, Patrick Sheridan, Wen Ma, ShinHyun Choi, and Wei D. Lu. Experimental Demonstration of a Second-Order Memristor and Its Ability to Biorealistically Implement Synaptic Plasticity. *Nano Letters*, 15(3):2203–2211, 3 2015.

[52] Karthik Krishnan, Tohru Tsuruoka, Cedric Mannequin, and Masakazu Aono. Mechanism for Conducting Filament Growth in Self-Assembled Polymer Thin Films for Redox-Based Atomic Switches. *Advanced Materials*, 28(4):640–648, 1 2016.

[53] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 1:1097–1105, 2012.

[54] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, 12 1989.

[55] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[56] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 5 2015.

[57] Myoung-Jae Lee, Chang Bum Lee, Dongsoo Lee, Seung Ryul Lee, Man Chang, Ji Hyun Hur, Young-Bae Kim, Chang-Jung Kim, and David H Seo. A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O(5-x)/TaO(2-x) bilayer structures. *Nature Materials*, 10(8):625–30, 2011.

[58] C. Lehmann, M. Viredaz, and F. Blayo. A generic systolic array building block for neural networks with on-chip learning. *IEEE Transactions on Neural Networks*, 4(3):400–407, 5 1993.

[59] Boxun Li, Yuzhi Wang, Yu Wang, Yiran Chen, and Huazhong Yang. Training itself: Mixed-signal training acceleration for memristor-based neural network. In *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 361–366. IEEE, 1 2014.

[60] Sicheng Li, Chunpeng Wu, Hai Li, Boxun Li, Yu Wang, and Qinru Qiu. FPGA Acceleration of Recurrent Neural Network Based Language Model. In *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines*, pages 111–118. IEEE, 5 2015.

[61] Seppo Linnainmaa. Taylor expansion of the accumulated rounding error. *BIT*, 16(2):146–160, 6 1976.

[62] Qi Liu, Shibing Long, Wei Wang, Qingyun Zuo, Sen Zhang, Junning Chen, and Ming Liu. Improvement of Resistive Switching Properties inZrO2-Based ReRAM With Implanted Ti Ions. *IEEE Electron Device Letters*, 30(12):1335–1337, 2009.

[63] Xiaoxiao Liu, Mengjie Mao, Beiye Liu, Boxun Li, Yu Wang, Hao Jiang, Mark Barnell, Qing Wu, Jianhua Yang, Hai Li, and Yiran Chen. Harmonica: A Framework of Heterogeneous Computing Systems With Memristor-Based Neuromorphic Computing Accelerators. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 63(5):617–628, 5 2016.

[64] Xiaoxiao Liu, Mengjie Mao, Beiye Liu, Hai Li, Yiran Chen, Boxun Li, Yu Wang, Hao Jiang, Mark Barnell, Qing Wu, and Jianhua Yang. RENO: A High-efficient Reconfigurable Neuromorphic Computing Accelerator Design *.

[65] J. W. Matthews, S. Mader, and T. B. Light. Accommodation of Misfit Across the Interface Between Crystals of Semiconducting Elements or Compounds. *Journal of Applied Physics*, 41(9):3800–3804, 8 1970.

[66] J.W. Matthews and A.E. Blakeslee. Defects in epitaxial multilayers: I. Misfit dislocations. *Journal of Crystal Growth*, 27:118–125, 12 1974.

[67] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2 2015.

[68] Gyeong Su Park, Young Bae Kim, Seong Yong Park, Xiang Shu Li, Sung Heo, Myoung Jae Lee, Man Chang, Ji Hwan Kwon, M. Kim, U In Chung, Regina Dittmann, Rainer Waser, and Kinam Kim. In situ observation of filamentary conducting channels in an asymmetric Ta2O5-x/TaO2-x bilayer structure. *Nature Communications*, 4:2382, 9 2013.

[69] Paolo Politi, GeneviÃÍve Grenet, Alain Marty, Anne Ponchet, and Jacques Villain. Instabilities in crystal growth by atomic or molecular beams. *Physics Reports*, 324(5-6):271–404, 2 2000.

[70] David A. Porter, K. E. Easterling, and Mohamed Y. Sherif. *Phase transformations in metals and alloys.* CRC Press, 2009.

