Neuromorphic Computing for Artificial Intelligence Introductory Report

Natan S. Szczepaniak

16th March 2020

Abstract

In this project we plan to manufacture a lateral memristive device characterised by a pinched hysteresis I-V graph using a TMDC with an oxygen ion migration mechanism. This device will then be used to mimic a fundamental mechanism of synapses in the brain which will then be exploited to realise physical neural network capable of Hebbian learning.A mixture of neuroscience, electrical engineering and condensed matter physics is combined to imitate the efficiency of nature and apply it to improve our contemporary computing methods.

1 Introduction

Artificial Intelligence is an ever-growing field of study due to the potential it holds in revolutionising optimisation methods drawing the interest of large corporations which see it as an effective way of increasing profits. In addition to this, AI is an interesting endeavour in attempting to replicate the function of our human brain, testing the idea of how our brains process information via mimicry of models found in neuroscience. Currently the study of artificial intelligence is predominantly an endeavour in creating artificial neural networks via software utilising traditional hardware, namely computers built using the von Neumann Architecture. [1] The problem with this architecture is the bottleneck posed by the data-bus between the Central Processing Unit (CPU) and the (working) random access memory (RAM). Majority of machine learning methods that exist today rely on a large and low latency memory meaning this bottleneck is more pronounced for these purposes. [2] A potential solution to this is a Neuromorphic computer architecture which utilises parallel/in-memory processing removing the bottleneck entirely. The potential of neuromorphic design is visible from the difference in size and power consumption of current supercomputers and the human brain. Our brains are capable of supercomputer level processing power with the energy consumption comparable to a domestic light bulb. It is true that our brains' processing methods are fundamentally different and more heuristic in nature than computers used today meaning replicating a human brain is a near impossible task. However, applying even a fraction of mechanisms observed in the brain to today's computing methods has the potential for a significant increase in efficiency and speed which makes this field of study worth exploring. Contemporary efforts in creating neuromorphic computer architectures a reality is led by companies such as IBM, Qualcomm, Samsung and Intel. Efforts include creating lower latency non-volatile RAM as well as physical neural networks using memristors. Memristors, first postulated and later created by Leon Chua in 1971 [3], lay at the heart of most neuromorphic designs. The existence of the

memristor was derived from the 6 possible relations between voltage, current, charge and flux of a device. A memristor is characterised by a constitutive relation between the charge q and flux ϕ. In this case, q and ϕ are derived mathematically and do not have a physical interpretation. [4] Memristors can be used as two terminal switches which change state based on the direction of current applied past a certain threshold. In a memristor, there is a low resistance and a high resistance state which correspond to an ON and OFF state. Once the resistive state is set, it does not need power to remain in this state unlike a traditional transistor-capacitor memory cell. This makes the device non-volatile meaning the "information" is not lost if the power is cut.

Applying memristors to computing allows us to recreate synaptic behaviour where the resistive state is altered by the current being passed through it. Multiple of these devices working in parallel can achieve sophisticated levels of computation and pattern recognition. The current flagship neuromorphic chips in industry include the IBM TrueNorth [5] (256M Synapses @ 73mW) as well as Intel's Loihi Chip [6] (130M Synapses @ 53mW). Unlike in traditional computing in these chips rely on exchanging bursts of electrical signals of different intensities in a similar way to the brain. Memristors can be used for the purpose of artificial intelligence in two ways. Firstly they can be utilised to create a new form of random access memory that is non-volatile and low latency, helping improve the efficiency of artificial neural networks programmed on traditional computers. Secondly, they can be used to create physical neural networks by integrating the memory into the processing part of the machine tackling the previously mentioned von Neumann bottleneck. In this project we will focus on the latter as it requires less memristive devices and is more within the realm of our capabilities given the time constraint. We will attempt to manufacture a device which demonstrates memristive behaviour and explore how we can apply it to mimic synaptic function

