UCLA
UCLA Electronic Theses and Dissertations

Title

A silicon-based self-programming synaptic resistor network for neuromorphic computing

Permalink

https://escholarship.org/uc/item/5kv1w5p5

Author

Nathan, Dhruva Sean

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

A silicon-based self-programming synaptic resistor network for neuromorphic computing

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Mechanical Engineering

By

Dhruva Sean Nathan

2022

© Copyright by

Dhruva Sean Nathan

2022

ABSTRACT OF THE DISSERTATION

A silicon-based self-programming synaptic resistor network for neuromorphic computing

by

Dhruva Sean Nathan

Doctor of Philosophy in Mechanical Engineering

University of California, Los Angeles, 2022

Professor Yong Chen, Chair

Compared to modern supercomputers, which consume roughly 10^6 W of power, the human brain requires only 20 W to function, and still exceeds the performance of supercomputers in many creative tasks. This stark difference in energy requirements is caused by a fundamental difference in computing architecture. Modern computers follow the Von Neumann architecture, in which transistors dedicated to logic and memory functions are physically separated, and the time and energy required to communicate between the two units constitutes a bottleneck which impedes performance in machine learning and optimization problems. On the other hand, in the human brain, logic, memory, and learning functions are integrated together in a single element:

the synapse. Without the Von Neumann bottleneck, the brain can achieve fast real-time learning, adaptability in complex environments, and massive parallelism. For the future of neuromorphic computing, it is important to develop an electronic device which mimics the synaptic function, so that large-scale circuits which mimic the neurobiological architecture can be developed. This work reports a silicon-based synaptic resistor (referred to as "synstor" hereinafter) which integrates logic, learning, and memory in a single device. The synstor is composed of a semiconducting silicon channel connected via Schottky contacts to titanium input and output electrodes, a thermal silicon dioxide, an aluminum oxide switching layer, and a tantalum oxide reference electrode. The large defect density in the switching layer attracts or repels charge carriers in the silicon channel to modify its conductance, and the defect density in turn can be modified by voltage pulses applied on its input and output electrodes (pre- and post-synaptic spikes). Synaptic resistor circuits could be scaled-up to facilitate mobile artificial intelligence systems with brain-like intelligence and adaptability in complex environments.

The dissertation of Dhruva Sean Nathan is approved.

Pei-Yu Chiou

Qibing Pei

Jason L. Speyer

Yong Chen, Committee Chair

University of California, Los Angeles

2022

Acknowledgements

This dissertation was made possible with the support and guidance of my advisors, colleagues, friends, and family.

I would like to thank my advisor, Professor Yong Chen, for his earnest support and guidance over many research projects, and for his mentorship in my own personal growth.

To my committee members, Professor Jason L. Speyer, Professor Pei-Yu Chiou, and Professor Qibing Pei, for their faith and guidance towards this effort.

To my present group members: Rahul Shenoy, Zixuan Rong, Dawei Gao, Jungmin Lee, and Atharva Deo, for their ongoing encouragement, support, and research collaboration.

To my former group members: Dr. Christopher Shaffer, Dr. Cameron Danesh, and Dr. Andrew Tudor, for generously sharing their expertise and for guidance during my early career.

To my parents, who have ardently and selflessly supported me throughout my graduate program.

To the support given by the Air Force Office of Scientific Research (AFOSR) under the program, “Intelligent Neuromorphic Network” (contract number: FA9550-15-1-0056) and “Avian-Inspired Multifunctional Morphing Vehicles” (contract number: FA9550-16-1-0087).

Table of Contents

List of Figures

Vita

2015	B.Asc Nanotechnology Engineering University of Waterloo
2018	M.S. Mechanical Engineering University of California, Los Angeles
2018-2022	PhD. Candidate and Graduate Student Researcher Mechanical & Aerospace Engineering Department University of California, Los Angeles

1. Introduction

Moore's law has overseen an exponential growth in the performance and energy efficiency of of transistor circuits, and has empowered supercomputers (e.g. Fugaku) to execute algorithms with computational speeds (~10^{17} floating-point operations per second, OPS) far higher than the human brain (~10^{16} OPS), leading to artificial intelligence (AI) systems that can outperform humans in specific tasks such as arithmetic computation, games[1], image/speech recognitions[2], and self-driving cars)[3-4]. Unfortunately, the supercomputers (Fugaku) consume much more power (~$3\times10^{7}\ W$) than the human brain (~$20\ W$) and have an energy-efficiency (~$10^{10}\ OPS/W$) five orders of magnitude inferior to the human brain (~$10^{15}\ O$PS/W) ([5]).

In modern computers, logic and memory transistors are physically separated, and operated in serial based on the Turing Model[6], as shown in Figure 1. The signal transmission between constitutes a large time and energy cost, referred to as the "Von Neumann bottleneck." Memory access itself consumes 100-1000 times more energy than a CPU operation[7]. Despite the improved computational architecture, connectivity, parallelism, data transmission, and energy efficiency, transistor-based computing circuits, such as the Fugaku supercomputer[8], graphics processing units (GPUs)[9], tensor processing units (TPUs), field-programmable gate arrays (FPGAs)[10], TrueNorth[11], Loihi[12], and Tianjic[13] are still subject to this limitation.

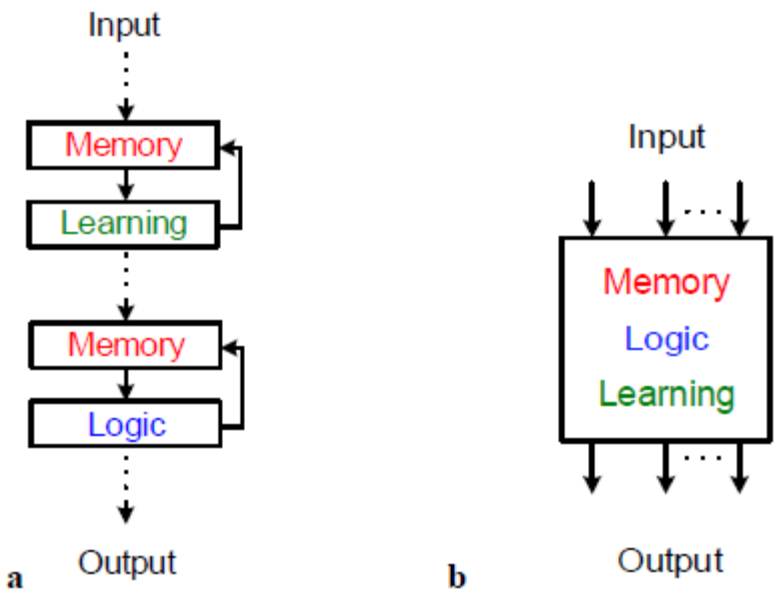

Figure 1: (Left) Based on the Turing model, a computer executes inference and learning algorithms on separated logic and memory units in serial mode with data transitions between them. (Right) A parallel computing scheme which integrates memory, logic, and learning.

Although massively parallel software neural networks are commonplace in the field of artificial intelligence, these networks are ultimately operated by conventional computers with the Von Neumann architecture, and a separation between logic and memory units that does not exist in biological systems. Artificial neural networks also suffer from the "curse of dimensionality," a range of phenomena which occur when classifying or analyzing large-dimensional datasets, which do not occur in low-dimensional datasets. In machine learning, it is increasingly common to have datasets with very large dimensions. For example, a simple grayscale small image with 50x50 pixels has 2500 dimensions. If the images are RGB-colored, then the dimensionality suddenly increases to 7500 (one dimension per color channel per pixel). As the dimensionality

of the space increases, the number of possible data permutations grows exponentially, and thus the amount of information gained from observing a single data point decreases. Therefore, the amount of training data required to make meaning models increases exponentially, creating unwieldy time and power demands for computing circuits which operate in the serial Von Neumann architecture.

Neuromorphic engineering refers to the development of bio-inspired computing architectures to avoid the limitations of conventional computer design. In the future, neuromorphic circuits may enable extremely low power signal processing and learning in embedded systems and edge devices. To date, neuromorphic devices such as memristors and phase change memory, have demonstrated analog conductance tuning, fast switching, long memory retention, and parallel signal processing. However, they have not demonstrated the important self-learning function of the brain, which can spontaneously program its own synaptic weights via Hebbian learning, without the need for external computation or peripheral circuits. Development of a device which can fully mimic the synaptic property holds important implications for machine learning and artificial intelligence (AI). Four major properties of the brain motivate the development of the synaptic device presented in this dissertation.

Firstly, the human brain has superior computing energy efficiency. The human brain concurrently infers and learns from massive information via ~10^{11} neurons and ~10^{14} synapses in parallel analog mode with a modest power consumption of ~20 W and computing energy efficiency of ~10^{15} OPS/W[5] that is significantly superior to the energy efficiencies of transistor-based computing circuits (~10^{5}–10^{12} OPS/W)[11–13] and neuromorphic circuits based on analog synaptic devices (~10^{10}–10^{14} OPS/W for inference and $\lesssim 10^{12}$ OPS/W for learning)[14–18] . Secondly, the human brain has superior real-time learning functionality. The human brain

implements learning and inference algorithms in massive parallel analog mode with an extremely high computing speed (~10^{16} OPS), facilitating the real-time learning by saving the learning time spent on offsite computers. Third, the human brain has superior accuracy, performance, and adaptability in changing environments. The real-time onsite learning functionality also enables the human brains to adapt to unpredictable challenges and create new functions instantly in dynamically changing environments. Although the computational precision of analog neural networks is significantly inferior to that of the digital computers, the synaptic conductance matrix $\boldsymbol{w}$ and inference algorithms of the brain are dynamically optimized in a real-time statistical learning process, leading to superior accuracy and performance in changing environments, such as the perceptions of nonstandard images and speeches, sensorimotor learning in complex environments, and medical diagnosis. Fourth, the human brain can implement the real-time learning algorithm in arbitrary environments, leading to general intelligence of the human brain.

The superior energy efficiency of the human brain is illustrated in Figure 2 below, and compared to the trend of microprocessor efficiency doubling every 18 months, predicted by Koomey's law[19]. Before 2005, the improvements primarily came from the transistor channel length reduction. The trend has stagnated in the past two decades due to leakage currents, increasingly complicated manufacturing processes, and short-channel effects such as drain-induced barrier lowering, and progress has instead come from advancements in architecture rather than at the device level. The development of devices which emulate the synaptic function are a key requirement to enable the highly efficient neuromorphic circuits with brain-like energy efficiency. To date, no neuromorphic devices have demonstrated the real time self-learning

function in a massively parallel network.

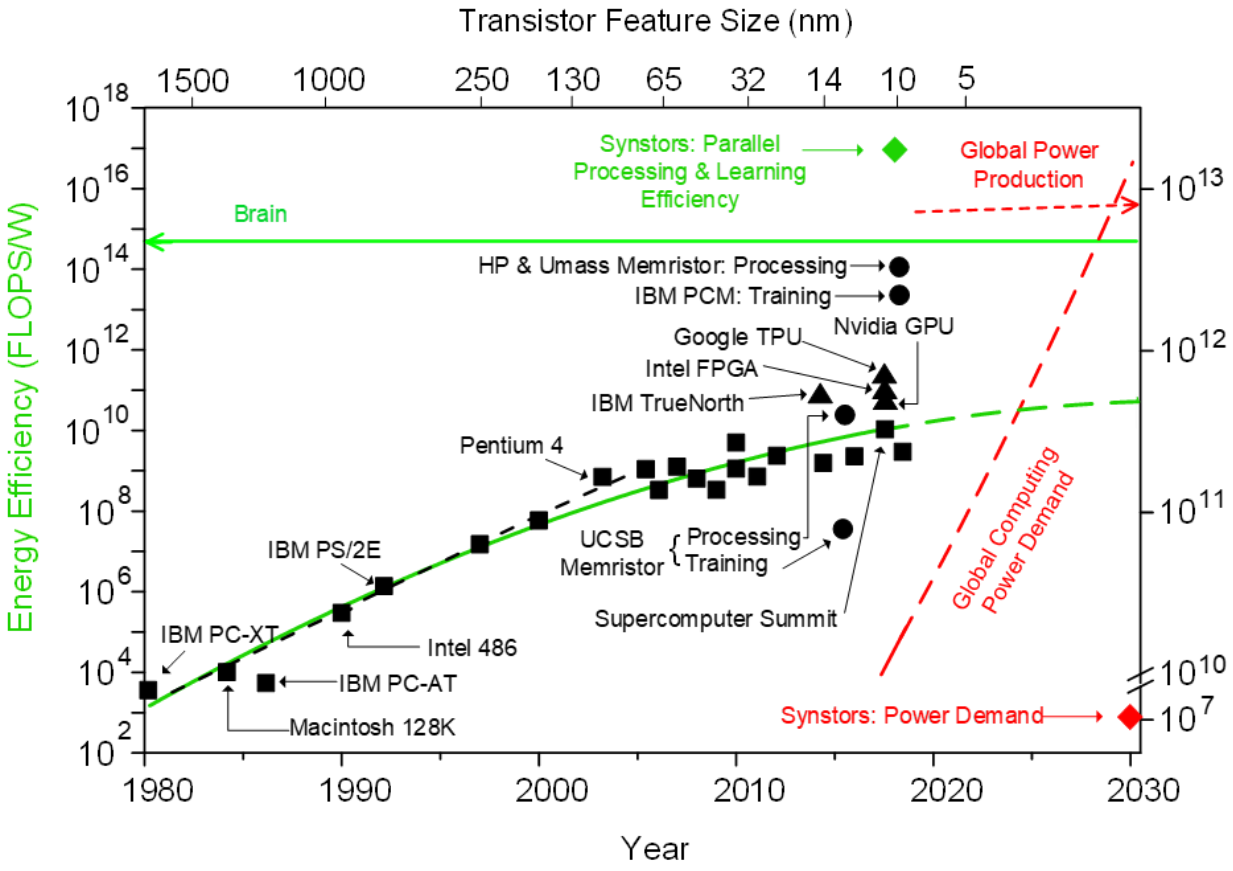

Figure 2: The energy efficiency of microprocessors (black squares), predicted to double every 18 months following Koomey's Law, with the dashed black line showing the trend prediction. The dashed red line indicates the global computing power demand[20].

The synaptic resistor (referred to as "synstor" hereinafter) is a new neuromorphic device which mimics the core synaptic function of the human brain. When an input voltage vector is applied to pre-synaptic neurons in a crossbar network, the synstors process the signal following Kirckoff's current law to produce an output signal, I, given by the sum of the currents across each synstor, where w represents the conductance (synaptic weight). When a potential occurs on

a pre- and post-synaptic neuron simultaneously, the conductance of the synstor w is modified in a Hebbian learning process, $w = \alpha V_i V_o$, with α as a learning coefficient.

Prior work on synstors has employed a p-type semiconducting carbon nanotube (CNT) channel and a $HfO_2/TiO_2/HfO_2$ charge trap heterojunction as a switching material. Previous work on carbon-nanotube (CNT)-based synstors have demonstrated the important real-time learning function. A 4x2 synstor crossbar array with two integrate-and-fire "neurons" was shown to perform speech recognition with an energy efficiency of ~10^{17} FLOPS/W, outperforming existing computers[20]. A 2x2 synstor crossbar array was used to optimize the lift and drag signals of a morphing wing in real time while inside a wind tunnel with a dynamic wind speed and angle of attack[21]. Lastly, a 4x4 synstor crossbar array was used to drive a drone to a target position, with no prior knowledge of the system, in a windy environment. The CNT-based synstor circuit demonstrated a performance at these tasks which exceeded that of humans performing the same task. Large-scale synstor networks could be used to solve the Von Neumann bottleneck and form the basis of a new computing platform with brain-like self-programming and adaptability in complex environments. To date, larger scale synstor networks have not been demonstrated. This is partly due to the unreliable processing of CNTS, whose variation in dimensions, chirality, and doping lead to significant device variability. This dissertation introduces a synaptic resistor based on Si, and shows improved scalability and uniformity, while maintaining the important real-time self-programming function.

1.1. Logic and Learning in Biological Synapses

The human brain, and other biological neural networks, have excellent energy efficiency and speed when processing, and especially when learning from, information with large signal dimensions in parallel. Creativity, adaptive learning, and pattern recognition are emergent

properties of advanced biological neural networks and are essential properties of an artificial general intelligence. Neuromorphic computing refers to the field of research in developing integrated circuits and systems whose architectures mimic biological neural networks. Therefore, it is useful to understand the basic operation principles of biological synapses and neurons.

The human brain is composed of some 10^{11} neurons and 10^{14} synapses. The synapses, an electrochemical junction connecting two neurons, is the basic information processing unit of the brain. Each neuron is connected to roughly 1000 other neurons via the synapses. Neurons collect and integrate potential pulses (termed pre-synaptic spikes) from neighboring synapses, and eventually trigger an action potential upon reaching a threshold (termed post-synaptic spikes). The spiking of this massively parallel network constitutes the basis of human learning and cognition.

In fact, the synapse is a junction, an empty space through which neurotransmitters can flow. Upon an action potential, "gates" in the neuron membrane are opened by the neurotransmitters, and an ionic current flows into the gates, propagating the signal. In electrical devices, the synaptic weight is typically represented by the device current or conductance, but biological synaptic weight is determined by how many neurotransmitter receptors are on the membrane of the post-synaptic neuron, and how many neurotransmitters are injected into the synaptic weight. A higher weight results in larger current, because more "gates" are open, or they are open for a longer time. Unlike electronic devices, the current is driven by diffusion instead of the action potential itself, which is simply required to activate the receptors in the post-synaptic neuron.

Learning in the brain is thought to be primarily the result of synaptic plasticity. Synapses exhibit spike-timing dependent plasticity (STDP), in which the rate of modification of the synaptic weight is proportional to the timing difference between pre- and post-synaptic spikes[22]. When the spikes are separated by a large time interval, it is likely that they are uncorrelated. When they occur together, it is likely that the pre-synaptic spike caused the post-synaptic spike, or vice versa, and the synaptic weight between the two neurons is strengthened. This is summarized by the famous quote of neuropsychologist Donald Hebb: "neurons that fire together, wire together[23]." In short, the timing difference between the pre-and post-synaptic spikes dictate the magnitude and direction of the weight change. To date, it is poorly understood how the neurobiology of synapses leads to higher-order properties such as instinct, creativity, and problem-solving.