[71] M. Prezioso, I. Kataeva, F. Merrikh-Bayat, B. Hoskins, G. Adam, T. Sota, K. Likharev, and D. Strukov. Modeling and implementation of firing-rate neuromorphic-network classifiers with bilayer Pt/Al2O3/TiO2âĹŠx/Pt Memristors. In *Technical Digest - International Electron Devices Meeting, IEDM*, pages 1–17. IEEE, 12 2016.

[72] 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*, 521(7550):61–64, 5 2015.

[73] 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*, 521(7550):61–4, 2015.

[74] A Prince. Ag-Ge-Si Ternary Phase Diagram Evaluation, 1988.

[75] U. Ramacher, J. Beichter, W. Raab, J. Anlauf, N. Brüls, U. Hachmann, and M. Wesseling. Design of a 1st Generation Neurocomputer. In *VLSI Design of Neural Networks*, pages 271–310. Springer US, Boston, MA, 1991.

[76] F Rollert, N A Stolwijk, and H Mehrer. Solubility, diffusion and thermodynamic properties of silver in silicon. *Journal of Physics D: Applied Physics*, 20(9):1148, 1987.

[77] F Rosenblatt. THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN. *Psychological Review*, 65(6):19–8.

[78] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 10 1986.

[79] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized Experience Replay. 11 2015.

[80] D. G. Schimmel. Defect Etch for Silicon Evaluation. *Journal of The Electrochemical Society*, 126(3):479, 1979.

[81] JÃijrgen Schmidhuber. Deep Learning in Neural Networks: An Overview. 2014.

[82] Jae-sun Seo, Binbin Lin, Minkyu Kim, Pai-Yu Chen, Deepak Kadetotad, Zihan Xu, Abinash Mohanty, Sarma Vrudhula, Shimeng Yu, Jieping Ye, and Yu Cao. On-Chip Sparse Learning Acceleration With CMOS and Resistive Synaptic Devices. *IEEE Transactions on Nanotechnology*, 14(6):969–979, 11 2015.

[83] Keisuke Shibuya, Regina Dittmann, Shaobo Mi, and Rainer Waser. Impact of Defect Distribution on Resistive Switching Characteristics of Sr ₂ TiO ₄ Thin Films. *Advanced Materials*, 22(3):411–414, 1 2010.

[84] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 10 2017.

[85] Karen Simonyan and Andrew Zisserman. VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION. 2015.

[86] Daniel Soudry, Dotan Di Castro, Asaf Gal, Avinoam Kolodny, and Shahar Kvatinsky. Memristor-Based Multilayer Neural Networks With Online Gradient Descent Training. *IEEE Transactions on Neural Networks and Learning Systems*, 26(10):2408–2421, 10 2015.

[87] J.S. Speck, M.A. Brewer, G. Beltz, A.E. Romanov, and W. Pompe. Scaling laws for the reduction of threading dislocation densities in homogeneous buffer layers A model of threading dislocation density in strain-relaxed Ge Scaling laws for the reduction of threading dislocation densities in homogeneous buffer layers. *Journal of Applied Physics Applied Physics Letters Journal of Applied Physics Applied Physics Letters Applied Physics Letters Journal of Applied Physics Journal of Applied Physics*, 801(10):102115–2293, 1996.

[88] Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models. 7 2015.

[89] K. Steinbuch. Die Lernmatrix. *Kybernetik*, 1(1):36–45, 1 1961.

[90] Dmitri B. Strukov, Gregory S. Snider, Duncan R. Stewart, and R. Stanley Williams. The missing memristor found. *Nature*, 453(7191):80–83, 5 2008.

[91] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel Emer. Efficient Processing of Deep Neural Networks: A Tutorial and Survey.

[92] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. 9 2014.

[93] Krzysztof Szot, Wolfgang Speier, Gustav Bihlmayer, and Rainer Waser. Switching the electrical resistance of individual dislocations in single-crystalline SrTiO3. *Nature materials*, 5(4):312–320, 2006.

[94] L. Vegard. Die Konstitution der Mischkristalle und die Raumfï£¡llung der Atome. *Zeitschrift fï£¡r Physik*, 5(1):17–26, 1 1921.