2 Theory

2.1 Biological Synapses

The human brain is composed of a combination of neurons connected by synapses. he memories we hold in our brain are for the most part believed to be chemical changes in the space between neurons which allow for spikes of current to flow between them. As different neurons are activated by external or internal stimuli, the connections between them get "stronger" in the form of a increased "synaptic weight" which is analogous to a decreased resistance and therefore current easily running through it. [7] This is a gross oversimplification of the real processes happening in the brain however this principle can be used to create a device mimicking this behaviour for future improvement. In biological systems, the process that changes the strength of these connections is called Spike Time Dependent Plasticity (STDP). In short it changes the weight of the synapse based on the relative timing of the spikes from the neurons connected to it. A common model to describe this behaviour is the PRE and POST neuron with a synapse connecting the two. If the PRE neuron fires before the POST neuron the synaptic weight increases, this is called Long Term Potentiation. Alternatively, if the POST neuron fires before the PRE neuron, the synaptic weight decreases, this is called Long Term Depression. The shorter the time between the two spikes from the two neurons, the higher the change in synaptic weight. This behaviour is seen in Figure 1.

After a post-synaptic current (PSC) in the POST

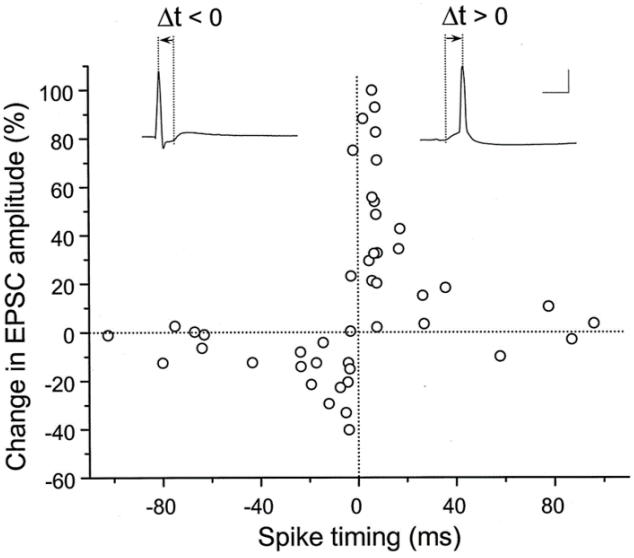

Figure 1: Plot of STDP behaviour in biological synapses in rats from neuroscientific paper. [8] On the vertical axis is the change in EPSC Excitatory Postsynaptic Current and on the horizontal axis is the relative time between the firing of the PRE and the POST neuron. [8]

neuron is created due to an initial firing in the PRE neuron, the synaptic weight restores back to normal, this corresponds to a memory fading away or forgetting in broad terms of speaking. This is why in the case of learning a new skill for example, the action needs to be repeated both on the short time and long time scale. In practice, billions of neurons are firing constantly in our brains at their respective "default" frequencies to maintain the synaptic weight and bursts of spikes between neurons cause the potentiation of the synapse. Naturally, there are different types of synapses that work in slightly different and correspond to different functions in our bodies however most of them follow the basic principle of potentiation and depression. There exist two different types of plasticity (changing synaptic weight) Short-Term Plasticity (STP) and Long-Term Plasticity (LTP). STP constitutes to short term memory which lasts 30 seconds to a minute. LTP can cause lasting and indefinite changes to the synaptic weight

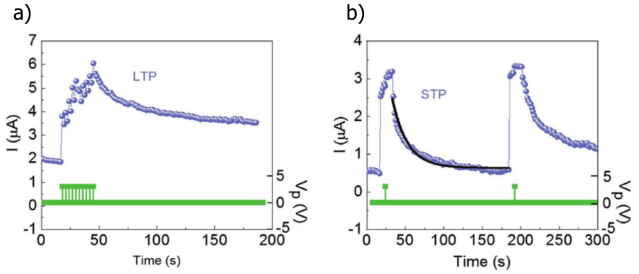

Figure 2: Visualisation of LTP (left) and STP (right) based on input spikes. The potentiation (increased conductivity) can be seen in blue whereas the input spikes are shown in green. LTP caused by successive spikes has a permanent change in resistive state whereas STP returns back to its original state. [9]

as seen in Figure 2.