1.2. Prior work in neuromorphic computing

With the approaching end of Moore's law, the energy efficiencies of transistor circuits are fundamentally limited by the energy cost on the signal transitions ($>10^{-11}$ *J*/bit), and are asymptotically saturated at ~10^{12} OPS/W[24]. Neuromorphic circuits based on analog electronic devices, such as synaptic transistors [25-26], memristors [14,15,16], and phase change memory (PCM) resistors[17,18] with integrated logic and memory functions circumvent the signal transitions between logic and memory units, and compute inference algorithms with energy efficiencies (~10^{10}–10^{14} OPS/W) significantly superior to those of conventional transistor-based circuits.

Nevertheless, the conductance matrixes ($\boldsymbol{w}$) of the neuromorphic circuits need to be modified by applying high writing voltages when a learning algorithm is executed, and the circuits execute

an inference algorithm with low reading voltages to avoid the change in $\boldsymbol{w}$. Thus, unlike a neurobiological network, a single neuromorphic circuit cannot concurrently execute inference and learning algorithms with the same voltage magnitudes. Moreover, due to inter-device variability, the analog neuromorphic circuits execute learning algorithms with much higher inaccuracy than the transistor-based digital computing circuits[27]. In order to execute algorithms accurately in the analog neuromorphic circuits, learning algorithms were executed in transistor-based digital computing circuits to obtain optimal conductance matrices of the neuromorphic circuits, then $\boldsymbol{w}$ were modified to optimal values in sequential writing and reading processes iteratively, which required separated memory and logic circuits, and signal transmissions between the circuits, thus limiting the energy efficiencies for learning to the range comparable or lower than those of the digital computing circuits ($\lesssim 10^{12}$ OPS/W)[14,17,18,24] . The energy and time consumption for computers to execute learning algorithms accurately from a big dataset with M-dimensional variables and combinatorial complexity increase exponentially versus M, referred to as the "curse of dimensionality"[28]. The tremendous costs of energy and time pose a major hindrance for edge computers in AI systems to execute learning algorithms onsite in real-time. Following the machine-learning protocol, the learning algorithms are usually executed based on big data technologies in offsite high-speed computers, with enormous power and time consumption to derive optimal inference algorithms that are then executed in edge computers in AI systems[1,3,4]. The derived inference algorithms can outperform humans in specific tasks under well-defined environments such as arithmetic computation, games[1], recognitions of standard images/speeches[2], and self-driving cars in normal environments[3,4], however, their performance in tasks beyond their learning domains, such as recognitions of nonstandard images/speeches, self-driving cars and robotic systems in unpredictable environments[29], and complex medical

diagnosis[30] are significantly inferior to human performance. The existing AI systems lack brain-like general intelligence beyond their learning domains and are unable to adapt to unpredictable challenges in changing environments.

Beyond these issues, existing neuromorphic circuits suffer from leakage or "sneak" current issues, which often necessitate the use of complicated peripheral circuits. Sneak current refers to currents flowing in unintended directions in the crossbar array as the various potentials applied on the pre- and post-synaptic neurons draw current from high-conductance synapses. The sneak current issue is especially difficult for devices, such as memristors, which require a large current for switching. In this regard, it is desirable for neuromorphic devices to have a low or zero current during switching, and to have a non-linear I-V curve, with small currents at low reading voltages. From an energy perspective, and to minimize sneak current, it is also useful to simply have synaptic devices with low conductance. This issue has been addressed by adding a selector device, such as a diode or transistor, in series with each synaptic element, which can be programmed externally to control what potential is experienced by the synapse, but this creates a complicated external circuit and increases the power consumption.

Selector devices are also required in many neuromorphic devices due to their linear tuning properties. A programming voltage applied across one device will equally be applied across all other devices in its respectful row or column, which makes it challenging to program individual devices. The solution is usually to use apply a half voltage, $+\frac{1}{2}V_t$ and $-\frac{1}{2}V_t$ respectively, on the row and column corresponding to the desired device, so that the tuning voltage V_t is experienced only by the desired device. This places a strict requirement on the neuromorphic device property that it should not be tuned at the half-voltage. It is desirable that the neuromorphic device's

tuning is a strong nonlinear function of the tuning voltage, in order to reduce the burden on external circuits.

1.3. Prior work in synaptic resistors

Our group has previously developed synaptic resistors based on carbon nanotubes, and used them for real time learning applications. The structure of the device is shown in Figure 3. The device is composed of a network of randomly oriented p-type semiconducting carbon nanotubes (CNTs), forming a channel, and connected via Schottky junctions to two Al input and output electrodes. The device also has a memory stack consists of a $HfO_2/TiO_2/HfO_2$ charge trap heterojunction structure and an Al bottom electrode which acts as a reference potential.

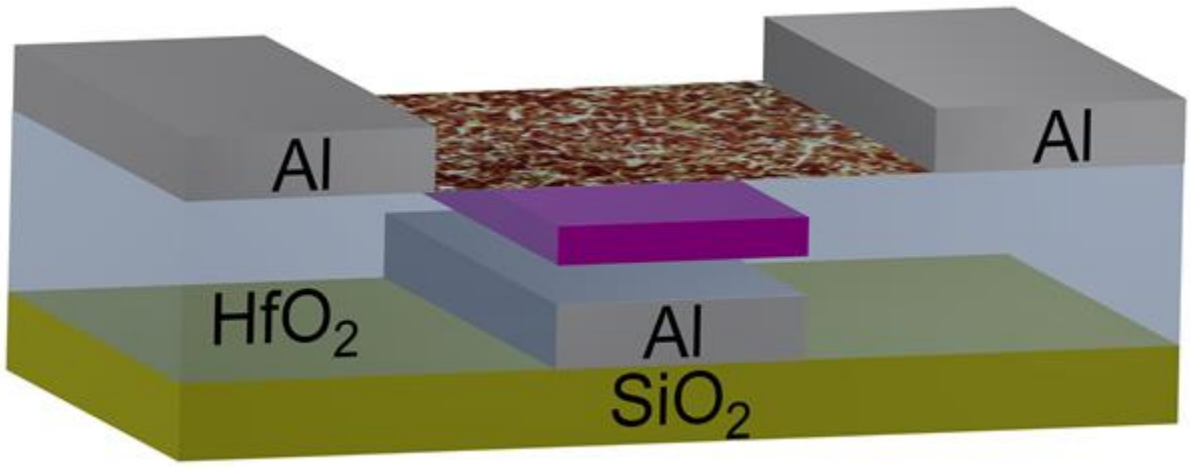

Figure 3: The device structure of a carbon-nanotube synstor. The synstor has Al input and output electrodes, connected via Schottky junctions to a random network of p-type semiconducting carbon nanotubes. The device has an Al reference electrode and a $HfO_2/TiO_2/HfO_2$ charge trap heterojunction as a switching layer.

The memory retention, endurance, band structure, and analog tunability have been previously investigated[20]. The CNT-based synstors have demonstrated the important real-time learning function. A 4x2 synstor crossbar array with two integrate-and-fire "neurons" was shown to perform speech recognition with an energy efficiency of ~10^{17} FLOPS/W, outperforming existing computers[20]. A 2x2 synstor crossbar array was used to optimize the lift and drag signals of a morphing wing in real time while inside a wind tunnel with a dynamic wind speed and angle of attack[21] . Lastly, a 4x4 synstor crossbar array was used to drive a drone to a target position, with no prior knowledge of the system, in a windy environment. The CNT-based synstor circuit demonstrated a performance at these tasks exceeding that of humans performing the same task.

The uniformity of a chip containing 400 synstors is described in Figure 4 below. The distribution of conductance is expressed as an error from the mean conductance. The data were obtained by applying a 10 ms -2 V pulse on the input electrode, of each device sequentially, and reading the current across the output electrode while the reference electrode is grounded. The devices were not pre-tuned before the measurement, so that they were in an equilibrium conductance state. The relative standard deviation, $\sigma/\overline{w}$, of the device conductance is 12.707.

The synstor is a memory device with analog tunability, and the synstors could be tuned to any analog value between their maximum and minimum conductances. However, there was a large variance in the equilibrium conductances, tuning rate, endurance, and retention of each device, making it difficult to implement a uniform crossbar array. Although the materials (99.9% pure semiconducting single wall CNTs) and processes were optimized to reduce variation in the device properties, the intrinsic variation in CNT densities, CNT directions, CNT diameters, CNT purity (semiconducting vs. metallic), CNT doping concentrations, charge

densities at Al/CNT interfaces, charge densities at CNT/HfO_2 interfaces can all hypothetically introduce variation in the final device property[20,31–34].

The variance in the device conductance, and tuning properties, led to difficulty in processing and learning from large-dimensional data. The insufficient reliability and scalability was a primary motivation to explore new device materials, and led to the development of the device presented in this work.

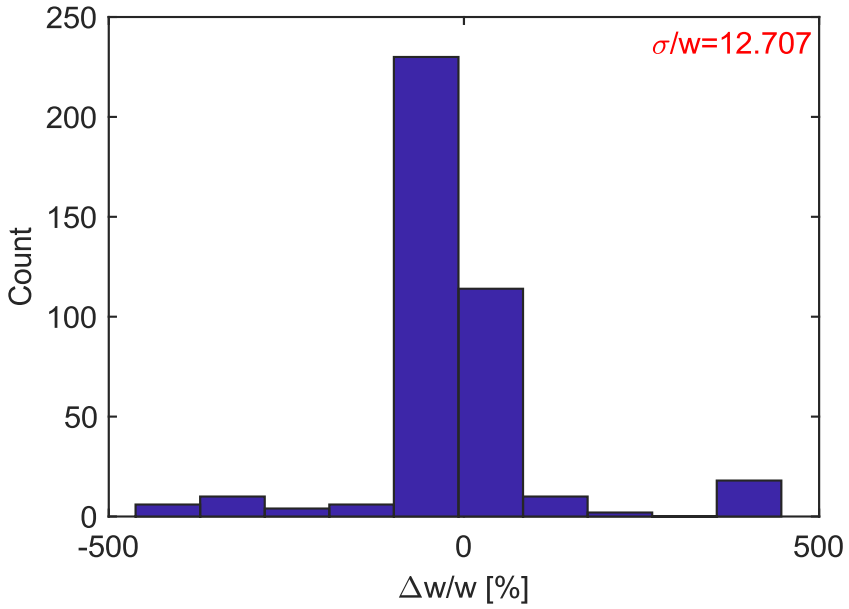

Figure 4: The conductance distribution of 400 carbon nanotube synstors on a chip, expressed as a percentage error from the mean conductance.

1.4. Research goals

The synaptic resistor should be a non-volatile memory device with analog conductance tunability. The memory retention should be such that the memory is not disturbed during reading operations, and while the circuit is at rest, and should be tunable by writing operations which consists of simultaneous writing voltages on the input and output electrode. This Hebbian-style learning will allow the circuit to implement arbitrary learning algorithms. The devices should have improved uniformity and scalability, and show the potential for large-scale integration, which has not been demonstrated in the previous generation of synaptic resistors.

The devices should be arranged in a crossbar circuit to enable vector-matrix multiplication operations between the input vector and the synaptic weight matrix for signal processing. The crossbar architecture will also allow for backpropagation of the output vector to coincide with the input vector and modify the synaptic weights via learning. To enable a selector-free architecture without sneak current, the device should have a low conductance compared to other neuromorphic devices. In general, the input current of the synstor is much lower than two-terminal neuromorphic devices, such as memristors, since the writing voltage itself is applied across the insulating oxide layers, rather than across the conductive or semiconductive channel[35].

Another requirement is that the device should be operated with the same magnitude of pulse for reading and for writing. In other neuromorphic devices, the reading and writing operations are performed with different voltage magnitudes, or across different terminals. For example, in memristors, reading is performed across the two terminals with a low voltage, to not disturb the memory state, while writing is performed using a high voltage. The high voltage is required to induce the formation or dissolution of a conductive filament (often by Joule heating) across the switching layer. In floating-gate transistors and other three-terminal neuromorphic devices, the

reading voltage is applied between the source and drain, while the writing voltage is applied between the gate and drain. The reading and writing voltages differ in magnitude here as well, with the writing voltage typically much larger (and erasing voltages larger still). The synstor is unique from other neuromorphic devices because it uses the same voltage magnitude for both reading and writing operations. During reading, a voltage pulse is applied across the input and output electrodes to induce a current proportional to the device conductance (Ohm's law). During writing, a simultaneous voltage pulse is applied across the input and output electrode, while the reference electrode is kept at ground. This operation changes the potential of the channel with respect to the grounded reference electrode, creating an electric field across the memory layer to induce switching. The coincident pulses are the same magnitude as the single pulse used for reading, and induce the redistribution of charge states in the defective switching layer, driving carriers towards or away from the silicon channel, which will modify the device conductance upon the next reading pulse. This operation mechanism is very useful for a highly parallel neural network. In a fully connected neural network composed of synstors at each crosspoint, no selector devices are required, and the writing voltage (voltage on the output electrode) will not affect the synstors which do not experience an equivalent input voltage. This operation is fundamentally different from other neuromorphic devices.

In this work, a synstor crossbar chip with the described properties has been designed and fabricated. The input electrodes of the device can be connected to sensors containing data of the system to be optimized, and the output electrodes are connected to an artificial integrate-and-fire "neuron" circuit, presented later in this dissertation.

1.5. Synaptic resistor & operation mechanism

The silicon synaptic resistor (synstor) device structure is shown in Figure 5. The synstor is composed of a 5µm-wide Si channel, and an input and output electrode with Schottky junctions to titanium silicide, connected by two Ti metal contact pads. Above the channel is a 13 nm SiO_2 thermal oxide, a 20 nm Al_2O_x switching layer, and a 18 nm TaO_y reference electrode. The switching layer is formed by an electron beam evaporation of 10 nm Al_2O_x and 10 nm Al, which results in a "metal-rich" oxide which serves as the basis of the memory function of the synstor. The synstor integrates spatiotemporal inference and learning in a single element, and functions as a biological synapse.

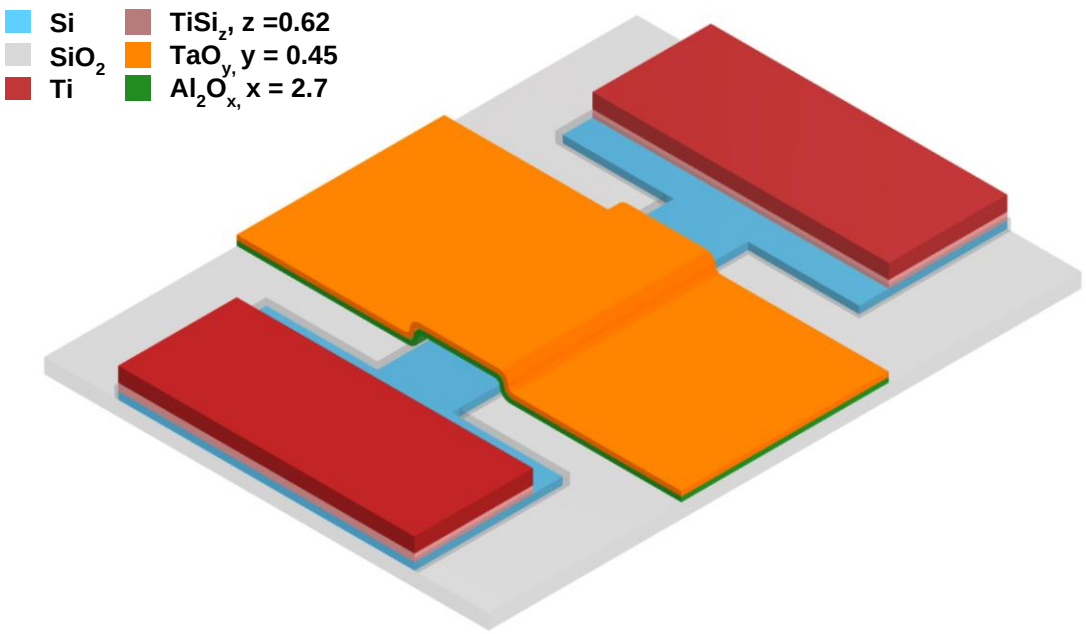

Figure 5: A synaptic resistor device schematic, showing a Si channel (blue), a thermal oxide (grey), a Ti input and output electrode (red), TiSi contacts to the channel (pink), an Al_2O_x switching layer (green), and a TaO_y reference electrode (orange).

The switching mechanism for the device is based on the charging and discharging of oxygen vacancies, V_o^{2+}, in the switching layer. The result of writing operations on the device is the redistribution electrons, hopping from the reference electrode or other nearby trap states, into the switching layer, which influences the carrier concentration in the channel via static field. A cross-sectional TEM image is shown in Figure 6, showing a 13 nm SiO_2 layer, a 20 nm Al_2O_x layer, and an 18 nm TaO_y layer. Proposed band diagrams for the synstor at various conductance states are shown in Figure 7 below.

