AFRL-RI-RS-TR-2023-077

MAGNETIC NANOELECTRONICS FOR BRAIN-INSPIRED COMPUTING (MN-BRIC): FROM MATERIALS TO CIRCUIT MODELS

UNIVERSITY OF ILLINOIS

APRIL 2023

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2023-077 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
JOSEPH E. VAN NOSTRAND
Work Unit Manager

/ S /
GREGORY J. HADYNSKI
Assistant Technical Advisor
Computing & Communications Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
APRIL 2023	FINAL TECHNICAL REPORT	JUNE 2021	DECEMBER 2022

4. TITLE AND SUBTITLE

MAGNETIC NANOELECTRONICS FOR BRAIN-INSPIRED COMPUTING (MN-BRIC): FROM MATERIALS TO CIRCUIT MODELS

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
N/A	FA8750-21-1-0002	61102F

5d. PROJECT NUMBER	5e. TASK NUMBER	5f. WORK UNIT NUMBER
		R34C

6. AUTHOR(S)

Dr. Shaloo Rakheja

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Board of Trustees of the University of Illinois Henry
Administration Building 506 S. Wright Street
Urbana IL 61801-3620

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Research Laboratory/RITB 525 Brooks Road Rome NY 13441-4505	AFRL/RI	AFRL-RI-RS-TR-2023-077

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This technical report summarizes the R&D efforts for the AFRL project “Magnetic Nanoelectronics for Brain-Inspired Computing (MN-BRIC): From Materials to Circuit Models”. This effort focused on spintronics technology, wherein they exploited the magnetization dynamics of magnetic materials, most prominently antiferromagnets (AFMs), to implement brain-inspired circuits and architectures. The project focused on developing physics-based models at various levels of the design hierarchy to fully quantify the potential and limits of spintronics for brain-inspired computing. This effort leveraged RPI’s complementary expertise in neuromorphic hardware simulator (NeMo) that can simulate neuromorphic architectures of arbitrary dimensions, allowing for novel architecture performance benchmarking. The major outcomes of the effort include (i) materials and device models of AFM neurons and ferromagnetic (FM) synapses, (ii) circuit-compatible SPICE models of these neurons and synapses, (iii) full architecture level benchmarking, including power, throughput, and area assessment of a neuromorphic architecture leveraging magnetic neurons and synapses and (iv) quantification of the impact of interconnects on architecture performance. The PI's results show that although spintronics hardware offers significant energy and latency advantages, for larger neuromorphic cores, the performance is dominated by interconnection networks. This limitation is overcome by architectural changes to the network or by using new interconnect materials that offer lower resistance and capacitance compared to copper/low-k interconnects, currently used in CMOS chips. Significantly, the work under this grant has also generated new concepts of brain-inspired computing at the thermodynamic limits based on diffusive and stochastic phenomena that can be obtained in magnetic devices.

15. SUBJECT TERMS

Magnetic, antiferromagnetic, device model, neuromorphic, nanoelectronics, thermodynamics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	32

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
JOSEPH E. VAN NOSTRAND	**N/A**

Table of Contents

List of Figures

List of Tables

1. Summary

This technical report summarizes the R&D efforts for the AFRL project "Magnetic Nanoelectronics for Brain-Inspired Computing (MN-BRIC): From Materials to Circuit Models". In this project, we focus on spintronics technology, wherein we exploited the magnetization dynamics of magnetic materials, most prominently antiferromagnets (AFMs), to implement brain-inspired circuits and architectures. The project focused on developing physics-based models at various levels of the design hierarchy to fully quantify the potential and limits of spintronics for brain-inspired computing. This effort leveraged RPI's complementary expertise in neuromorphic hardware simulator (NeMo) that can simulate neuromorphic architectures of arbitrary dimensions, allowing for novel architecture performance benchmarking. The major outcomes of the effort include (i) materials and device models of AFM neurons and ferromagnetic (FM) synapses, (ii) circuit-compatible SPICE models of these neurons and synapses, (iii) full architecture level benchmarking, including power, throughput, and area assessment of a neuromorphic architecture leveraging magnetic neurons and synapses and (iv) quantification of the impact of interconnects on architecture performance. Our results show that although spintronics hardware offers significant energy and latency advantages, for larger neuromorphic cores, the performance is dominated by interconnection networks. This limitation is overcome by architectural changes to the network or by using new interconnect materials that offer lower resistance and capacitance compared to copper/low-k interconnects, currently used in CMOS chips. Our work under this grant has also generated new concepts of brain-inspired computing at the thermodynamic limits based on diffusive and stochastic phenomena that can be obtained in magnetic devices.