The difference between these two types of synaptic plasticity is the number of spikes and the timescale they are fired at. The idea that repetition helps build stronger connections in the brain was first postulated by Donald O. Hebb in 1949 [10], "...the persistence or repetition of a reverberatory activity tends to induce lasting cellular changes that add to its stability.". The well known Pavlov's dog experiment can also be explained using this mechanism. This concept is the foundation for Spiking Neural Networks SNN in neural network research today however it is still at its early stages with the first real implementation taking place in September 2017 by BrainChip [11].

Additionally, another type of cell worth mentioning that regulates the synaptic weight is the astrocyte however, we have decided not to focus on this as we agreed it is not needed to demonstrate STDP in our device. The aforementioned mechanisms are very similar to the inner workings of a memristor as it is able to change conductance based on the direction of current running through it making it ideal for this purpose.

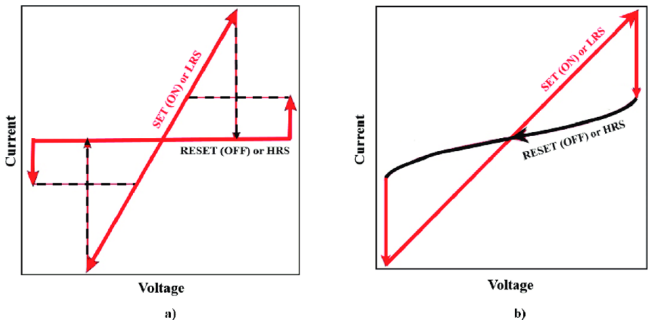

Figure 3: Graph (a) demonstrates a unipolar memristor and (b) shows a bipolar memristor. [12]

2.2 Memristors

The memristor, as mentioned previously, is the fourth fundamental component relating charge and flux, with the other three components being: the resistor, capacitor and inductor. The variables that govern these components are current i, voltage v, charge q and flux ϕ. The relation describing Memristance M is

$$d\phi = Mdq \tag{1}$$

The units of M are ohms same as resistance however, because M has q dependence, it means the component is non-linear. The variables q and ϕ have no physical interpretation as they are merely derived mathematically as mentioned in Leon Chua's Paper. [4]

There exist two types of memristors, unipolar and bipolar. Bipolar memristors switch resistive states based on the direction current applied past a threshold whereas unipolar memristors switch resistive states just as long as the threshold is reached. The switching mechanism of a memristive device (both unipolar and bipolar) can be seen in Figure 3.

A pinched hysteresis loop I-V graph is characteristic of a memristor. As seen from Figure 3, in the case of the more popular and interesting bipolar memristor, when a threshold voltage is applied in the positive direction, the device switches from the high resistance state (HRS) to the low resistance state (LRS). In order to switch the device back to a HRS, a threshold voltage needs to be applied in the opposite direction. The state of the memristor is the information stored. We can see that this is non-volatile memory as the pinched hysteresis loop passes through the origin meaning the information (resistive state) remains when no voltage is applied. Furthermore, the resistive state of the memristor can be read by sending signals below the threshold which lets us establish a read-write with only two terminals.