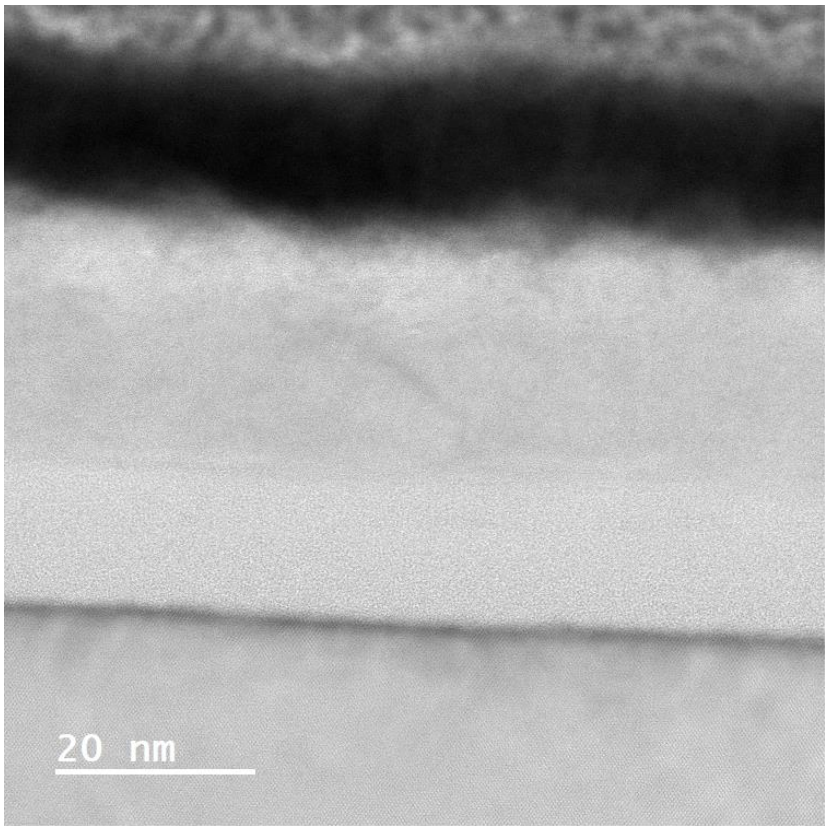

Figure 6: A cross-sectional TEM image showing the active region of the device. A Si channel is capped by a 13 nm thermal oxide, a 20 nm AlO_x layer, and an 18 nm TaOy layer.

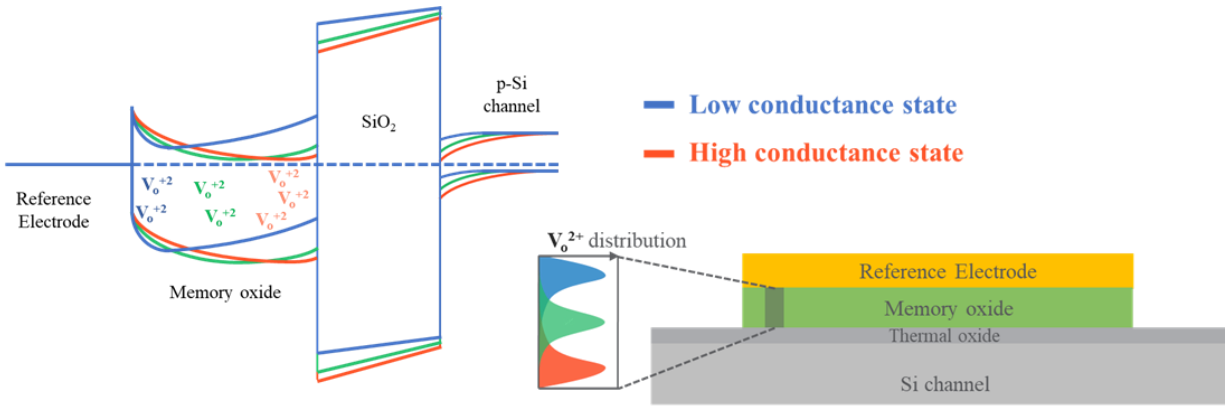

Figure 7: The band structure of a Si synstor in its low, equilibrium, and high conductance states, based on the defect distribution in the switching layer.

Amorphous aluminum oxide (a-Al_2O_3) films, grown by various deposition techniques, have a large defect density composed of O and Al vacancies and interstitials, and H interstitial centers. The defect distribution has recently been studied in a-Al_2O_3 and α-Al_2O_3 by Dicks et al [36] using density functional theory calculations. The presence of this charge could be undesirable in many digital electronic devices, but could be desirable in various memory devices, such as the synstor reported here. The proposed switching mechanism for the synstor is shown in Figure 8. At equilibrium, oxygen vacancies have a neutral charge. When the fermi level of the channel is raised with respect to the reference electrode, electrons hop from the oxide to the reference electrode, and the vacancies take a positive charge[37,38]. When the fermi level of the channel is lowered with respect to the reference electrode, electrons hop from the reference electrode to the oxide. After learning, the new charge state attracts or repels electrons in the n-type Si channel to modify its conductance. The presence of oxygen vacancies in the memory stack is confirmed by X-ray photoelectron spectroscopy (XPS) and energy dispersive x-ray spectroscopy (EDX)

analysis, presented later in this dissertation. The presence of vacancies is also inferred based on hysteresis in the current-voltage measurements appearing as a function of the voltage drop between the channel and reference electrode.

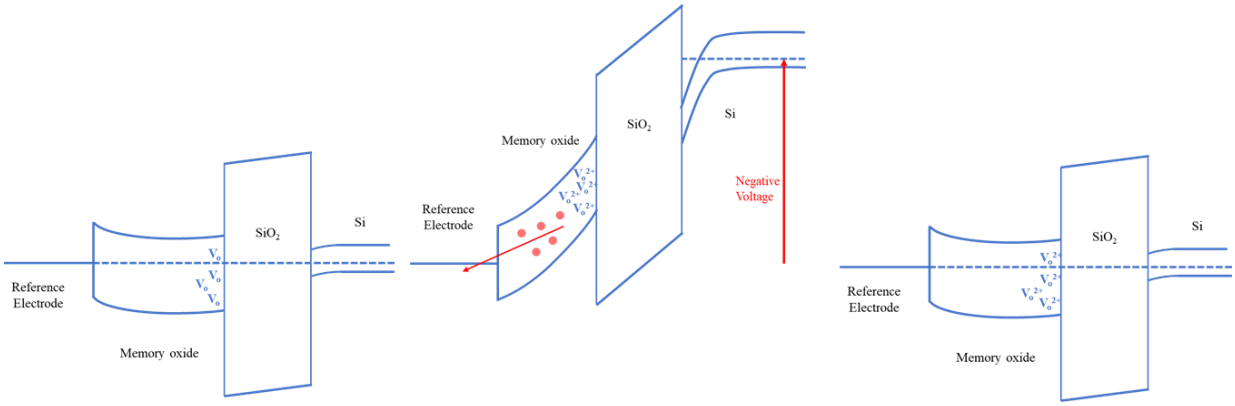

Figure 8: The proposed switching mechanism for the synstor. At equilibrium, oxygen vacancies have a neutral charge. When the fermi level of the channel is raised with respect to the reference electrode, electrons hop from the oxide to the reference electrode, and the vacancies take a positive charge. After learning, the new charge state attracts electrons in the n-type Si channel to decrease its conductance.

A synstor crossbar array is shown in Figure 9, connected to M pre-synaptic inputs and N post-synaptic outputs. For inference, a wave of voltage pulses, $V_i^m(t)$, in the m^{th} presynaptic neuron is processed by a synapse connected with the m^{th} presynaptic and n^{th} postsynaptic neurons, and induces a current in the n^{th} postsynaptic neuron[22], $I^{nm} = \kappa * (w^{nm} V_i^m)$ where w^{nm} denotes the synaptic weight (conductance), κ denotes a temporal kernel function, and $\kappa * (w^{nm} V_i^m)$ represents the temporal convolution between κ and $w^{nm} V_i^m$. For spatiotemporal parallel

inference, a wave of voltage pulses in presynaptic neurons induces a collective current via synapses in the nth postsynaptic neuron, which can be expressed as,

$$I^n(t) = \sum_m \kappa^{nm} * (w^{nm} V_i^m) \qquad \text{(Equation 1)}$$

and the current induces voltage pulses, $V_o^n(t)$, in the nth postsynaptic neuron. When the voltage pulse is fired in the postsynaptic neuron ($V_o^n \neq 0$), the postsynaptic current $I^n = 0$. The w^{nm} matrix is also modified concurrently by the spatiotemporal waves of voltage pulses in the presynaptic and postsynaptic neurons for learning,[20,22,23]

$$\frac{dw^{nm}}{dt} = \alpha V_i^m V_o^n \qquad \text{(Equation 2)}$$

where α denotes the conductance modification coefficient, and V_i^m and V_o^n voltage pulses have the same amplitudes and durations. w^{nm} is modified when $V_i^m = V_o^n$, with the learning coefficient $\alpha > 0$ in Hebbian learning, and $\alpha < 0$ in anti-Hebbian learning. α is a function of the timing difference between and pulses in the learning based on synaptic spike-timing-dependent plasticity (STDP). Based on Equation 2, general correlative learning algorithms in machine learning[20] can also be implemented. Following Equation 2, when $V_i^m \cdot V_o^n = 0$ (e.g. $V_i^m \neq 0$ and $V_o^n = 0$ during inference), $\frac{dw^{nm}}{dt} = 0$, i.e. w^{nm} remains nonvolatile for memory. By integrating the analog convolutional processing (Equation 1), correlative learning (Equation 2), and nonvolatile memory functions in a single synapse, the brain circumvents the fundamental limitations such as physically separated memory units, data transmission between memory and logic units in computers, and concurrently executes the inference (Equation 1) and learning (Equation 2) algorithms in a neural network in analog parallel mode. The previously reported CNT synstor demonstrated this function

and operation with an energy efficiency more than five orders of magnitudes higher than that of the Summit supercomputer[20].

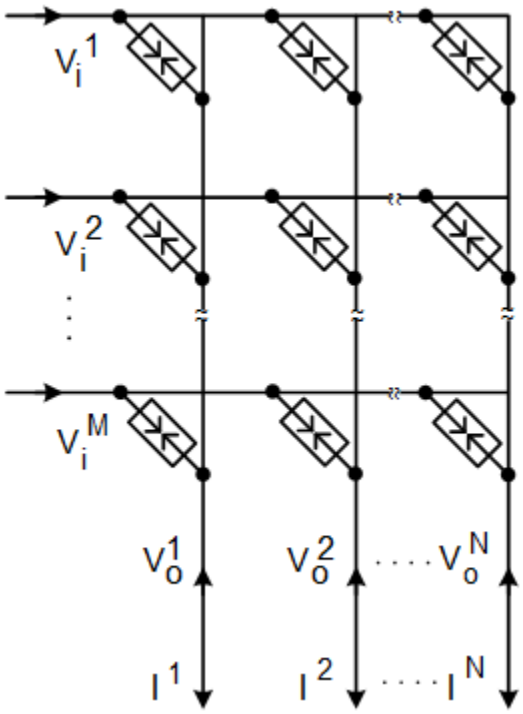

Figure 9: An MxN synapse crossbar array with M presynaptic inputs and N post-synaptic outputs. V_i^m denotes an input potential on the m^{th} presynaptic neuron, V_o^n denotes a potential on the n^{th} post-synaptic neuron, and I_n denotes a current flowing into the n^{th} post-synaptic neuron.

2. Methods

2.1. Device fabrication

The synstor is fabricated on a p-type silicon-on-insulator (SOI) wafer with $10^{15}cm^{-3}$ boron doping, a 145nm device layer, a 1µm buried oxide (BOX) layer, and a 700µm insulating handle layer. The process flow is shown in Figure 10. A photoresist was spin-coated on the Si surface, then exposed by ultraviolet (UV) photolithography (Karl Suss MA6), and developed

(AZ300MIF). The 5μm silicon channels are then etched by dry plasma etching (Technics FRIE, 1:4 O_2:CF_4, 100W). The photoresists are stripped by acetone, isopropanol, and de-ionized water. The SOI wafers are then cleaned using the standard RCA cleaning process[39]. The first step, RCA-1 (5:1:1 H_2O:H_2O_2:NH_4OH at 80°C) cleans organic residue from the surface. The next step is a 50:1 H_2O:HF dip to remove the native silicon oxide. The final step, RCA-2 (5:1:1 H_2O:H_2O_2:HCl at 80°C) cleans away metal ions and reduces further contamination by hydrogen passivation.

The wafers are then oxidized by thermal oxidation at 1000°C for 2 minutes, which results in a 13 nm oxide layer based on reflectometry (Nanospec 210). Since the silicon dry etch process does not have perfect selectivity of silicon to oxide, the process is performed after the dry etch process.

Next, another photolithography is performed to pattern the input and output (IO) electrodes. The negative photoresist is developed and baked, and then used as a dry etch mask to etch the exposed thermal oxide. The oxide in the IO electrode pattern is etched away, but the oxide in the channel and active area are protected by photoresist. Dry etching is used instead of the more typical wet etching of oxide by hydrofluoric acid (HF) to prevent the HF undercut profile which would leave exposed Si in the active region. After the dry etching (Oxford RIE, 1:1 CHF_3:Ar, 100W), the wafers are loaded into an electron beam evaporation (CHA Industries Mark 40) and the 300 nm Ti IO electrodes are deposited by liftoff. The photoresist is stripped by n-methyl-2-pyrrolidone (NMP) at 75°C, and the wafers are treated by a short oxygen plasma to descum the surface (Technics FRIE, 100 mtorr O_2, 50 W).

Then, a 40 nm Al_2O_3 sacrifical layer is grown on the surface by atomic layer deposition (Fiji Ultratech ALD). Then, the wafers are annealed in forming gas (5% hydrogen in N_2) at 460°C for 30 minutes to induce the reaction of titanium and silicon to form titanium silicide. The Al_2O_3 sacrificial layer protects the metal on the wafer from unwanted oxidation and hydrogenation during the anneal. After the anneal, the sacrificial layer is selectively etched away using a 3% tetramethylammonium hydroxide aqueous solution. Then, a layer of 50 nm Ti contact pads are deposited on the 500x500 μm silicon and metal pads by liftoff to improve the electrical contact.

Finally, the memory layer and reference electrode (10:10:8 nm Al_2O_3:Al:Ta) are deposited by a single photolithography and e-beam evaporation (CHA Industries Mark 40). E-beam evaporation, rather than the more typical sputtering method, is used for the memory oxide for two reasons. First, e-beam evaporation results in high defect generation in the deposited film due to the differing vapor pressures of aluminum and oxygen, and this non-stoichiometric film is desirable from a memory perspective. Secondly, deposited the memory layer and reference electrode in a single process results in a self-aligned structure. Depositing the reference electrode in a separate process would result in lithographic misalignment, and some electrode metal directly contacting the silicon dioxide, which would shield the channel from charges stored in the memory layer. Lastly, depositing all three layers without breaking vacuum minimizes the interface layers formed between them.

After the liftoff of the memory and reference electrode stack, the wafers are passivated by a thermal evaporation of parylene-C (SCS PDS-2010 Parylene). This permanent capping layer protects the sensitive silicon, silicon oxide, and memory layers. A final photolithography and dry etch (Technics FRIE, 100 mtorr O_2, 100 W) is used to etch the parylene on top of the contact pads only, so electrical contact can be made to the synstors.

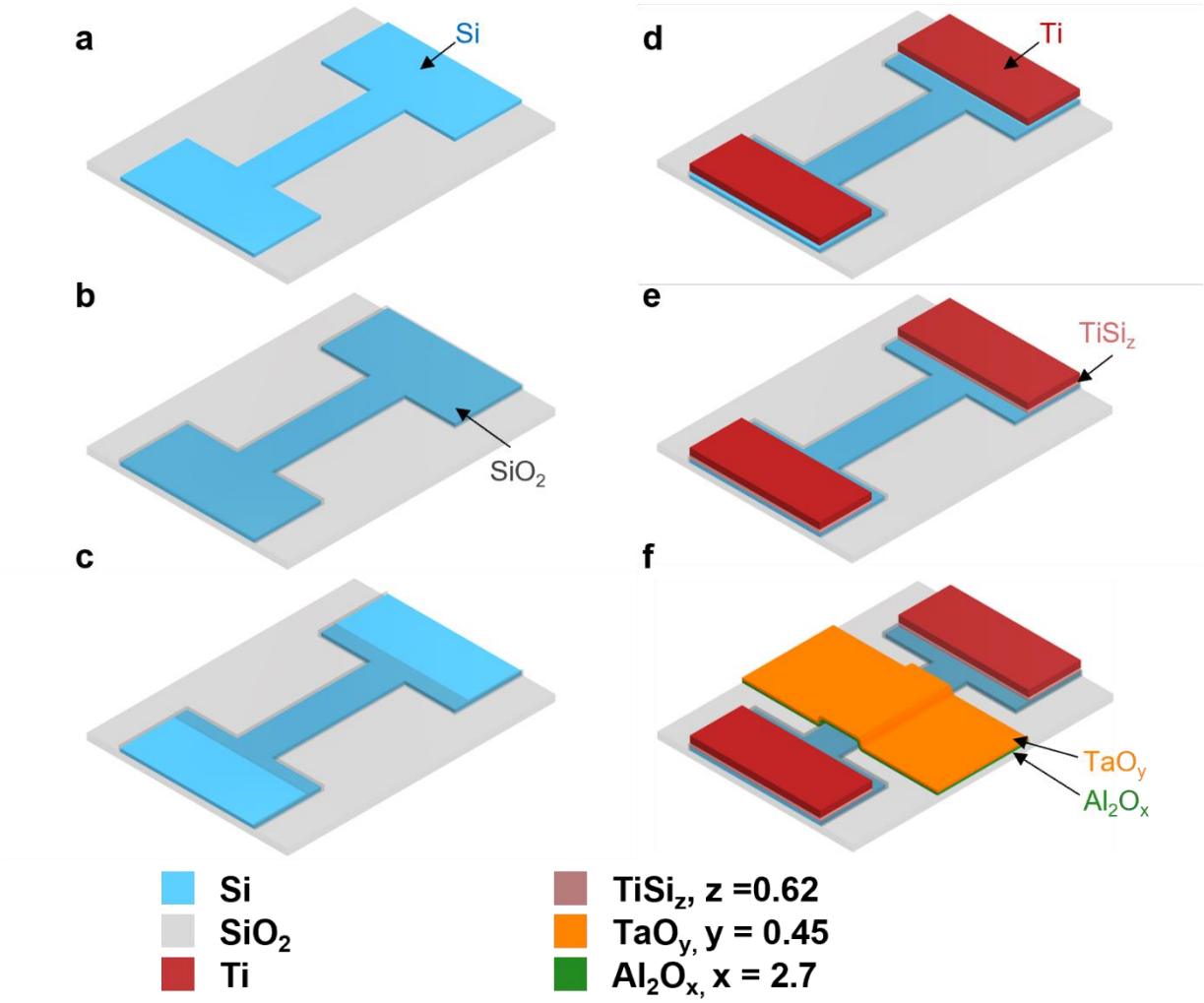

Figure 10: Process flow for the synstor fabrication. (a) A Si channel is etched into a p-type SOI wafer, and (b) oxidized. (c) Thermal oxide in the contact area is dry etched. (d) Ti input and output electrodes are deposited by e-beam evaporated and (e) annealed to form titanium silicide Schottky contacts. (f) Finally, the memory stack and reference electrode are deposited by e-beam evaporation and liftoff.