2. Introduction

The human brain is widely regarded as the ultimate computing engine with extremely high energy efficiency, reliability, and learning and cognitive capabilities. The aspiration to design computing platforms with attributes approaching those of the brain is universal and long-standing. The area of neuromorphic or brain-inspired computing[1] has its origins in the pioneering work of Carver Mead and his collaborators in the 1980s [1]. Though much progress has been made since then, neuromorphic systems have yet to demonstrate the cognitive functionality of mainstream AI methods (e.g., deep nets) or the energy efficiency approaching that of the brain. A key reason is that there exists a huge mismatch between the properties of mainstream (CMOS) devices vs. those of the brain. To overcome this challenge, in this project, we evaluated the full-stack performance of neuromorphic systems that exploit the neuro-synaptic dynamics in magnetic devices. Unlike CMOS devices, magnetic devices demonstrate dynamics that are similar to those of biological neurons and can be tuned from several 10's of gigahertz to near-terahertz [2]. Moreover, magnetic devices acting as neurons and synapses, the building blocks of neuromorphic systems, possess low area ($< 1\ \mu m^2$) and operate at significantly lower energy ($<10^{-3}$ pico-joules) compared to their counterparts. Therefore, these materials can serve as the computational primitives of a biologically inspired hardware that promises to emulate the highly efficient and low-power cognitive capabilities of the human brain in hardware.

New devices can be classified into two main categories, depending on whether they realize synaptic behavior or the nonlinear transfer function of neuron in hardware. Non-volatile memristors, such as resistive random-access memory (RRAM), phase-change memory (PCM), and ferroelectric random-access memory (FeRAM) function of hardware synapses. Similarly, memristive dynamics is also feasible in magnetic materials, which was first experimentally observed by Kyrysteczko in 2012 [3]. On the other hand, neuron hardware emulators typically utilize silicon transistor, although more recently spin dynamics in magnetic tunnel junctions (MTJs) have also been explored [4]. In the last five years, significant research progress has been reported in antiferromagnetic spintronics [5-7]. Unlike ferromagnets that have been the active elements in all spintronic devices, whether it is for memory applications or neuromorphic / brain-inspired computing applications, antiferromagnets traditionally served a secondary role wherein they were used mainly to create exchange bias[2] in a proximal ferromagnet. However, antiferromagnets have unique magnetization dynamics [8] that make them particularly useful for emulating biological neurons in hardware. For example, antiferromagnets can display coherent oscillations, incoherent oscillations (e.g., spiking, bursting, chirping), and memristive switching.

Meanwhile, ferromagnetic devices have been used for memory and neuromorphic computing; however, in current methods, their brain-inspired spin dynamics are not truly harnessed. For example, most neuromorphic architectures only exploit the memristive dynamics in ferromagnetic devices. But our group and others have shown that the intrinsic noise in ferromagnets can also be harnessed for stochastic resonance and stochastic facilitation [9,10] — phenomena in which noise promotes neural synchronization and information transmission in the brain.

[1] We will use the term "neuromorphic" and "brain-inspired" interchangeably throughout this document.

[2] Exchange bias refers to the pinning of the magnetization direction of a proximal ferromagnet that is in contact with an antiferromagnet. This effect is caused due to the exchange interaction at the interface between a ferromagnet and an antiferromagnet. A soft ferromagnet film will have its interfacial spins pinned if it is strongly exchange coupled to the antiferromagnet in a bilayer or multilayer structure.

Our research adopts a materials-centric approach to computing and combined with a rigorous mathematical underpinning breaks new ground in theory-driven discovery and design of magnetic nanostructures with functional capabilities that go beyond state-of-the-art (SOTA) approaches for brain-inspired computing.

Specifically, this project harnesses the intrinsic brain-inspired spin dynamics in magnetic nanostructures, including antiferromagnetic and ferromagnetic materials and devices, to implement circuits and architectures that can provide a higher degree of cognition (e.g., processing of both spatial and temporal tasks) at much smaller circuit area and high energy efficiency compared to SOTA CMOS and neuromorphic technologies.

The key technical outcomes of this research are as follows.