There has been a plethora of attempts at modelling the pinched hysteresis loop behaviour of the theoretic idea of a memristor. Two main mechanisms currently being researched are Molecular & Ionic thin films as well as Spin and Magentic memristors. For our purposes and capability within the lab, we will be focusing on creating a Ionic thin film memristor using a Transition Metal Dichalcogenide (TMDC) between two electrodes. The TMDC that we are going to be using is HfO_2 as it's memristive properties have been demonstrated in the past and there are people in the department which have experience with it's manufacturing process. [13]

There exist two configurations of Metal-Insulator-Metal MIM devices, vertical stack as well as lateral, the difference between these can be seen from Figure 4 . There are several different mechanisms that can be used to model the memristor behaviour. In our project, we aim to create a bipolar (MIM) using the oxygen ion mechanism present in HfO_x . In a device with HfO_x placed in between Ti and TiN electrodes, the resistive state changes due to the creation of oxygen vacancy "filaments" via the migration of oxygen (Figure 4). As seen from Figure 5, when a positive voltage is applied at the Ti terminal, the oxygen ions migrate towards it creating oxy-

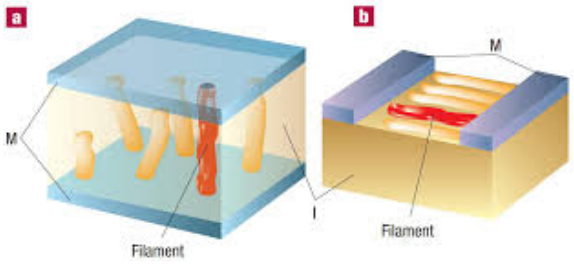

Figure 4: Filamentisation of oxygen vacancies in vertical stack (a) and lateral (b) devices. [15]

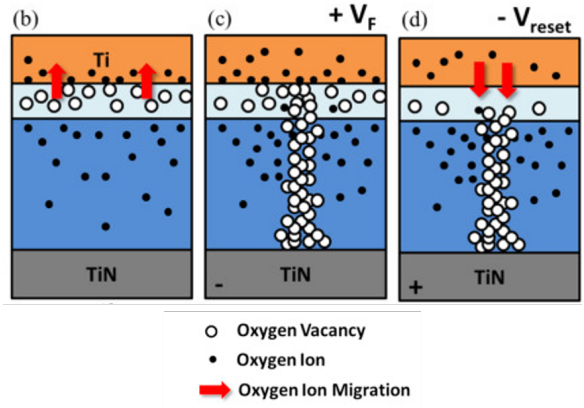

Figure 5: Demonstration of Oxygen Ion migration in a MIM device with Ti and TiN electrodes and the filamentisation of oxygen vacancies. [16]

gen vacancies in the oxide decreasing its resistance as filaments are formed. The Ti electrode works as an oxygen ion reservoir as it absorbs the ions from the oxide layer. [14] The oxygen vacancy filament is conductive allowing electrons to flow in turn reducing the resistance into the LRS. When an opposite voltage is applied, the oxygen ions return to the oxide restricting the flow of current and increasing the resistance to the HRS. It is worth mentioning that the expected IV characteristic of our device is most likely to include some noise as the ions diffuse in the oxide sporadically. The diffusion of oxygen ions can be described by considering driven-fick diffusion and thermophoresis.

Another aspect that is important to mention is the longevity of the device as the transport of ions between the Ti and HfO_x layers causes permanent changes in the material. Previous attempts at creating a similar device have achieved 2 million cycles before the oxygen ions started getting trapped in the Ti electrode giving rise to permanent oxygen vacancy filaments leading to the failure of the device. Improvements can be made, such as making the Ti electrode thicker or blanket capping the device with HfO_x+Al_2O_3 which are possibilities we might explore if our device is prone to failure. [14]

2.3 Memristors as Synapses

Memristors show attributes similar to that of biological synapses in the brain. Based on previous attempts at using memristors as synapses [17,18] we have concluded that the best way of reproducing the behaviour of a synapse is modelling STDP in the device by sending spikes of signal through (like mentioned in section 2.1) in attempt of getting behaviour similar to that shown in Figure 1. The Exact graph characteristics of Figure 1 would be difficult to reproduce exactly with how primitive our device is compared to the complexities of a biological synapse so we have to rely on LTP and LTD as a benchmark. As the device itself will not function as a synapse by itself, it has to be embedded into a circuit which will allow for this behaviour to come through. A good starting point for a circuit which uses memristors to achieve STDP is the 1T1R circuit (1 Transistor 1 Resistor) shown in Figure 6.