2.2. System Integration

The wafers are cleaved in 30x30 mm chips, which interface into a custom-build adapter (Ironwood Electronics) via compression insert and pogo pins (spring-loaded pins). The adapter contains 430 pins each with 1 mm pitch. The pins are arranged into a 20x20 grid for input electrode pads, 20x1 grid for output electrode pads, and 10x1 grid for reference electrode pads. Each pogo pin contacts a corresponding 500x500 µm metal pad on the chip side. An optical image of the completed chip is shown in Figure 11, with the parylene passivation layer omitted for clarity.

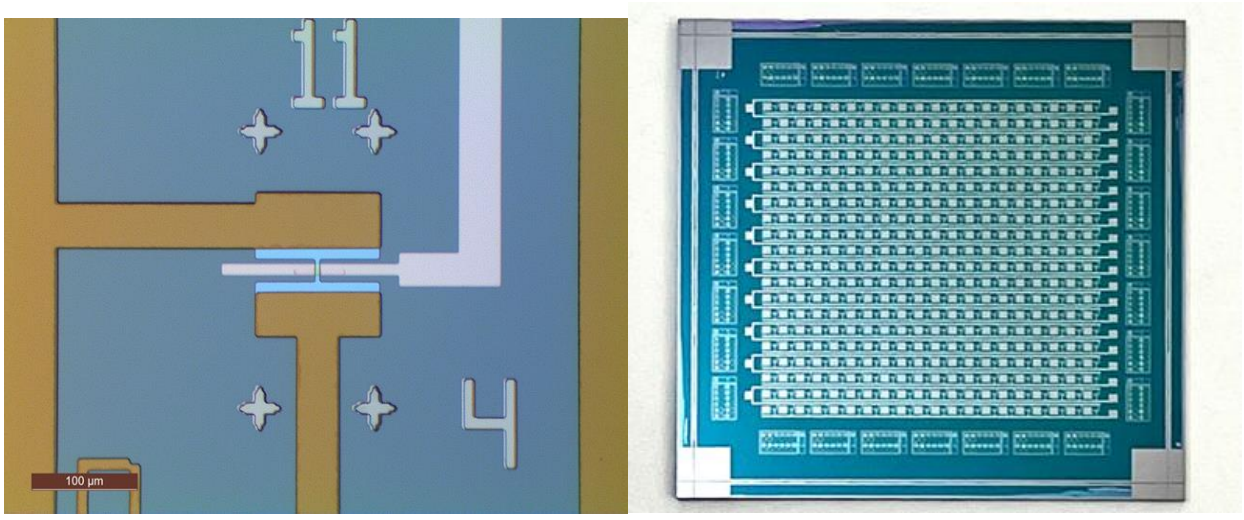

Figure 11: (Left) Microscope image showing the active region of a synstor in a single crosspoint. (Right) Optical image of a chip with 400 synstors arranged in a 20x20 crossbar.

The chip layout groups together all output, and reference electrodes within a row. The crossbar is formed when the chip is inserted into the adapter, which groups together all input electrodes within a column. Shorted or stuck devices can be isolated from the crossbar by

simply removing the corresponding pogo pins. Similarly, arbitrary crossbar sizes smaller than 20x20 can be constructed by removing all the unnecessary pogo pins. This is usually also important to minimize sneak current. The adapter used to interface with the synstor chip is shown in Figure 12.

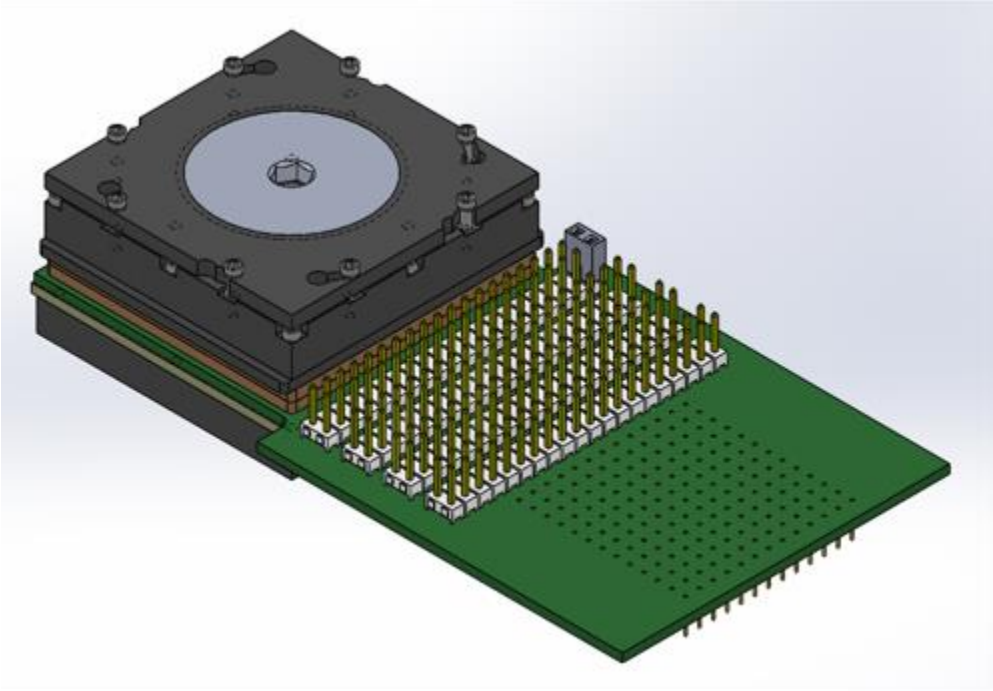

Figure 12: A custom pogo pin adapter for interfacing the synstor chip with external circuits

The adapter connects to a PGA socket on a custom printed circuit board (PCB) to characterize and operate the synstors. The adapter also comes with a conductive path to the backside of the chip for applying substrate bias. Even without removing any pogo pins, individual devices can be accessed using jumpers on the PCB which connect to each row and column, but in that case the sneak current should be considered.

The PCB serves as the interface between the synstor chip and a compactRIO controller (National Instruments, NI-9063) with Zynq-7020 FPGA. The FPGA controls 32 analog output (AO) channels, 32 analog input (AI) channels, and 32 digital input/output (DIO) channels. Voltage pulses are applied to the input, output, and reference electrodes using the AO channels, and device currents are converted into voltages on the PCB and measured by the AI channels. The device currents were converted to voltages through an operational amplifier (Microchip Technologies, MPC6022), in an inverting op-amp configuration. The output voltage of the op-amp, V_{out}, was converted to the device conductance following $w = \frac{I}{V_i} = -\frac{V_{out}}{V_i R_f}$, where R_f is a feedback resistor between the inverting input and output terminals of the op-amp. This system was used for the electrical characterization and measurements shown in this work. Leakage current measurements were performed on a semiconductor parameter analyzer (Keithley 4200).

2.3. An Integrate-&-Fire "Neuron" Circuit

Integrate-and-fire neuron circuits were designed, as shown in Figure 13, to emulate the basic functions of biological neurons according to the Hodgkin-Huxley neuron model.[40] Current from synstors, I, flows into a capacitor, C_{IF}, thus increasing its potential . A leakage current, I_L, flows through the resistor, R_L, decreasing V_I. V_I is proportional to the integration of $I - I_L$ with respect to time. When V_I reaches a threshold value, a Schmitt trigger composed of transistors M_1– M_6 is switched back and forth to generate an output pulse from the output channel, V_f. The output pulse resets V_I back to zero by switching transistors M_7, M_8, and M_9, and the capacitor C_I restarts the integration of the current. The transistors in the circuit are operated in their subthreshold regions. A "neuron" circuit with C_{IF}=10 nF, R_L=0.5 $M\Omega$, V_L=−30 mV, and R_{inv} =5 $M\Omega$ was tested by applying a series of 10 ns-wide pulses, with varied firing rates over a resistor with a resistance of 5 MΩ to inject a current I to C_{IF}, and the average firing rate of the output pulses triggered from the

circuit, r_f, is plotted as a function of the average magnitude of the current, I, in Figure 14. The firing rate of the pulses output from the "neuron" circuit, r_f, is plotted versus the average current, $\bar{I}$, input to the "neuron" circuit (open circles). The experimental data were fitted by a sigmoid function with $r_f = \frac{r_s}{1+e^{-\chi[\bar{I}-I_0]}}$ with $r_s = 152\ Hz$, $\chi = 0.26$ /nA, and $I_0 = 16.0$ nA.

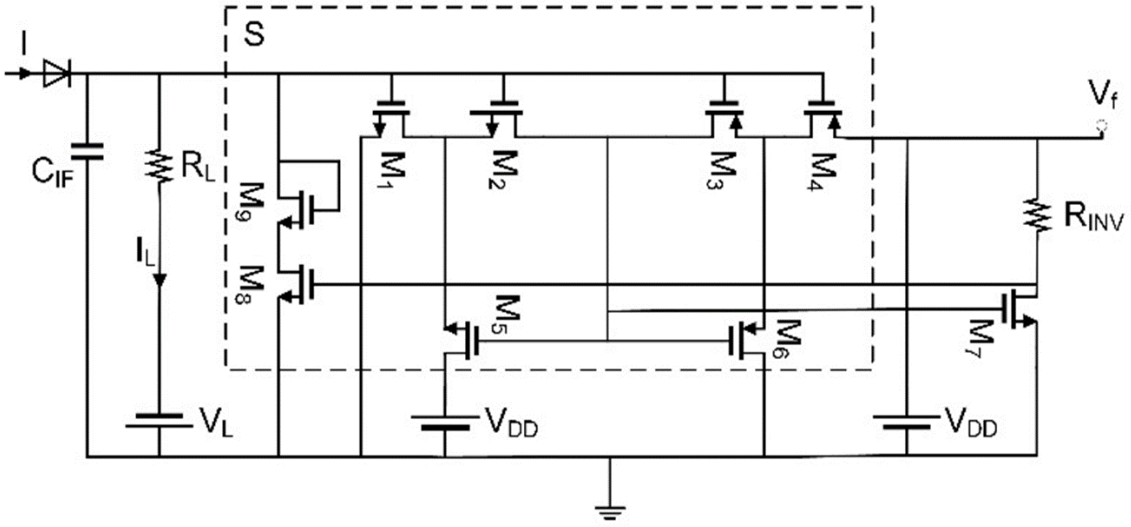

Figure 13: An-integrate-and fire "neuron" circuit

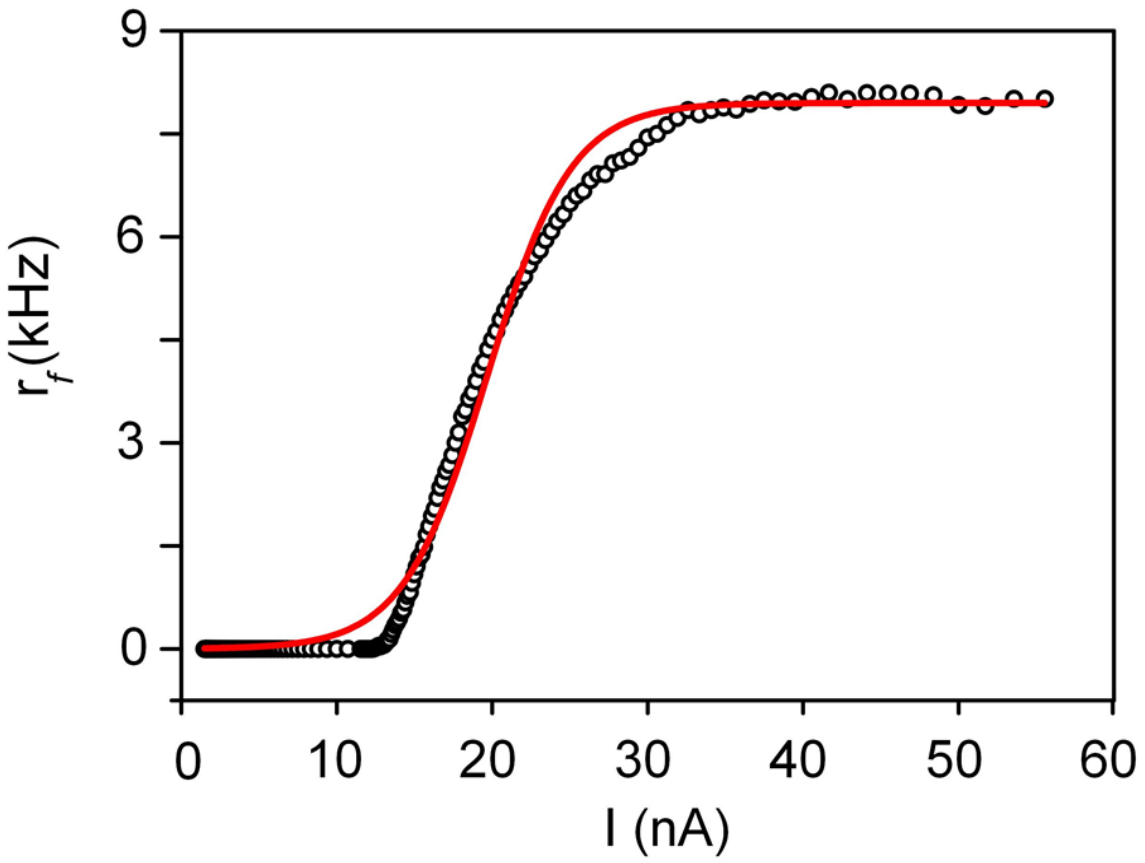

Figure 14 The firing rate of the pulses output from the "neuron" circuit, r_f, is plotted versus the average current, $\bar{I}$, input to the "neuron" circuit (open circles). The experimental data were fitted by $r_f = \frac{r_s}{1+e^{-\chi[\bar{I}-I_0]}}$ (red line) with $r_s = 152\ Hz$, $\chi = 0.26$ /nA, and $I_0 = 16.0$ nA.

3. Results

3.1. Current-Voltage Measurements

The synstor was tested by applying a continuous triangular voltage sweep, V_i, on its input electrode, while measuring the current through the grounded output electrode and grounding the reference electrode. The nonlinear rectifying I-V_i curves demonstrate the formation of a Schottky contact between the Ti input electrode and the n-type Si semiconducting channel. The

I-V_i sweep exhibits hysteresis, which is a consequence of applying a voltage larger than the Schottky barrier height, which induces a voltage drop between the channel and grounded reference electrode (ie. across the charge trap layer).

The synstor was further tested by applying a continuous triangular voltage sweep, V_{ref}, across the reference electrode, while applying a constant reading bias, V_i, on the input electrode, and measuring the current across the grounded output electrode. The results of the I-V_i and I-V_{ref} sweeps are shown in Figure 15 below. The I-V_{ref} sweeps shows a significant hysteresis owing to the large defect density in the charge trap layer. When the reference electrode voltage is positive, electrons in the memory oxide hop to the reference electrode and leave an increased density of positively charged O vacancies, V_O^{2+}, which results in an increase in the carrier concentration in the n-type semiconducting channel. When the reference electrode voltage sweeps back to negative, the V_O^{2+} density is reduced, and the channel is shifted to a depleted state, which remains in place until further programming (supported by memory retention tests in section 3.2). The dielectric constant of Al_2O_3 is much higher than SiO_2, so the conductance change observed is attributable to changes in the Al_2O_3 switching layer during the sweep.

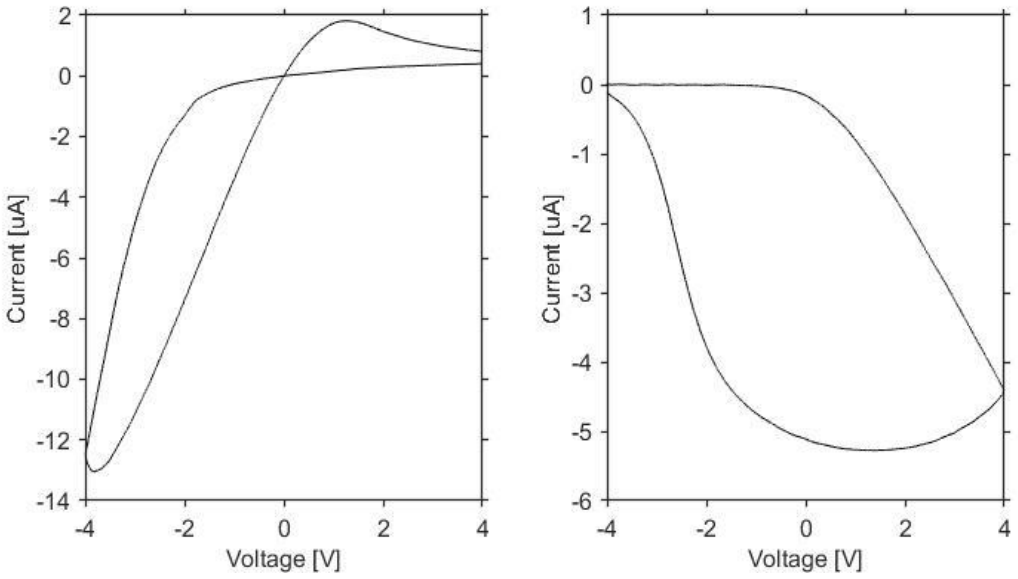

Figure 15: (Left) I-V_i measurement obtained by applying a continuous triangular voltage sweep, V_i, on the synstor input electrode, while measuring the current through the grounded output electrode and grounding the reference electrode. (Right) I-V_{ref} measurement obtained by applying a continuous triangular voltage sweep, V_{ref}, across the reference electrode, while applying a constant reading bias, V_i, on the input electrode, and measuring the current across the grounded output electrode.