- For the hardware emulation of neurons, we numerically solved for the spin-torque-induced magnetization dynamics in metallic and insulating antiferromagnets and obtained closed-form expressions for the output as a function of the input spin current. Specific materials analyzed include nickel oxide (NiO), chromium oxide (Cr_2O_3), and a Weyl semi-metal manganese tin (Mn_3Sn).
- We also explored the magnetization dynamics in the pulsed limit (i.e., when the magnet is excited with a pulse of spin current). This regime is particularly useful for low-power applications and in spiking neural networks.
- We investigated memristive dynamics in magnetic tunnel junctions for their use as a synapse in neural networks.
- We developed compact models of interconnects for the neural network which allowed us to estimate the overhead of data communication or signaling in large-scale networks for solving data-intensive tasks.
- We scaled up the models so that the hardware costs of neuromorphic computing could be readily assessed across the full stack, that is ranging from materials hardware costs to the system-level hardware costs.
- We evaluated the hardware costs for realistic workloads which include the LeNet model for MNIST and Fashion-MNIST. A small LeNet (SL) and a large LeNet (LL) were used. Our results conclusively show that bulk of the hardware costs are associated with connections between neurons, rather than the neurons themselves. This indicates that architectures with shorter interconnects or new interconnect materials with lower resistivity at scaled nodes are important research directions to explore in the future.

2.1 Overview of brain-inspired computing

The elementary basis of intelligence in biological systems are the neurons and synapses and their complex interconnections forming neural circuits. *Figure 1* shows a biological neuron. Learning and memory are dynamical in nature and evolve due to the interaction of neural circuits with external environment. The human brain contains 10^{11} neurons with 10^{15} synapses, while it consumes 20 W power to perform complex learning, adaptation, and inferencing functions in massively parallel structures [11]. Even with noisy signals, the brain can operate with extreme efficiency unlike our conventional computers whose performance degrades significantly when noisy inputs are presented to them. That is, conventional computers rely on high-precision data that incurs huge costs in energy and delay to handle real-world uncertainties. Neuromorphic computing that looks to biology to inspire new ways to process data. At the heart of neuromorphic systems are materials and devices that implement neuronal functions, e.g., memory elements, variable weights in artificial synapses, neurons with nonlinear dynamics, and dendrites.

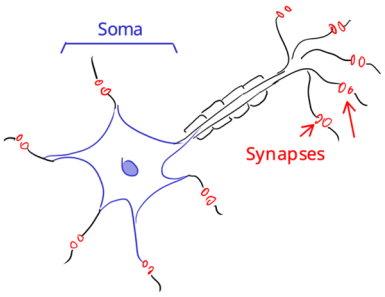

Figure 1. A biological neuron.

2.2 Overview of magnetic materials and devices for neuromorphic computing

Magnetic materials possess ordered arrangement of spins and can be classified as ferromagnetic, ferrimagnetic, or antiferromagnetic. The difference between the three classes is illustrated in *Figure 2*. In the case of ferromagnets, all atomic spins are aligned along the same direction, while in antiferromagnets, spins on the neighboring atoms are aligned in an anti-parallel manner with respect to each other such that there is no net magnetic moment. In the case of ferrimagnets, although the spins on neighboring atoms are arranged in an anti-parallel manner, the two spins do not cancel each other out, resulting in a net macroscopic magnetic moment.

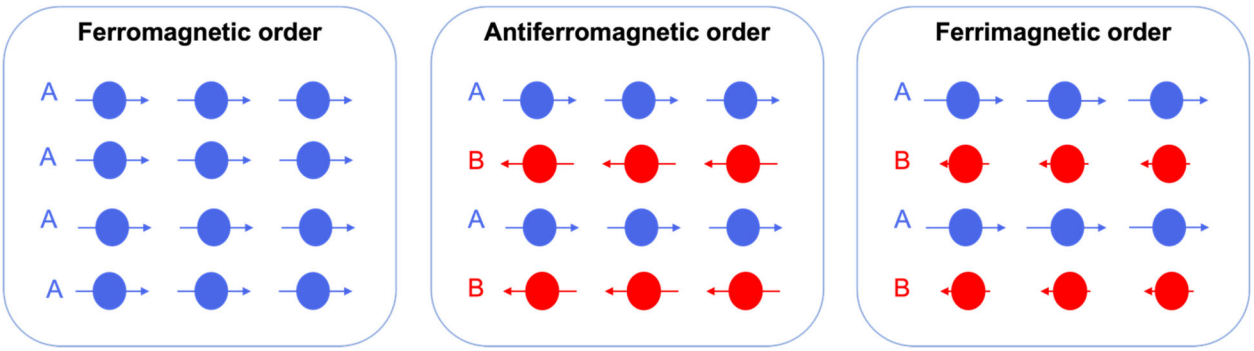

Figure 2. Examples of magnetic ordering.