Once the STDP characteristic is achieved it can be used for the creation of a physical neural network modelling a spiking neural network for various different architectures and applications.

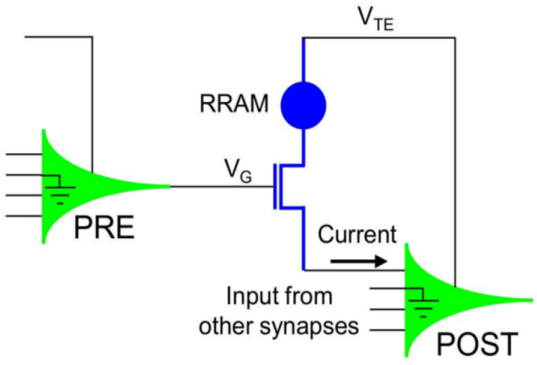

Figure 6: Abstract plan for circuit that will be replicated and adjusted for our purposes.Here, the transistor and memristor (RRAM) coloured in blue are connected between two CMOS neurons. [19]

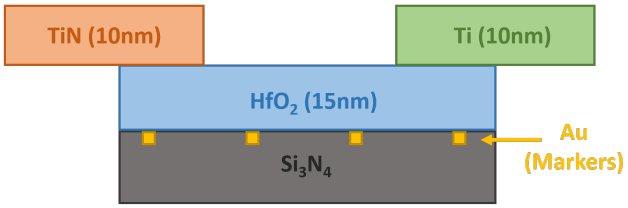

Figure 7: Our proposed lateral HfO_2 device. Au markers on the substrate serve no purpose to the mechanism.

3 Experimental

3.1 Memristor Manufacturing

For our experiment we plan to create a lateral and bipolar memristor using HfO_2 combined with Ti and TiN electrodes for the purposes of recreating STDP characteristics of real biological synapses which are useful for artificial intelligence applications. From our comprehensive research in this rather new field of study, this has not been attempted before using the manufacturing methods we are planning on using. This makes our experiment unique to others. In addition to this, due to the experimental nature of our project we expect to discover limitations of these methods as well as other interesting properties we may decide to investigate. We will also explore the possibilities of creating a stacked memristor depending on the timescale of the project and how much progress is made.

Our proposed device is composed of a flake of HfO_2 exfoliated onto a Si_3N_4 Chip with Au coordinate markers with two electrodes sputtered on top: one Ti and one TiN. Our current planned method begins with using mechanical exfoliation of a HfS_2 crystal using scotch tape onto the Si_3N_4 1x1cm substrate. Next, we would use Atomic Force Microscopy (AFM) to see the distribution of flakes on the chip in order to work out the best positioning for the electrodes. We will then use AutoCAD 2019 pre-loaded with a template of the substrate with calibrated coordinate markings and the AFM images aligned to plan where which flakes we will use and where we put the electrodes. After the electrodes are drawn on in the software, this will be exported to an EBL (Electron Beam Lithography) machine which will expose the HfS_2 ready for DC Sputtering of the electrodes on the surface. Next, we will use selective photooxidation to convert the HfS_2 into the oxide HfO_x. From previous work in our department of photooxidising HfS_2 , we know that the optimal wavelengths of the laser are (UV) λ_{ir}=375nm and (visible) λ_{ir}=264nm with a typical energy density of 53mJ/μm^2 at an exposure of 1 to 2 seconds per point [13]. We will experiment with oxidising different parts of the flakes and plan on testing what effect a layer of unoxidized HfS_2 in the device would have.