The leakage current across the switching layer was similarly tested using a semiconductor parameter analyzer (Keithley 4200) and a probe station. The synstor, as a non-volatile memory device, should have minimal leakage current during reading, writing, and at rest. The leakage was tested by applying a triangular voltage sweep, V_{ref}, across the reference electrode, while grounding the input and output electrode, and measuring the current across the grounded output electrode. A probe station was used to eliminate the chance of leakage sneak paths from

adjacent devices. The leakage current is shown in Figure 16 below. The leakage is 6 orders of magnitude lower than the device current at -4 V.

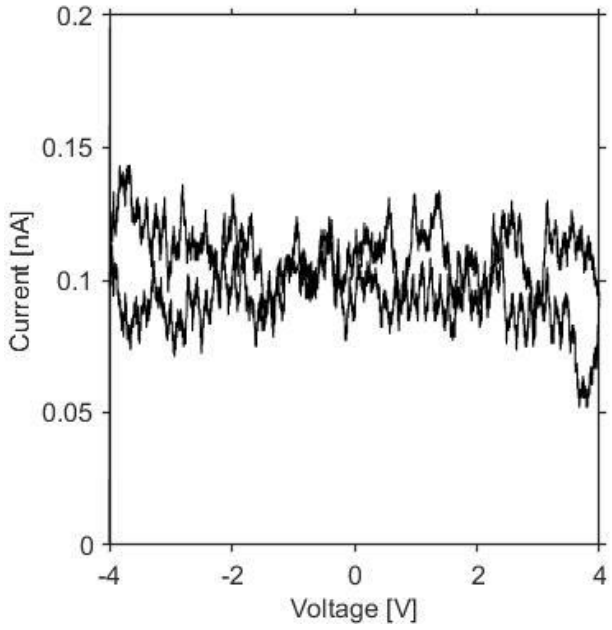

Figure 16: Leakage current measurement obtained by applying a continuous triangular voltage sweep, V_{ref}, on the synstor reference electrode, while grounding the input and output electrode and measuring the current through the grounded output electrode

3.2. Memory Endurance & Retention

Memory endurance and retention are two common figures of merit for non-volatile memory devices. The endurance is obtained by writing and erasing the device memory in rapid cycles, and measuring the decay in tuning range. The synstor is an analog memory device which can be tuned to arbitrary conductances between its minimum and maximum, but only the decay in the minimum and maximum conductances are measured here.

The synstor endurance is shown in Figure 17. A train of 50 -3 V (+3 V) 10 ms coincident pulses on the input and output electrodes was used to turn on (turn off) the synstors. After each cycle, a 10 ms -3 V read pulse on the input electrode was used to sample the conductance, and then the opposite cycle was performed. The minimum conductance was within the system noise of the measurement system. The minimum and maximum conductances were stable over 45000 cycles.

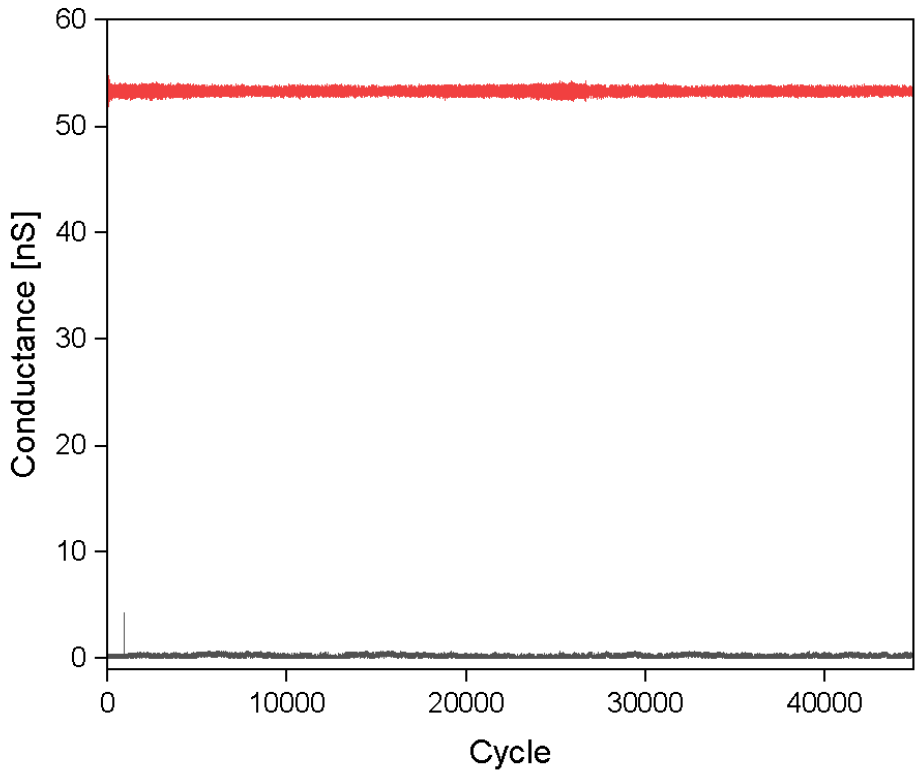

Figure 17: The endurance of the synstor is plotted as a function of cycle number. The minimum and maximum conductances are shown in black and red respectively, over 45000 tuning cycles. Each tuning cycle consists of a train of 10 ms ±3 V tuning pulses on the input and output electrodes, followed by a single 10ms -3 V read pulse on the input electrode.

In conjunction with the endurance test, the memory retention test measures how long a programmed state takes to decay. Non-volatile memory devices, such as SONOS memory, have a lifetime on the order of 100 years[41]. The retention results are shown in Figure 18 below. For each programmed conductance state, the retention is obtained by a +1 V 10ms read pulse on the input electrode, while the output and reference electrodes are grounded. The programming was performed by pulses on the reference electrode with various tuning magnitudes from 0 to -5.5V. The read pulses have a period of 100 seconds, and the chip is placed in an electrically and thermally isolated environment. 50 different analog conductance states are shown, within the range 0-100 nS, and the retention is measured over 10^4 seconds. The curves are fitted to the form $w = at^k$, where a and k are constants. The retention is extrapolated over a 10 year period to show that the conductance stances can still be distinguished from one another even with some decay. The device retention is appropriate for dynamic real-time learning applications where the environment is constantly changing, and the optimal strategy needs to be learned and re-learned quickly by the neural network for best performance.

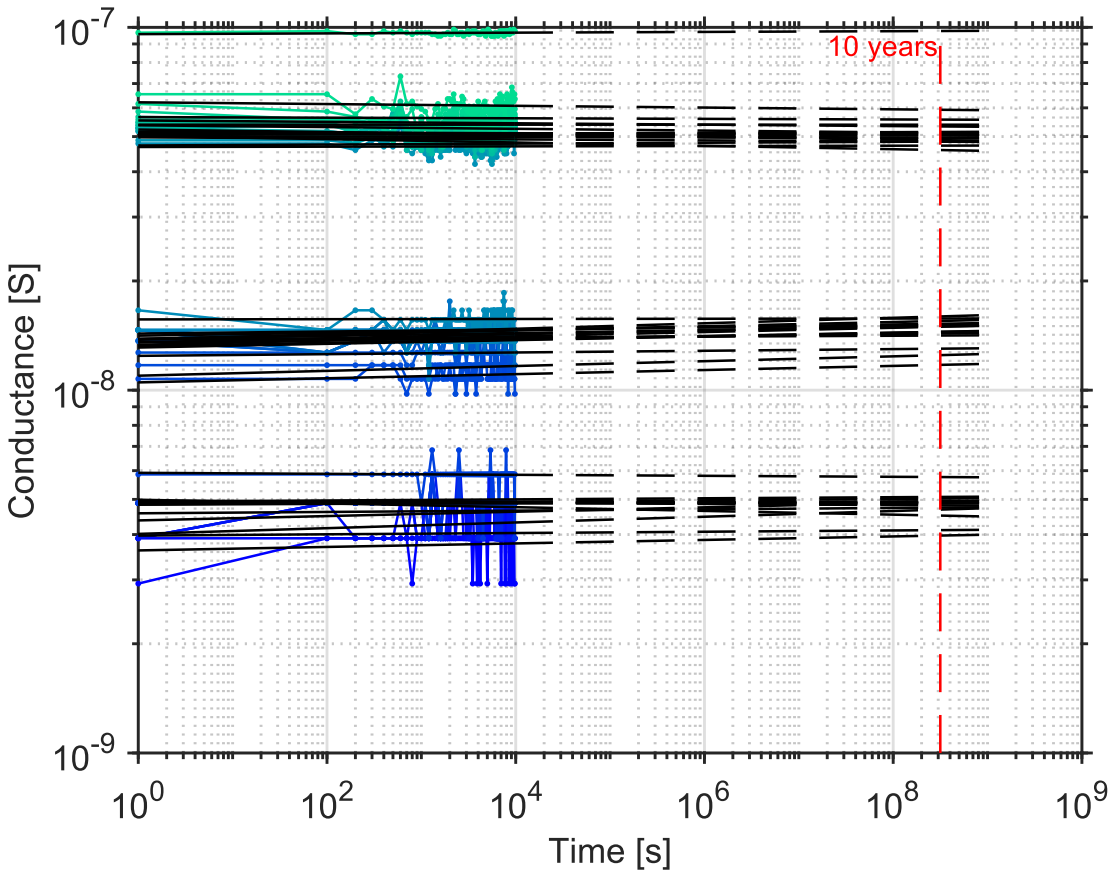

Figure 18: Memory retention of the synaptic resistor. 50 analog conductance states are measured and in blue, separated into four conductance groups, and fitted by exponential functions shown in black. The retention is extrapolated over 10 years, showing that the individual conductances are still resolved.

3.3. Nonlinear analog conductance tuning

The synstor network is typically operated in pulse mode, where the input pulse frequency is based on sensor inputs and the output pulse frequency is based on a learning algorithm. If an input pulse occurs on a synstor, it will process the signal by generating a current following Equation 1. If an input pulse and an output pulse occurs on a synstor at the same time, the

current across the channel will be zero, and the channel potential will be modified with respect to the reference electrode, inducing a shift in the charges of the memory layer, following Equation 2. This constitutes the learning function of the synstor. In Figure 19 below, the relative conductance change is plotted versus pulse trains of various combinations of input and output pulse magnitudes and polarities, with a duration of 100 pulses and a pulse width of 10 ms.

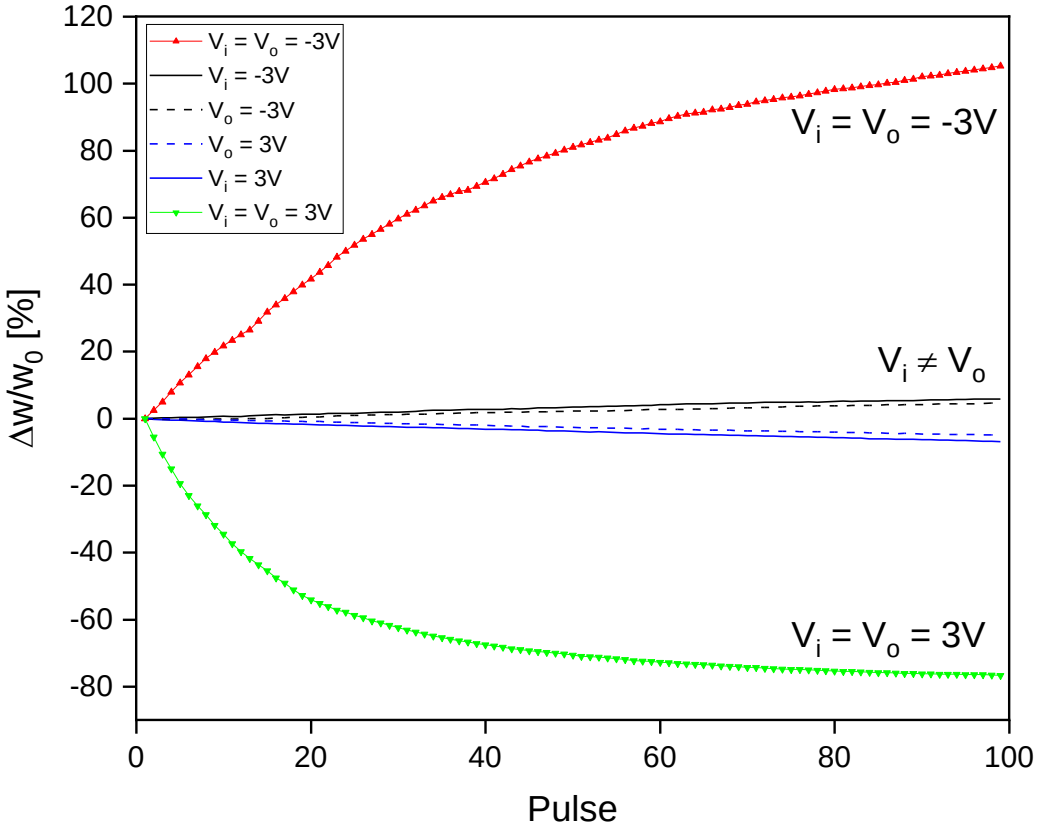

Figure 19: The non-linear analog conductance change of the synstor is plotted as a function of pulse number, for a train of 10 ms pulses. The conductance changes caused by coincident -3 V (red triangles) and +3 V (green triangles) coincident pulses is much larger than the conductance

change caused by +3 V input or output pulses individually (black and blue solid lines respectively), or by -3 V input or output pulses individually (black and blue dashed lines respectively).

The conductance changes caused by coincident -3 V (red triangles) and +3 V (blue triangles) coincident pulses is much larger than the conductance change caused by +3 V input or output pulses (black and blue solid lines respectively) individually, or by -3 V input or output pulses (black and blue dashed lines respectively) individually. As can be seen, the synstor experiences very little change in conductance when a pulse train of either polarity is applied on either the input or output electrodes. This corresponds to the signal processing mode of the synstor. On the other hand, the conductance change is much larger when two pulses occur simultaneously on the input and output electrode, similar to STDP in the brain. Non-linear tuning, or conductance change as a function of the timing and polarity difference between the input and output pulse, enables synstor circuits to have real-time learning and parallel signal processing and learning.

The non-linear tuning is also shown as a function of pulse amplitude in Figure 20 below. The pulse trains had a duration of 100 pulses and a pulse width of 10 ms, and voltage magnitudes ranging from -3 V to +3 V. When the coincident pulse amplitude is small, the tuning is also small. At voltages larger than 1.2 V, the positive (green triangles) and negative (red triangles) coincident pulses cause much larger conductance change compared to the individual input pulses (solid line) or output pulses (dashed line). The conductance change scales exponentially as a function of the magnitude of the coincident tuning pulse. This corresponds to the defects (negative charges) migrating further or closer to the silicon channel upon larger magnitude of

tuning pulses. The defects in turn repel the carriers in the n-type silicon channel to modify its conductance. The presence of the Schottky contact, and the non-linear relationship between the threshold shift and current magnitude in the subthreshold regime, contribute respectively to the insignificant conductance change when current flows through the channel and to the non-linearity of conductance change with respect to voltage amplitude.

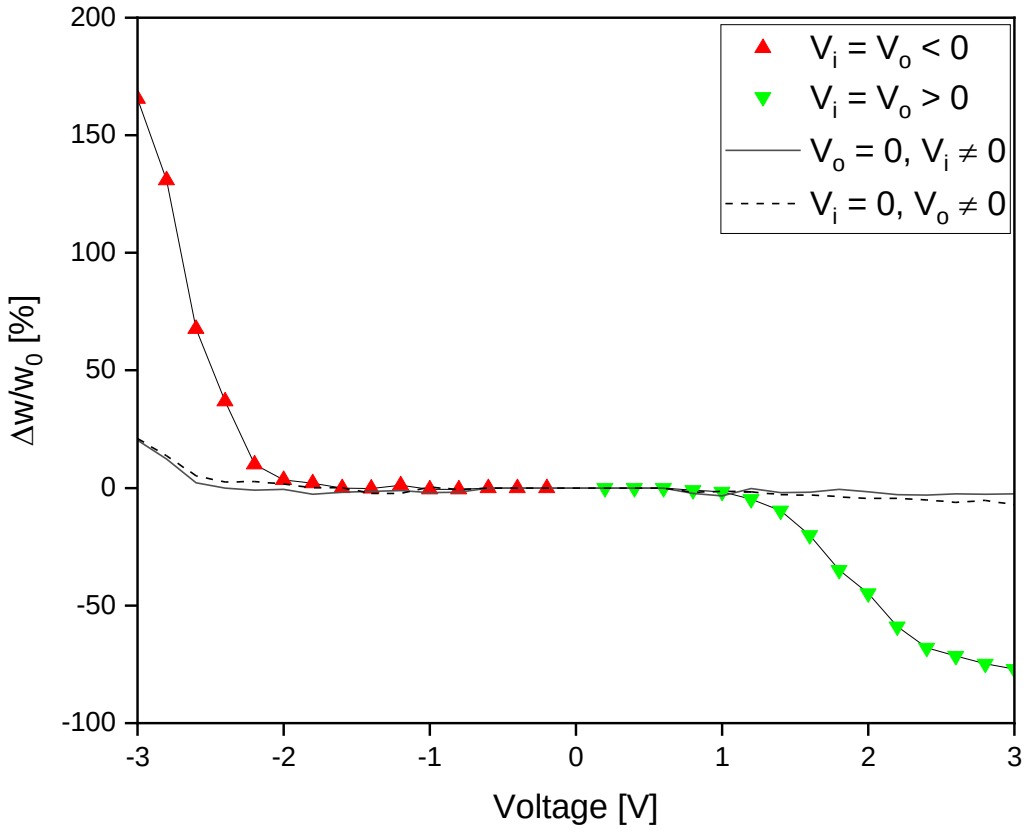

Figure 20: The non-linear analog conductance change of the synstor is plotted as a function of tuning voltage. When a train of 100 10 ms tuning pulses is applied on the input electrode (solid line) or output electrode (dashed line) only, the conductance change is small. When a train of positive input and output pulse are applied simultaneously (green triangles), the synstor is turned

off as a function of voltage. When a train of negative input and output pulse are applied simultaneously (red triangles), the synstor is turned on as a function of voltage.