[95] Zhongrui Wang, Saumil Joshi, Sergey E. SavelâĂŹev, Hao Jiang, Rivu Midya, Peng Lin, Miao Hu, Ning Ge, John Paul Strachan, Zhiyong Li, Qing Wu, Mark Barnell, Geng-Lin Li, Huolin L. Xin, R. Stanley Williams, Qiangfei Xia, and J. Joshua Yang. Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nature Materials*, 16(September), 2016.

[96] Rainer Waser, Regina Dittmann, Georgi Staikov, and Kristof Szot. Redox-Based Resistive Switching Memories - Nanoionic Mechanisms, Prospects, and Challenges. *Advanced Materials*, 21(25-26):2632–2663, 7 2009.

[97] A F Wells. Structural inorganic chemistry. *Nature*, 229(5285):453, 1971.

[98] Chao Weng, Dong Yu, Shinji Watanabe, and Biing-Hwang Fred Juang. Recurrent deep neural networks for robust speech recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5532–5536. IEEE, 5 2014.

[99] Zihan Xu, Abinash Mohanty, Pai-Yu Chen, Deepak Kadetotad, Binbin Lin, Jieping Ye, Sarma Vrudhula, Shimeng Yu, Jae-sun Seo, and Yu Cao. Parallel Programming of Resistive Cross-point Array for Synaptic Plasticity. *Procedia Computer Science*, 41:126–133, 1 2014.

[100] J Joshua Yang, Dmitri B Strukov, and Duncan R Stewart. Memristive devices for computing. *Nature nanotechnology*, 8(1):13–24, 2013.

[101] Yuchao Yang, Peng Gao, Siddharth Gaba, Ting Chang, Xiaoqing Pan, and Wei Lu. Observation of conducting filament growth in nanoscale resistive memories. *Nature Communications*, 3:732, 2012.

[102] Yuchao Yang, Peng Gao, Linze Li, Xiaoqing Pan, Stefan Tappertzhofen, Shin-Hyun Choi, Rainer Waser, Ilia Valov, and Wei D. Lu. Electrochemical dynamics of nanoscale metallic inclusions in dielectrics. *Nature Communications*, 5:377–383, 6 2014.

[103] Yuchao Yang and Wei Lu. Nanoscale resistive switching devices: mechanisms and modeling. *Nanoscale*, 5(21):10076–92, 2013.

[104] Jung Ho Yoon, Jeong Hwan Han, Ji Sim Jung, Woojin Jeon, Gun Hwan Kim, Seul Ji Song, Jun Yeong Seok, Kyung Jean Yoon, Min Hwan Lee, and Cheol Seong Hwang. Highly Improved Uniformity in the Resistive Switching Parameters of TiO ₂ Thin Films by Inserting Ru Nanodots. *Advanced Materials*, 25(14):1987–1992, 4 2013.

[105] Byoung Kuk You, Myunghwan Byun, Seungjun Kim, and Keon Jae Lee. Self-Structured Conductive Filament Nanoheater for Chalcogenide Phase Transition. *ACS Nano*, 9(6):6587–6594, 6 2015.

[106] Shimeng Yu and Pai-Yu Chen. Emerging Memory Technologies: Recent Trends and Prospects. *IEEE Solid-State Circuits Magazine*, 8(2):43–56, 2016.

[107] Shimeng Yu, Pai-Yu Chen, Yu Cao, Lixue Xia, Yu Wang, and Huaqiang Wu. Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect. In *2015 IEEE International Electron Devices Meeting (IEDM)*, pages 1–17. IEEE, 12 2015.

[108] Shimeng Yu, Yi Wu, Rakesh Jeyasingh, Duygu Kuzum, and H.-S. Philip Wong. An Electronic Synapse Device Based on Metal Oxide Resistive Switching Memory for Neuromorphic Computation. *IEEE Transactions on Electron Devices*, 58(8):2729–2737, 8 2011.

[109] Chen Zhang, Peng Li, Guangyu Sun, Yijin Guan, Bingjun Xiao, and Jason Cong. Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks. In *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays - FPGA '15*, pages 161–170, New York, New York, USA, 2015. ACM Press.