3.2 Artificial Synapses

Once the electrodes are connected and the correct parts are oxidised we plan on connecting the device to an AC power supply and seeing the IV character-

istic of it using an oscilloscope in search of the desired pinched hysteresis loop. Once memristive behaviour is established, we will place the device into a 1T1R circuit and investigate further for STDP behaviour described in Section 2.1. Although unlikely, if everything goes well and time restrictions allow us, we will attempt to use this device in a extremely simple physical neural network composing of a small number of these devices. This has been demonstrated before using other types of memristors by producing convolutional neural networks being able to identify handwritten digits from the well known MNIST database. Our implementation would be closer to a simple one-layer perceptron however that is enough to demonstrate Hebbian learning showing how suitable our device is for this purpose on a small scale.

4 Current Progress

Currently, with the help of a postgraduate researcher we have been able to exfoliate the HfS_2 onto the Si_3N_4 base. After using AFM to capture images of the flakes at different magnifications we imported these files into AutoCAD. We stacked these photos on top of each other (in a similar fashion to satellite imagery) to create a full image of the chip with areas of interest having higher resolution for more precise electrode placement.

We selected a small area around the desired flakes where more fine EBL operations can be made shown by a light blue circle in Figure 8. We used separate layers to draw on the thinner and thicker electrodes as this will be passed on to the EBL machine. For the thicker electrodes we can use a faster less precise setting to save on time it takes to draw them on. Our next step in the lab would be to draw on the shown electrode shapes and use DC sputtering to deposit the Ti and TiN on the device.

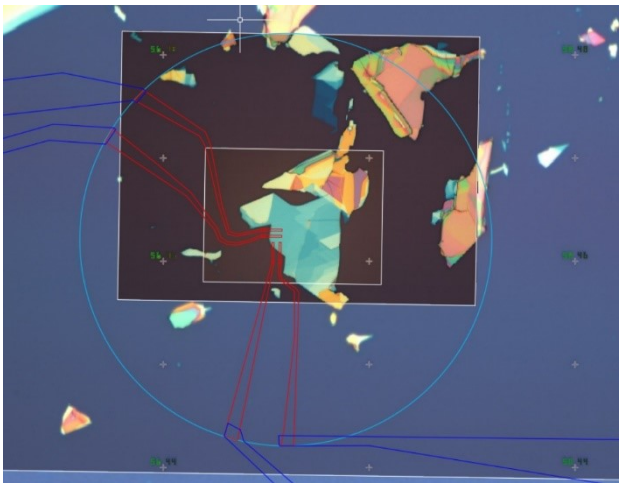

Figure 8: Screenshot of layered images in AutoCAD with electrodes drawn in. Red lines show fine electrodes closer to the flake whereas dark blue lines show shape of thicker electrodes which will be used as input-output.

5 Conclusions

In summary, we plan on creating a reliable memristive device using HfO_2 and test its performance. Next, we will use this device in a 1T1R circuit to model the behaviour of a biological synapse described in Section 2.1 in the form of a STDP Figure 1. Then, if time allows, we will apply these artificial synapses to a simple 1 layer physical neural network to demonstrate Hebbian learning. We are aware that this plan of action is rather optimistic, and we don't expect to complete all of these things as problems and unexpected scenarios are inevitable this is why we have included three possible outcomes that would be a satisfactory outcome of the project. These are: Creating a memristor and testing its performance; Using the Device to recreate STDP; Using the STDP behaviour to produce a physical neural network capable of artificial intelligence tasks.

Glossary

STDP – Spike Time-Dependent Plasticity
EPSC - Excitatory Postsynaptic Current
LTP – Long Term Potentiation
STP – Short Term Potentiation
PSC – Post Synaptic Current
HRS – High Resistive State
LRS – Low Resistive State
TMDC - Transition Metal Dichalcogenide
MIM – Metal Insulator Metal
AFM - Atomic Force Microscopy
EBL – Electron Beam Lithography

References

[1] J. L. H. David A. Patterson, *Computer Organization and Design: The Hardware/Software Interface*. Elsevier Science, 4 ed., 11 2011.