3.4. Device Uniformity

A major concern for previous generations of synstors, based on materials other than silicon, was the scalability. Although individual devices demonstrated the synapse-like property, their conductances and tuning properties had a large variation, making it difficult to operate large-scale neural networks. In contrast, silicon and silicon processing have been studied exhaustively in the semiconductor industry for decades, which have allowed the creation of very large scale integration (VLSI).

In Figure 21, the relative standard deviation of the Si synstor is compared to the previously reported CNT synstor. The relative standard deviation is reduced by a factor of 29.

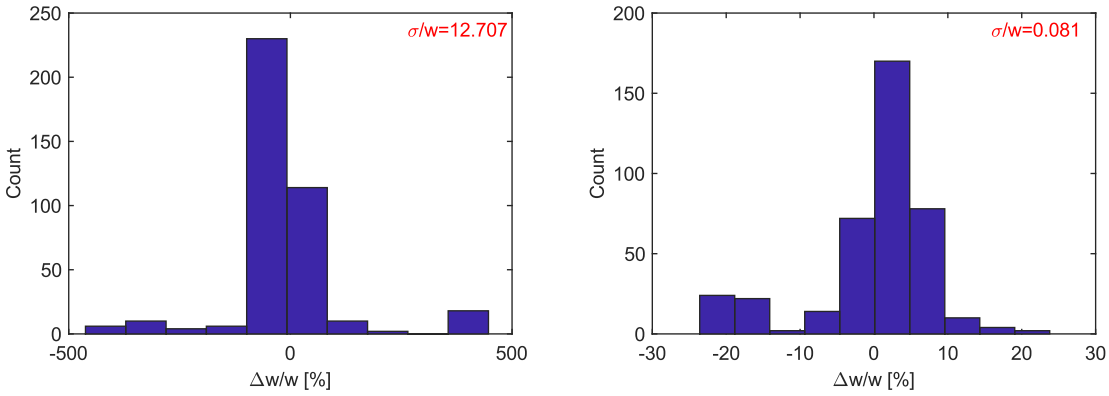

Figure 21: The conductance distribution of a carbon nanotube synstor chip (left) vs. a Si synstor chip (right), expressed as expressed as a percentage error from the mean conductance. The relative standard deviation of the Si chip is smaller by a factor of 29.

To demonstrate the uniformity and analog scalability of the chip, 300 silicon synstors were tuned, using a train of 10 ms ±3 V pulses, to 100 different targeted analog conductance values. In Figure 21a, the vertical and horizontal position of each marker respectively show each target conductance, and the mean conductance of the corresponding population. The height of each marker corresponds to 3σ. In Figure 21b, the height of the bars indicate 3σ, and the width of the bars indicate the error between the target conductances and the mean of the population. As can be seen, this population of 300 synstors can be tuned with high precision to many discrete analog conductance values separated by less than 1 nS. The error between the mean values and the target values is much smaller than the separation between each target values, which shows that the populations are resolved. The synstor could thus be useful in a neural network where precise analog memory values are required.

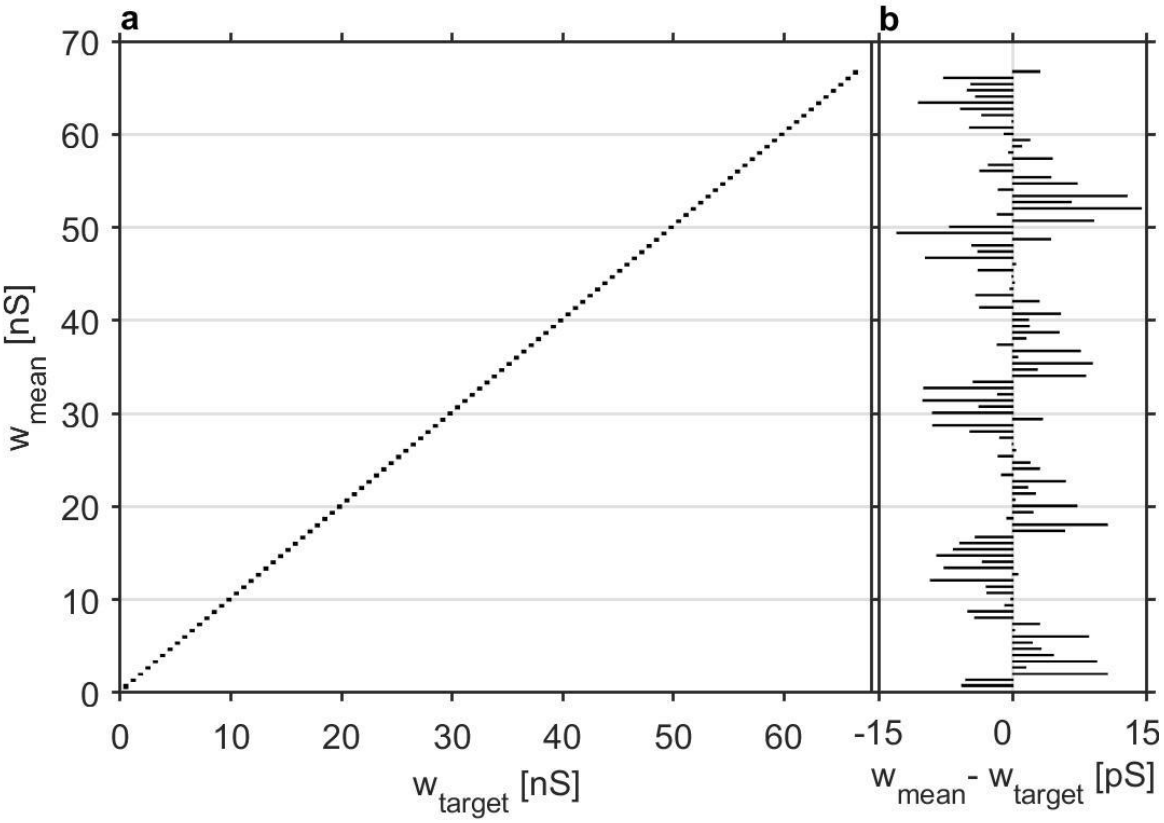

Figure 22: (a) The analog tunability of 300 synstors on a chip is shown. The synstors are tuned to 100 target conductances, shown on the x-axis, with the center of each marker and the height of each marker on the y-axis indicating the mean of each distribution and three standard deviations respectively. (b) The error between the mean of each population and the corresponding target value is shown along the x-axis, while the height of each marker is three standard deviations of the population.

3.5. Schottky Junction

The Schottky junction between the input electrode and channel, and the output electrode and channel, is an essential feature which gives the synstor its nonlinear analog tuning property.

When a single input or output pulse is applied on an electrode to read the conductance, the voltage drops primarily across the contact, and the channel potential remains unchanged relative to the reference electrode. The Schottky junction in the synstor is formed between the single crystal silicon and an amorphous titanium silicide layer. Devices with ohmic contacts, such as synstors prior to the forming gas anneal to form titanium silicide, exhibit poor non-linear tuning.

Titanium disilicide, $TiSi_2$, is a common silicide for CMOS contacts in the semiconductor industry owing to its thermal stability and low resistivity[42]. Titanium silicide has two phases commonly used in industry. When annealing a contact between Ti and a crystalline silicon substrate between 400-500°C, an amorphous $TiSi_2$ layer is formed. Further annealing at 400-500°C results in the C49-$TiSi_2$ phase (60-70μΩcm), and annealing above 700°C results in the C54-$TiSi_2$ phase (15-20μΩcm). This work uses an annealing temperature of 460°C, in a forming gas environment (4% H_2 in N_2), in order to use the band offset of the amorphous TiSi phase to Si to form a Schottky junction. This is unlike the ohmic contacts used in CMOS, where $TiSi_2$ are typically in contact with degenerately doped silicon.

The cross-section of the titanium-silicon interface after forming gas annealing was analyzed by transmission electron microscope (TEM), shown in Figure 22. The diffraction patterns in Figure 23 are produced by selected area electron diffraction (SAED), a companion technique to TEM which indicates the degree of crystallinity. The diffraction pattern shows a polycrystalline Ti layer, an amorphous TiSi interlayer, and a single crystal Si layer.

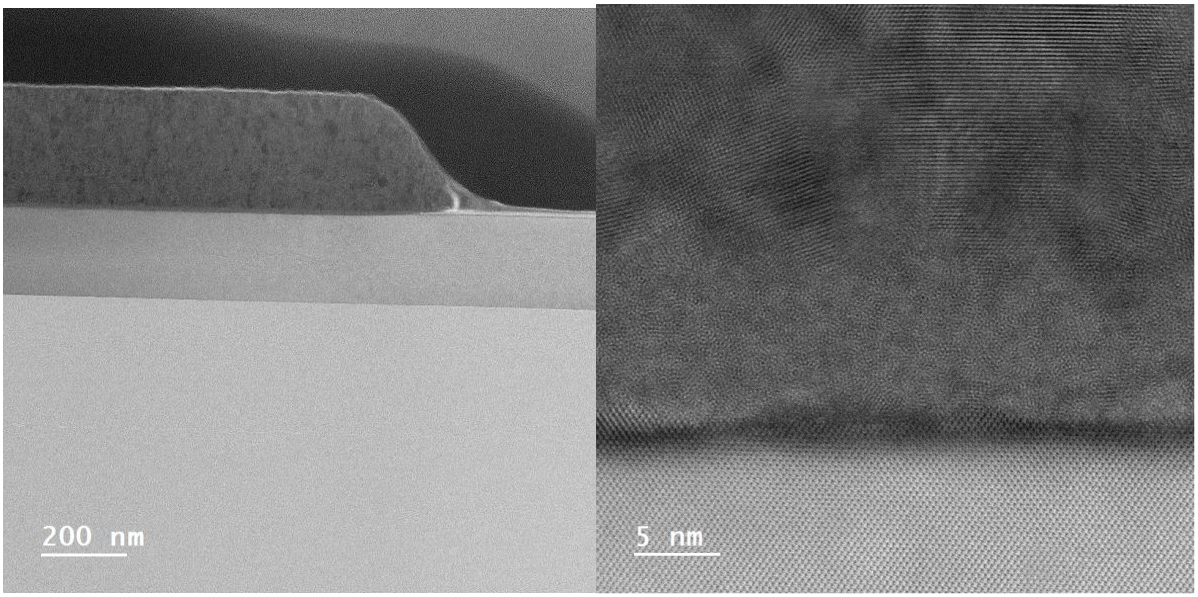

Figure 23: (Left) Transmission electron microscope image (TEM) image showing the cross-section of the input electrode and channel of the synstor. Titanium metal and single crystal silicon are separated by an interface layer of amorphous titanium silicide. (Right) A magnified image of the interface showing the crystallinity of all three layers.

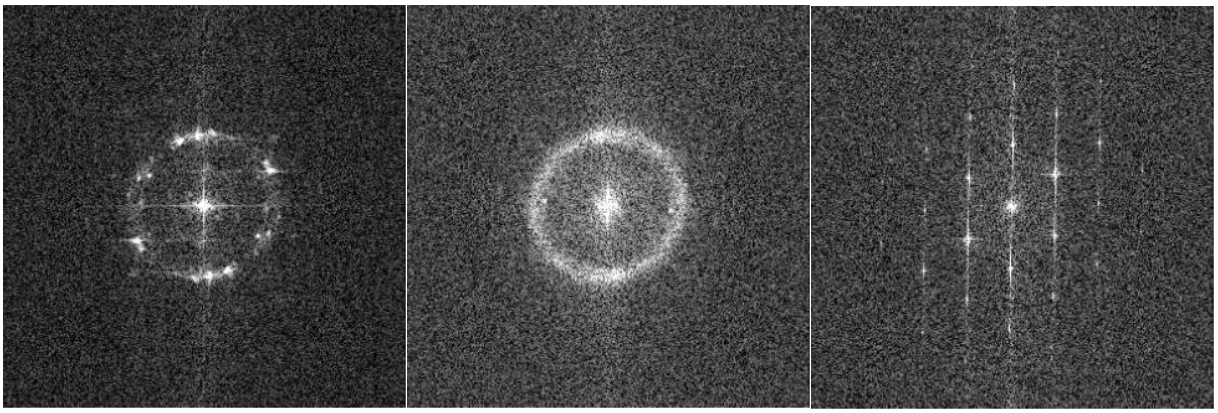

Figure 24: The diffraction pattern of (left) a polycrystalline titanium layer, (middle) amorphous titanium silicide layer, and (right) single crystal silicon layer. The images are produced by selected area electron diffraction (SAED)

Energy dispersive x-ray spectroscopy (EDX) was performed in conjunction with the TEM analysis to obtain a chemical composition profile of the images taken. The ratio of x-rays gathered from Ti and Si atoms in the sample can be used to estimate the stoichiometry of the TiSi film. a TEM image and corresponding EDX depth profiles for Ti, Si, and O are shown in Figure 24 below. The colors of the EDX maps correspond to the device layer colors used in Figure 5. The depth profiles are calculated based on the ratio of elements, by atomic weight, averaged across each horizontal cross section of the image. Based on this analysis, the amorphous interlayer layer is a-$TiSi_z$ with z=0.62. It has been reported in the literature that amorphous TiSi, despite not having a fixed stoichiometry, forms a Schottky contact with p-type crystal Si with a barrier height 0.57-0.59eV[43].

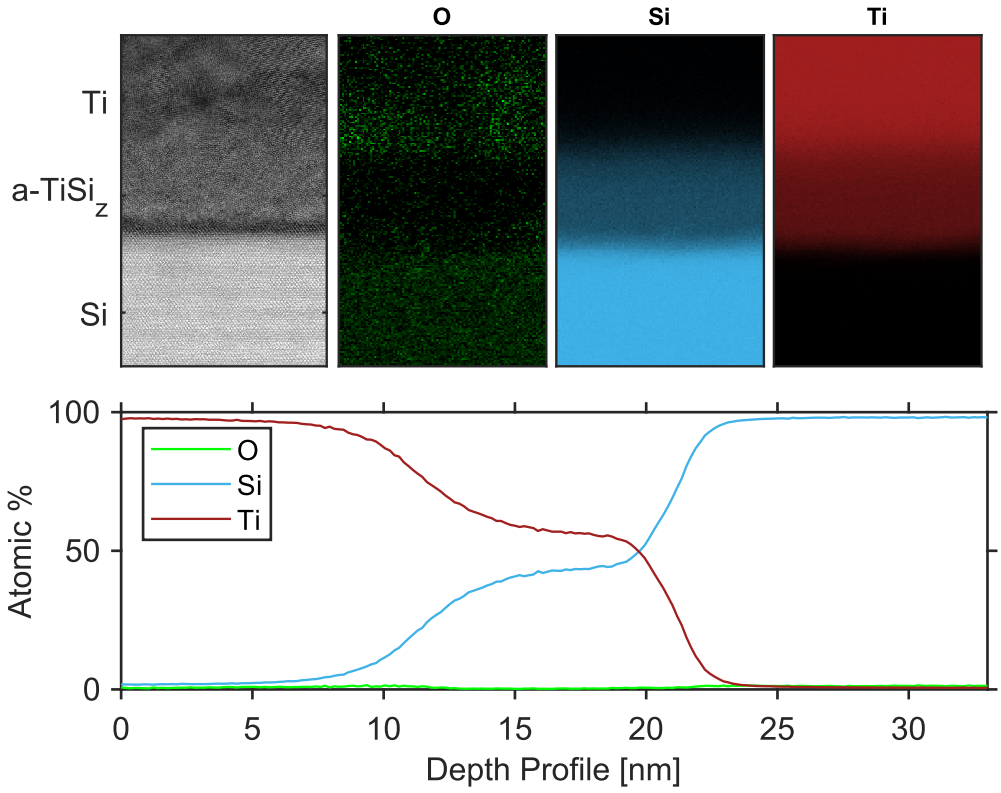

Figure 25: A TEM image of the interface between the Ti input electrode and Si channel (top-left), corresponding EDX maps for O, Si, and Ti (top-right), and corresponding depth profile by EDX analysis (bottom). The value of z in a-TiSi$_z$ is 0.62, based on the depth profile analysis.

3.6. EDX analysis

Elemental analysis by EDX was performed on the memory structure to gain information on the defect distribution, especially in the Al_2O_x layer. The ratio of x-rays gathered from Al and O atoms in the sample was used to estimate the stoichiometry of the Al_2O_x film. a TEM image and corresponding EDX depth profiles for O, Si, Al, and Ta are shown in Figure 25 below. The colors of the EDX maps correspond to the device layer colors used in Figure 5.