[2] G. Pratl and P. Palensky, "Project ars - the next step towards an intelligent environment," pp. 55– 62, 07 2005.

[3] L. Chua, "Memristor-the missing circuit element," *IEEE Transactions on Circuit Theory*, vol. 18, no. 5, pp. 507–519, 1971.

[4] L. Chua, "Resistance switching memories are memristors," *Applied Physics A*, vol. 102, pp. 765–783, 3 2011.

[5] A. S. Cassidy, J. Sawada, P. Merolla, J. V. Arthur, R. Alvarez-Icaza, F. Akopyan, B. L. Jackson, and D. S. Modha, "Truenorth: A high-performance, low-power neurosynaptic processor for multi-sensory perception, action, and cognition," 2016.

[6] M. Davies, N. Srinivasa, T. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y. Weng, A. Wild, Y. Yang, and H. Wang, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.

[7] J. J. Langille and R. E. Brown, "The synaptic theory of memory: A historical survey and reconciliation of recent opposition," *Frontiers in Systems Neuroscience*, vol. 12, p. 52, 2018.

[8] G.-q. Bi and M.-m. Poo, "Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type," *Journal of Neuroscience*, vol. 18, no. 24, pp. 10464–10472, 1998.

[9] S. Majumdar, H. Tan, Q. Qin, and S. Dijken, "Energy-efficient organic ferroelectric tunnel junction memristors for neuromorphic computing," *Advanced Electronic Materials*, 01 2019.

[10] G. L. Shaw, "Donald hebb: The organization of behavior," in *Brain Theory* (G. Palm and A. Aertsen, eds.), (Berlin, Heidelberg), pp. 231–233, Springer Berlin Heidelberg, 1986.

[11] W. G. Wong, "Electronic design." https://www.electronicdesign.com/technologies/embedded-revolution/article/21805572/brainchip-enters-ai-territory-with-spiking-neural-network, accessed 20 March, 2017.

[12] A. Yesil, F. Gül, and Y. Babacan, *Emulator Circuits and Resistive Switching Parameters of Memristor*, pp. 41–61. 04 2018.

[13] N. Peimyoo, M. Barnes, J. Mehew, A. De Sanctis, I. Amit, J. Escolar, K. Anastasiou, A. Rooney, S. Haigh, S. Russo, M. Craciun, and F. Withers, "Laser-writable high-k dielectric for van der waals nanoelectronics," *Science advances*, vol. 5, p. eaau0906, January 2019.

[14] S. Kumar, Z. Wang, X. Huang, N. Kumari, N. Davila, J. P. Strachan, D. Vine, A. L. D. Kilcoyne, Y. Nishi, and R. S. Williams, "Oxygen migration during resistance switching and failure of hafnium oxide memristors," *Applied Physics Letters*, vol. 110, p. 103503, Mar 2017.

[15] M. A. R. Waser, "Nanoionics-based resistive switching memories," *Nature Materials*, pp. 833–840, 11 2007.

[16] D. R. U. T.Y.Tseng, "Metal oxide resistive switching memory: Materials, properties and switching mechanisms," *Ceramics International*, vol. 43, pp. 547–556, 8 2017.

[17] H.-K. He, R. Yang, H.-M. Huang, F.-F. Yang, Y.-Z. Wu, J. Shaibo, and X. Guo, "Multi-gate memristive synapses realized with the

lateral heterostructure of 2d wse2 and wo3," *Nanoscale*, vol. 12, pp. 380–387, 2020.

[18] L. M. B.Irem, "Neuromorphic computing with multi-memristive synapses," *Nature Communications*, vol. 9, no. 1, p. 2514, 2018.

[19] S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z. Wang, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Neuromorphic learning and recognition with one-transistor-one-resistor synapses and bistable metal oxide rram," *IEEE Transactions on Electron Devices*, vol. 63, no. 4, pp. 1508–1515, 2016.