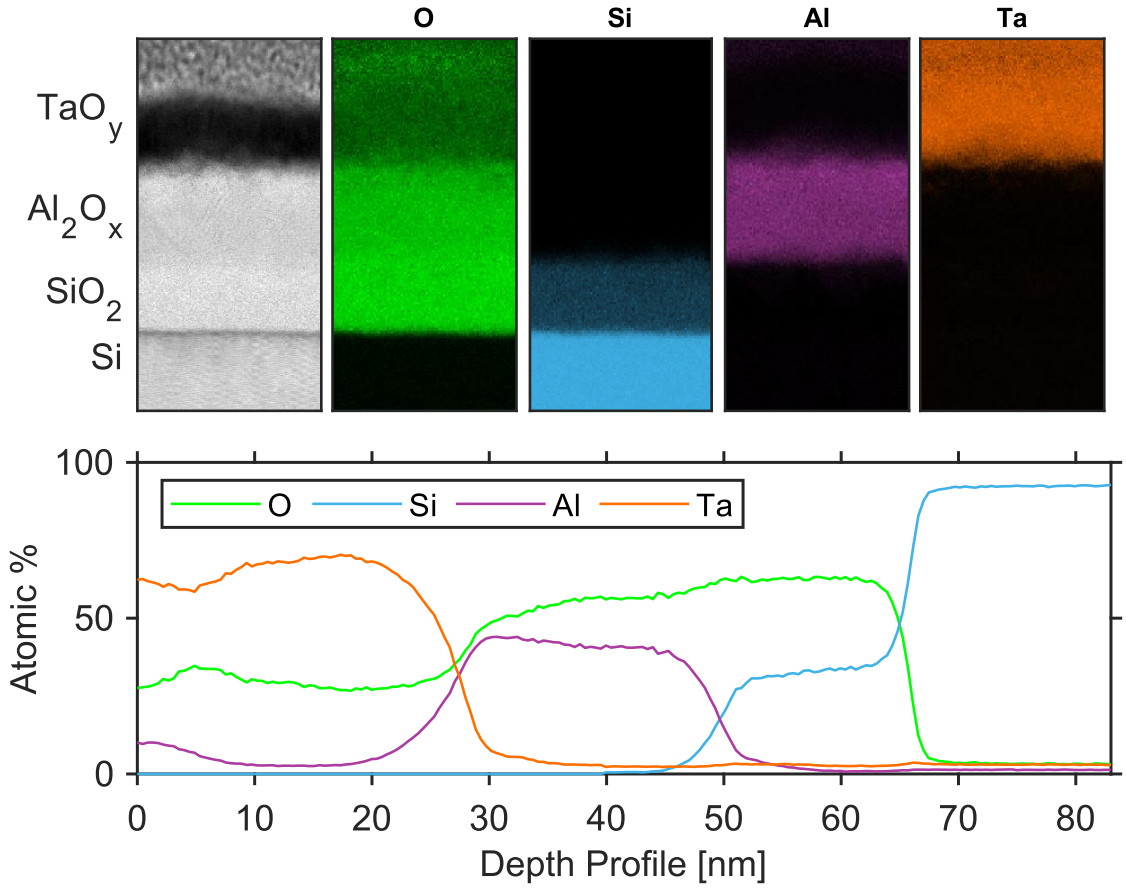

Figure 26: A TEM image (top-left) of the memory stack, corresponding EDX maps for O, Si, Al, and Ta (top-right), and corresponding depth profile by EDX analysis (bottom).

It is reported in the literature that there is an overlap between between Si kα (1.739keV) and Ta M (1.709keV) peak energies. This is a common issue for analysis of semiconductor devices containing refractory metals. For the range of low energies preferred for EDX, only M-line x-rays from metals such as Ta and W are generated, and these have a significant overlap with the Si-K x-rays. For the depth profile in Figure 26, and the calculation of the TaO_y stoichiometry, it is assumed that the peak at ~1.71-1.74eV is entirely attributable to Ta. While it

is difficult to analytically separate the EDX signals from the two elementals over this energy range, it is likely that there is only trace Si presence in this region of the device. X-ray photoelectron spectroscopy (XPS) was done in parallel to this analysis, which showed only trace presence of Si below the SiO_2 layer. The accompanying XPS analysis is also presented in this dissertation.

The oxygen vacancy profile in the Al_2O_x layer was analyzed based on the chemical profile obtained by EDX. The Al_2O_x stoichiometry is presented in the form Al_2O_{3-x}, where x = 0.3. The stoichiometry was estimated by taking the average ratio of the Al to O signal, by atomic weight, over the entire layer. This analysis shows the presence of O vacancies relative to stroichiometric Al_2O_3. These O vacancies are the result of the deposition method, e-beam evaporation, which inherently evaporates Al and O at different vapor pressures, and the redox reaction between the as-deposited AlO_x and Al switching layers. The O vacancies are the primary contributor to the switching behavior of the synstor. When a voltage drops between the reference electrode and semiconducting Si channel, the concentration and distribution of charged O vacancies will shift to reprogram the conductance (synaptic weight), and said distribution remains non-volatile when the voltage drop is removed and the device is operated in its signal processing mode.

Similarly, the depth profile analysis gives a value of y = 0.45 for TaO_y. This oxide is known to have no band gap at room temperature[47], thus its conductivity is suitable as a reference electrode.

3.7. XPS analysis

A standard sample of unpatterned $Si/SiO_2/Al_2O_x/Al/Ta$ was pre-processed under the same conditions as the regular synstor fabrication, and analyzed under x-ray photoelectron

spectroscopy (XPS) to study its chemical composition. 16 sputter cycles of 30 seconds each were used to etch through the device layers for each measurement. The sputter cycles used Ar(2000)+ clusters at 20 kV, and each sputter cycle etched through a thin layer of the device stack. Peaks were fitted by Gaussian-Lorentzian profiles after a Shirley background correction. The survey spectra from the first 12 cycles are shown in Figure 1 below. The final 4 cycles are not shown because they contain only the single crystal silicon. Peaks are identified from O, Al, Ta, and Si.

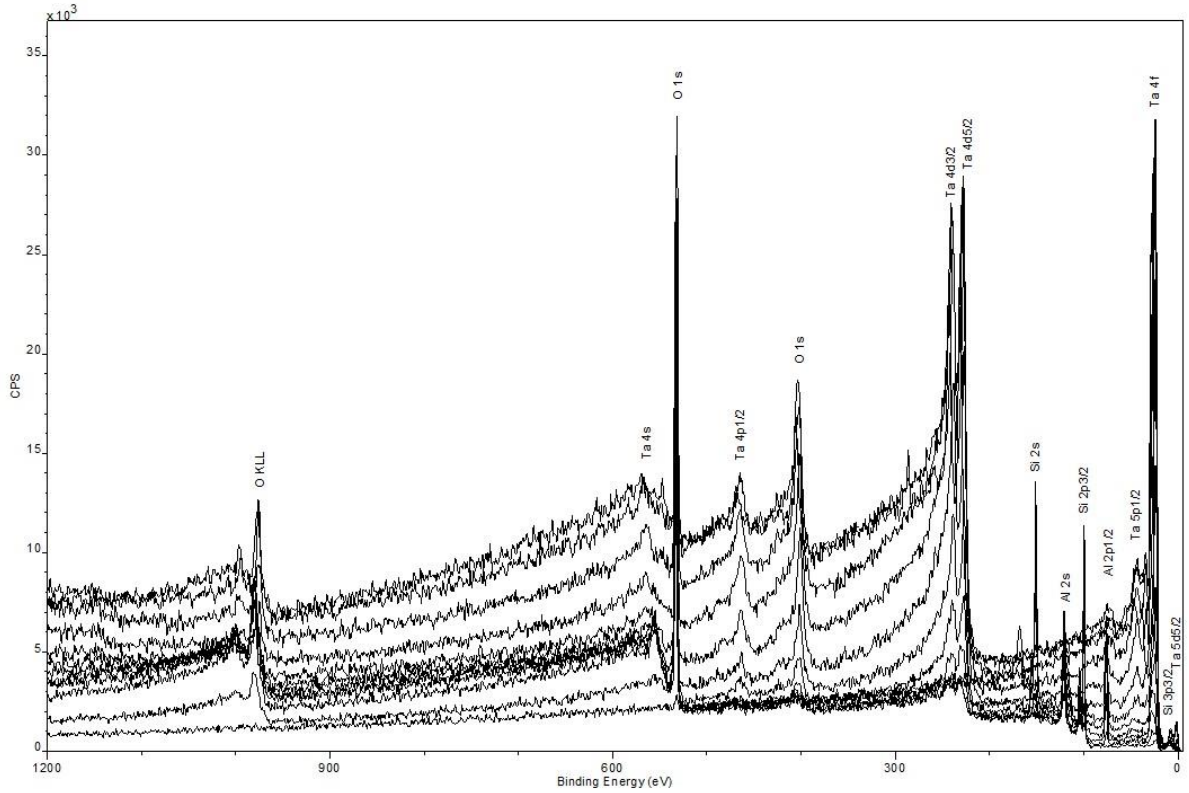

Figure 27: X-ray photoelectron spectroscopy (XPS) survey spectra for the top layers of the synstor, corresponding to the AlOx/Al/Ta structure. Peaks for Ti, Si, Al, and Ta are identified.

The elemental composition of each layer is shown in Figure 1 below. Cycles 12-16 correspond to the silicon channel. One important observation is that the Ta signal appears throughout the device, although EDX imaging from TEM confirms that it exists only on the top of the memory stack. This artifact occurs because the Ta atom is very heavy and will be implanted into the device rather than sputtered after each sputter cycle.

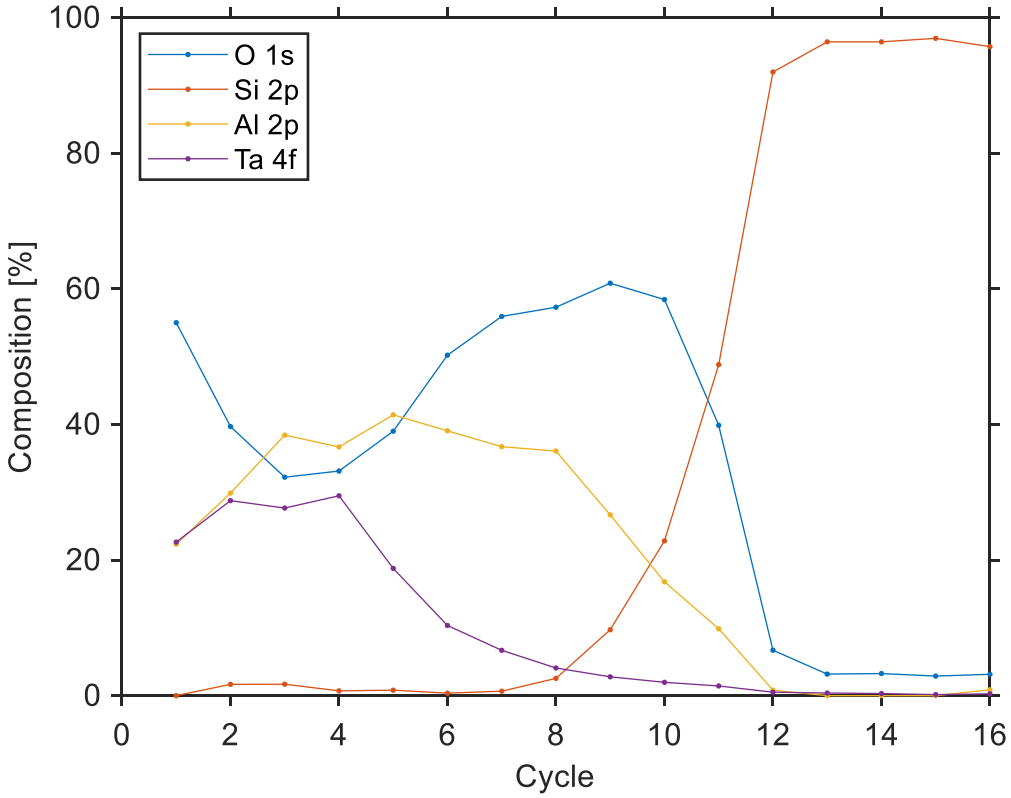

Figure 28: Elemental distribution of the memory structure by cycle number.

Analysis of individual elemental spectra provides an indication of the defect distribution in the memory structure. The binding energy of O 1s peaks in the switching layer were deconvoluted into three major peaks[35,43,44]. The binding energy 531.1±0.2 eV is attributed to the

constituent O species in the amorphous Al_2O_x matrix, ie. O bonded to cations in a stoichiometric ratio. The binding energy 532.5±0.5eV is associated with defective oxides and O vacancies (ie. the photoelectron is emitted from a neighboring O atom with a shifted binding energy), and the binding energy 533.6±0.5eV is attributed to M-OH and H_2O species in the Al_2O_x film. A quantitative analysis of the defects contributing to the memory function based on XPS alone is difficult. This analysis has several limitations, the most important of which is that the sputtering of each layer may itself contribute to the defect density in the oxide layer. Secondly, Ta is poorly sputtered and may in fact be implanted deeper into the oxide by the Ar cluster, and it is challenging to decouple the defective memory oxide from the artificial contribution of the various Ta suboxides found in every subsurface layer. However, in Figure 1 below, showing the O 1s spectra from the first 6 cycles, a qualitative trend can be observed. The fitted peak corresponding to stoichiometric oxide is dominant near the Ta surface. Approaching the SiO_2/Al_2O_x interface (cycle 6), however, the stoichiometric O peak shrinks, and the contribution from the defective oxide increases substantially. It can be inferred that there is a large defect density in the Al_2O_x film near the SiO_2 interface, which may modify the silicon channel carrier concentration via static field.

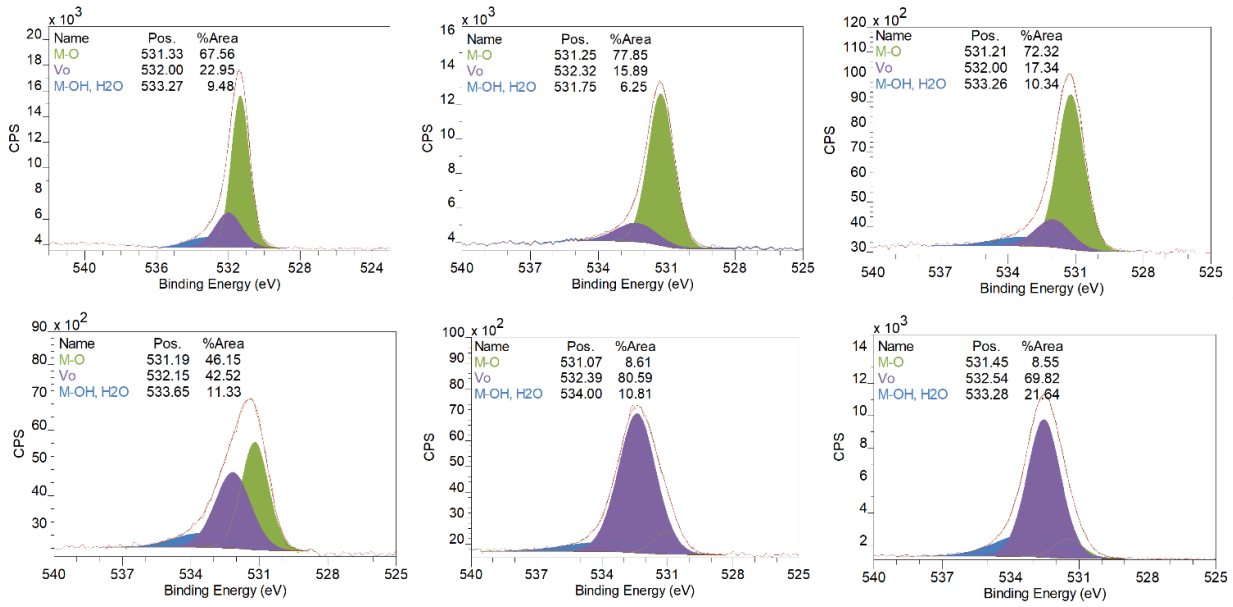

Figure 29: O 1s spectra for the 1st through 6th cycles of the XPS analysis, corresponding roughly to the Al_2O_x/TaO_y sublayer. The green fitted peak corresponds to O in the oxide matrix, while the purple and blue fitted peaks are attributed to defective oxides, and hydroxides or water in the film, respectively.

Al 2p spectra in the 4th through 7th sublayers, corresponding roughly to the Al_2O_x layers of the memory structure, are shown in Figure 2 below. The Al 2p peak at 72.5±0.3eV corresponding to metallic Al matches well with literature values[45]. From EDX, the metallic Al film deposited by e-beam evaporation is oxidized. Thus, cycles 4 and 5 can be viewed as a "metal-rich" oxide, which may also be an indication of a large density of vacancies. The 2nd major fitted peak at 75.6±0.2eV is more difficult to analyze practically, because the binding

energy for all different suboxides of Al are close to overlapping[45].

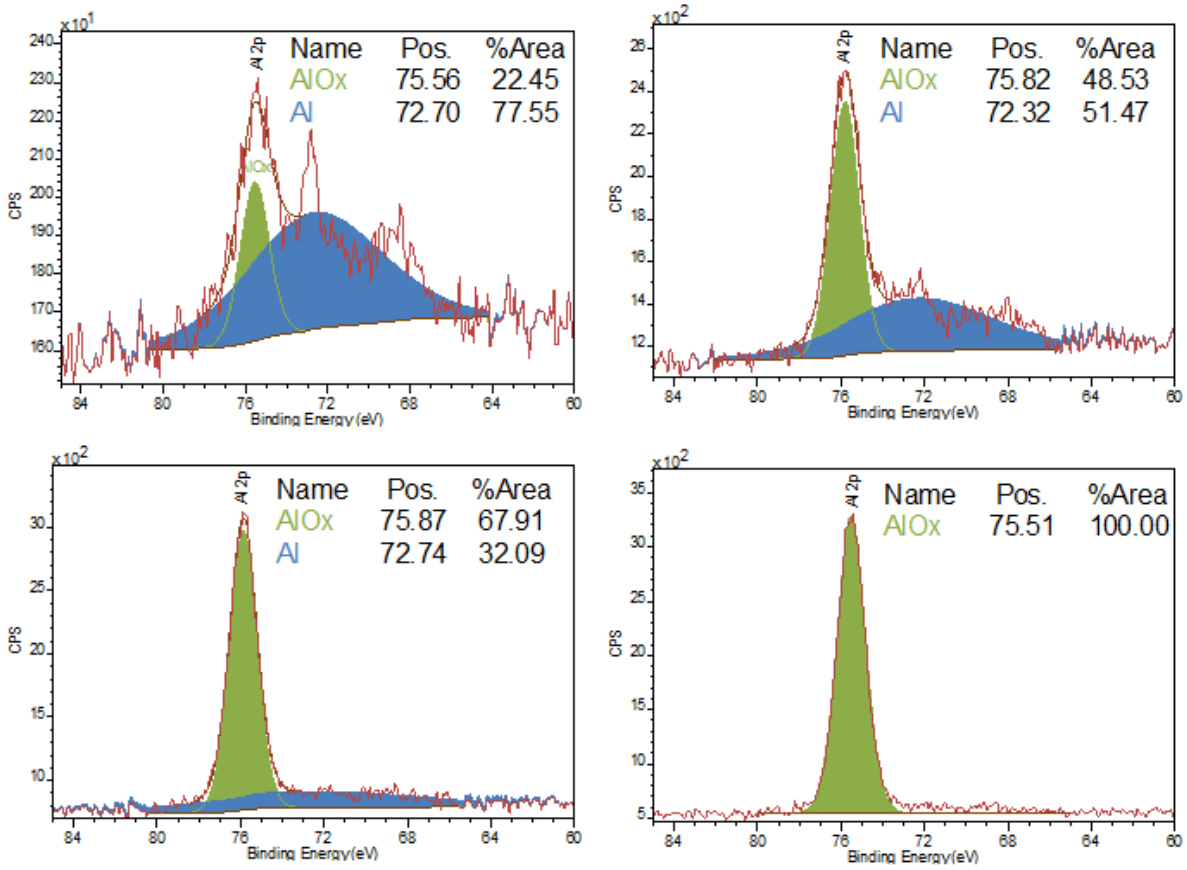

Figure 30: Al 2p spectra for the 3rd through 6th cycles of the XPS analysis, corresponding roughly to the $Al2O_x$ sublayer of the device. The blue fitted peak corresponds to metallic Al, while the green fitted peak corresponds to all Al suboxides present in the film.

Given the variation in the literature and experimental error, it is difficult to conclude which suboxides of Al are dominant in this device based on XPS analysis alone.

3.8. Control devices

To demonstrate the validity of the proposed mechanism, control devices with $Si/SiO_2/Al$ and $Si/SiO_2/Al_2O_3/Pd$ device stacks were fabricated. The purpose of the control devices were to study the memory effect of Al and Al_2O_3 by themselves, without the possibility of a redox reaction. Pd was selected in the second control device as an inert reference electrode metal.

The synstors were fabricated on a similar p-type SOI wafer. The photolithography, channel etching, RCA cleaning, and thermal oxidation processes were equivalent to those described in Methods. Prior to the Ti electrode deposition, the thermal oxides underneath the developed photolithography patterns were etched using buffered hydrofluoric acid (Buffered Oxide Etch, BOE 6:1). After the 300 nm Ti electrodes were deposited, the photoresist was stripped by NMP, and the wafers were treated by a short oxygen plasma to descum the surface (Technics FRIE, 100 mtorr O_2, 50 W). The wafers were then annealed in forming gas (5% hydrogen in N_2) at 460°C for 30 minutes. For the first control device, the Al reference electrode was patterned by photolithography, and liftoff in an acetone and isopropanol solution. For the second control device, a 10 nm layer of Al2O3 was deposited by atomic layer deposition (ALD), and then the Pd reference electrode was patterned by photolithography and liftoff. The Pd reference electrode was then used as an etching mask to etch the Al2O3 into a self-aligned pattern underneath the reference electrode.

The discrepancy in memory properties between the synstor with AlO_x/Al memory stack is highlighted by comparing Figure 29 and 30 below with Figure 15. The devices were tested using triangular voltage sweeps on the input or reference electrodes, in an analogous method to the process described in 3.1. The non-linear rectifying $I\text{-}V_i$ curves in both control devices indicate the formation of Schottky contacts, since the process for the input and output electrodes are

identical. However, the reference electrode sweeps show a distinct lack of hysteresis compared to the real device, since the control devices contain only Al_2O_3 and Al respectively. Without the possibility of the previously described switching mechanism, the control devices simply behave like field-effect transistors with Schottky junctions.

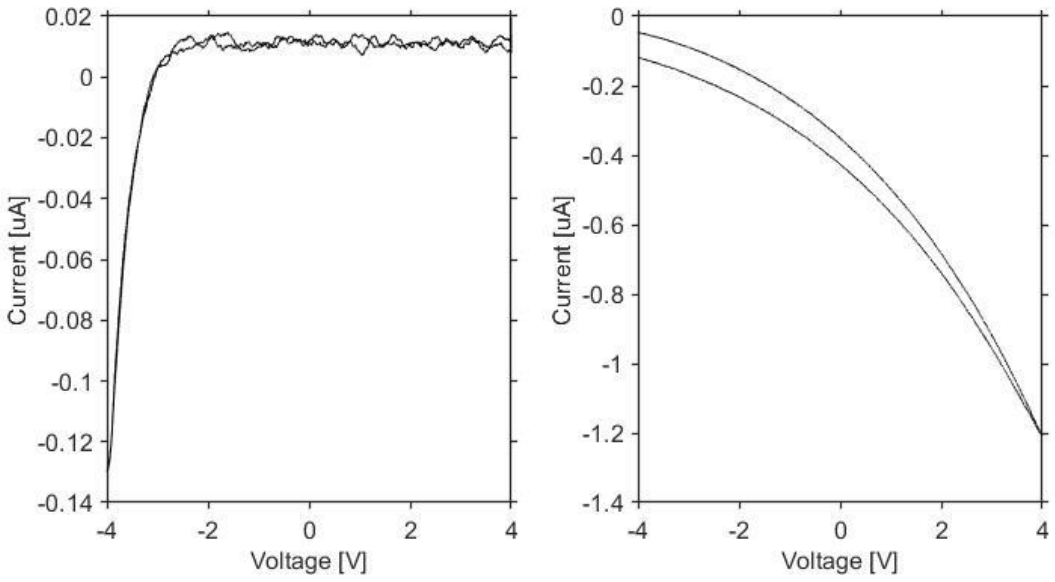

Figure 31: I-V_i (left) and I-V_{ref} (right) measurements obtained by applying a continuous triangular voltage sweep on a Si/SiO_2/Al control device.

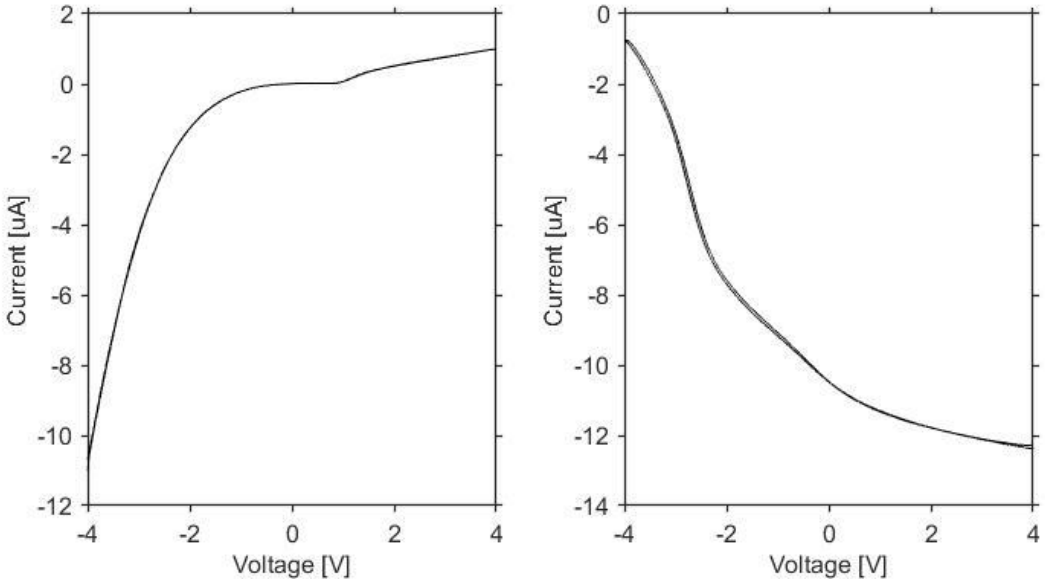

Figure 32 I-V_i (left) and I-V_{ref} (right) measurements obtained by applying a continuous triangular voltage sweep on a $Si/SiO_2/Al_2O_3$/Pd control device.

4. Conclusions and Recommendations

The synaptic resistor developed in this work integrates signal processing, memory, and learning functions into a single element, similar to a biological synapse. A Schottky contact between Si and a-$TiSi_z$ was used to control the voltage drop across the channel, and thus the voltage drop between the channel and reference electrode. In this design, the synstor can operate in two distinct modes: a learning mode (following Equation 2), where the input and output electrodes are both raised to an equivalent potential, and signal processing mode (following Equation 1), where only a single electrode is raised to that potential. Switching was achieved using an oxygen deficient Al_2O_x switching layer with x=2.7. The presence of a high defect density was inferred based on hysteresis in current-voltage measurements, and later confirmed

analytically. XPS analysis indicated the presence of a metal-rich (ie. O vacancy rich) oxide layer. These vacancies constitute the basis of the memory mechanism of the device. When the device is operating in its learning mode, a voltage drops between the reference electrode and semiconducting Si channel, such that the concentration of charged O vacancies will shift to reprogram the conductance (synaptic weight), and the charges remain non-volatile when the voltage drop is removed and the device is later operated in its signal processing mode. The synstor circuit was shown to have improved uniformity over previously the published synstor materials, demonstrating a network of 300 devices with 100 discrete analog conductance states between 0 and 100 nS. The error between the targeted conductance and the mean of each population was less than 15pS. The relative standard deviation of the Si synstor was 29 times larger than the previously reported CNT synstor.

A synstor crossbar circuit emulates neurobiological networks by executing inference and learning algorithms concurrently with the ultra-high energy efficiency, and circumvents the fundamental limitations of energy consumptions in existing electronic circuits such as physically separated logic and memory units, data transmission between memory and logic, the execution of the inference and learning algorithms in serial mode in different circuits, and the signal transmissions between the inference and learning circuits. Like human brains, the circuit is operated in analog parallel mode.

Transistor-based computing circuits can be scaled up, and their computing speeds can approach the speed of the human brain for offline learning, but their power consumption also escalates to ~10^4–10^8 W[21], thus preventing onsite real-time learning. The energy efficiency of a synstor circuit increases with the numbers of synstors connected in parallel with each neuron, and could be further improved by scaling up the crossbar array[21]. With its high energy efficiency,

it is scaled-up synstor could overcome the power hurdle of real-time learning in AI systems, be embedded in mobile robotic systems and powered by batteries.

As an analog device, synstor circuits cannot execute algorithms with the accuracy of digital computers, however, the synaptic conductance matrix $\boldsymbol{w}$ can be dynamically optimized in the real-time learning process, improving the accuracy of the analog inference algorithm, leading to the superior performance and adaptability of AI systems for broad applications in changing environments.

In summary, by circumventing the fundamental limitations of computers, synstors emulate a neurobiological network able to concurrently execute inference ($I^n(t) = \sum_m \kappa^{nm} * (w^{nm} V_i^m)$) and learning ($\frac{dw^{nm}}{dt} = \alpha V_i^m V_o^n$) algorithms in real-time. The synstor conductance matrix $\boldsymbol{w}$ does not need to be pre-programmed, and can be modified toward its equilibrium matrix $\boldsymbol{w}$ to spontaneously optimize the system performance.

While only individual devices have been analyzed in this dissertation, synstor circuits could potentially be used to overcome the "curse of dimensionality" and "Von Neumann bottleneck" issues which currently hinder progress in the field of artificial neural networks. Large-scale synstor circuits could be used to enable AI robotic systems with high energy efficiency, and real-time adaptability in complex environments. There is "plenty of room at the bottom" to miniaturize synstor size, scale up synstor circuits, optimize their materials and fabrication processes, improve their energy efficiency, speed, power consumption, and uniformity for concurrent inference and learning from "big data" in intelligent systems.

5. References

1. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
2. Park, D. S. *et al.* SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv [eess.AS]* (2019).
3. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
4. James, C. D. *et al.* A historical survey of algorithms and hardware architectures for neural-inspired and neuromorphic computing applications. *Biologically Inspired Cognitive Architectures* **19**, 49–64 (2017).
5. Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology*. (Penguin, 2005).
6. Turing, A. M. Mind. *Mind* **59**, 433–460 (1950).
7. Mutlu, O., Ghose, S., Gómez-Luna, J. & Ausavarungnirun, R. Processing data where it makes sense: Enabling in-memory computation. *Microprocess. Microsyst.* **67**, 28–41 (2019).
8. Nakao, M., Ueno, K., Fujisawa, K., Kodama, Y. & Sato, M. Performance Evaluation of Supercomputer Fugaku using Breadth-First Search Benchmark in Graph500. *2020 IEEE International Conference on Cluster Computing (CLUSTER)* (2020) doi:10.1109/cluster49012.2020.00053.
9. Zhao, H., Wang, Q., Wang, J., Wan, B. & Li, S. VM Performance Maximization and PM Load Balancing Virtual Machine Placement in Cloud. in *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)* 857–864 (2020).
10. Shawahna, A., Sait, S. M. & El-Maleh, A. FPGA-Based Accelerators of Deep Learning Networks for Learning and Classification: A Review. *IEEE Access* **7**, 7823–7859 (undefined

2019).

11. Merolla, P. A. *et al.* Artificial brains. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).
12. Davies, M. *et al.* Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro* vol. 38 82–99 (2018).
13. Pei, J. *et al.* Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature* **572**, 106–111 (2019).
14. Prezioso, M. *et al.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61–64 (2015).
15. Li, C. *et al.* Analogue signal and image processing with large memristor crossbars. *Nature Electronics* **1**, 52–59 (2017).
16. Hu, M. *et al.* Memristor-based analog computation and neural network classification with a dot product engine. *Adv. Mater.* **30**, (2018).
17. Ambrogio, S. *et al.* Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **558**, 60–67 (2018).
18. Eryilmaz, S. B. *et al.* Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array. *Front. Neurosci.* **8**, 205 (2014).
19. Koomey, J., Berard, S., Sanchez, M. & Wong, H. Implications of Historical Trends in the Electrical Efficiency of Computing. *IEEE Ann. Hist. Comput.* **33**, 46–54 (2011).
20. Danesh, C. D. *et al.* Synaptic Resistors for Concurrent Inference and Learning with High Energy Efficiency. *Adv. Mater.* **31**, e1808032 (2019).
21. Shaffer, C. M. *et al.* Self-programming synaptic resistor circuit for intelligent systems. *Advanced Intelligent Systems* 2100016 (2021).

22. Dan, Y. & Poo, M.-M. Spike timing-dependent plasticity of neural circuits. *Neuron* **44**, 23–30 (2004).

23. Hebb, D. O. *The organization of behavior: A neuropsychological theory*. (Psychology Press, 2005).

24. Xu, X. *et al.* Scaling for edge inference of deep neural networks. *Nature Electronics* **1**, 216–222 (2018).

25. Diorio, C., Hasler, P., Minch, A. & Mead, C. A. A single-transistor silicon synapse. *IEEE Trans. Electron Devices* **43**, 1972–1980 (1996).

26. Javey, A. & Kong, J. *Carbon Nanotube Electronics*. (Springer Science & Business Media, 2009).

27. Janke, D. & Anderson, D. V. Analyzing the Effects of Noise and Variation on the Accuracy of Analog Neural Networks. in *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)* 150–153 (2020).

28. Bellman, R. Dynamic programming. *Science* **153**, 34–37 (1966).

29. Endsley, M. R. Situation Awareness in Future Autonomous Vehicles: Beware of the Unexpected. in *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* 303–309 (Springer International Publishing, 2019).

30. Strickland, E. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum* **56**, 24–31 (2019).

31. Tans, S. J., Verschueren, A. R. M. & Dekker, C. Room-temperature transistor based on a single carbon nanotube. *Nature* **393**, 49–52 (1998).

32. Chen, B. *et al.* Highly Uniform Carbon Nanotube Field-Effect Transistors and Medium Scale Integrated Circuits. *Nano Lett.* **16**, 5120–5128 (2016).

33. Muckley, E. S., Nelson, A. J., Jacobs, C. B. & Ivanov, I. N. Multimodal probing of oxygen and water interaction with metallic and semiconducting carbon nanotube networks under ultraviolet irradiation. *JPEN J. Parenter. Enteral Nutr.* **6**, 025506 (2016).

34. Svensson, J. & Campbell, E. E. B. Schottky barriers in carbon nanotube-metal contacts. *J. Appl. Phys.* **110**, 111101 (2011).

35. Bolat, S. *et al.* Synaptic transistors with aluminum oxide dielectrics enabling full audio frequency range signal processing. *Sci. Rep.* **10**, 16664 (2020).

36. Dicks, O. A., Cottom, J., Shluger, A. L. & Afanas'ev, V. V. The origin of negative charging in amorphous Al2O3 films: the role of native defects. *Nanotechnology* **30**, 205201 (2019).

37. Ágoston, P., Erhart, P., Klein, A. & Albe, K. Geometry, electronic structure and thermodynamic stability of intrinsic point defects in indium oxide. *J. Phys. Condens. Matter* **21**, 455801 (2009).

38. Erhart, P., Klein, A. & Albe, K. First-principles study of the structure and stability of oxygen defects in zinc oxide. *Phys. Rev. B Condens. Matter* **72**, 085213 (2005).

39. Kern, W. The Evolution of Silicon Wafer Cleaning Technology. *J. Electrochem. Soc.* **137**, 1887 (1990).

40. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117**, 500–544 (1952).

41. Sze, S. M., Li, Y. & Ng, K. K. *Physics of Semiconductor Devices*. (John Wiley & Sons, 2021).

42. Clevenger, L. A. *et al.* Formation of a crystalline metal-rich silicide in thin film titanium/silicon reactions. *Thin Solid Films* **289**, 220–226 (1996).

43. Liauh, H. R., et al. "Electrical and microstructural characteristics of Ti contacts on (001)

Si." *Journal of applied physics* 74.4 (1993): 2590-2597.

44. Liu, A. *et al.* Eco-friendly water-induced aluminum oxide dielectrics and their application in a hybrid metal oxide/polymer TFT. *RSC Adv.* **5**, 86606–86613 (2015).
45. Jo, J.-W. *et al.* Highly stable and imperceptible electronics utilizing photoactivated heterogeneous sol-gel metal-oxide dielectrics and semiconductors. *Adv. Mater.* **27**, 1182–1188 (2015).
46. Tago, T., Kataoka, N., Tanaka, H., Kinoshita, K. & Kishida, S. XPS study from a clean surface of Al2O3 single crystals. *Procedia Engineering* **216**, 175–181 (2017).
47. Bondi, Robert J., et al. "Electrical conductivity in oxygen-deficient phases of tantalum pentoxide from first-principles calculations." *Journal of Applied Physics* 114.20 (2013): 203701.