Magnetic materials have been used to emulate brain-like functionality since the early developments of neural network theory. The seminal works of Karl Steinbuch (1962) on *Lernmatrix* [12] and Hopfield on Ising model [13] with tunable coupling between neighboring spins, mimicking synaptic functions, laid the foundation for artificial neural networks (ANNs) and spin glass systems. The last several years have witnessed significant progress in the fabrication and commercialization of spin-transfer-torque (STT)-driven memory, thereby also opening the door to neuromorphic applications that rely heavily on high-density and low-power memory [14].

Various spin-dependent phenomena such as STT [15], spin-orbit torque (SOT) [16], and magneto-electricity [17] have been utilized to realize compact neurons and synapses in hardware. Magnetic memristors based on domain wall (DW) movement in two-terminal (2T) and three-terminal (3T) magnetic tunnel junctions (MTJs) have also been experimentally reported. Antiferromagnetic synapses using SOT as their switching mechanism are also experimentally demonstrated.

However, one of the biggest challenges, despite these impressive experimental investigations, is a lack of a theoretical framework and models that allow us to use these magnetic devices in circuits and architectures and enable the full stack assessment of the opportunities and challenges of magnetic devices in the context of neuromorphic computing. In addition, many new types of spin dynamics that resemble neuro-synaptic dynamics of biological neurons have not been fully studies in these materials, which means that the full potential of these materials cannot be harnessed at the system-level.

For system-level simulations, we combine comprehensive physical models of magnetic neurons and synapses, as well as interconnects, with a specialized parallel discrete event simulation model called Doryta, developed at Rensselaer Polytechnic Institute (RPI). Doryta is a deterministic, parallel spiking neural network simulation platform that is able to execute real neuromorphic applications in simulation and has been validated against existing spiking neural network tools.

3. Methods, Assumptions and Procedures

a. Antiferromagnetic neurons

Antiferromagnets such as NiO, Cr_2O_3, and alloys of Mn (e.g., Mn_3Sn, Mn_3Ir), etc. are a class of magnetic materials that are internally magnetic on a microscopic scale but possess negligible net magnetization on a macroscopic scale owing to their atomic arrangement [18-20]. In principle, they can be used to realize non-linear signal generators and detectors, operating in the gigahertz to the terahertz frequency spectrum. Such AFM-based signal generators have been theoretically shown to emulate spiking neurons in hardware within a compact form factor.

Antiferromagnets can be classified as collinear or noncollinear. In the former case, the spins are arranged 180-degrees with respect to each other, while in the latter case, the spins form a triangular orientation (120-degree orientation). Examples of collinear and non-collinear antiferromagnets is shown in *Figure 3*.

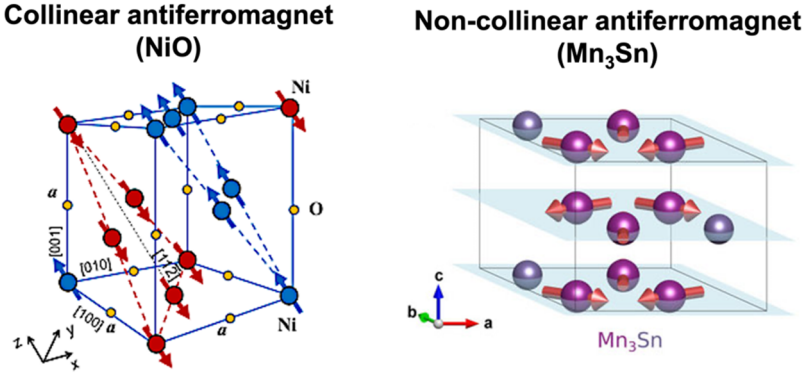

Figure 3. A unit cell of a collinear antiferromagnet (NiO) and noncollinear antiferromagnet (Mn_3Sn).

Figure 4(a) shows an example of an artificial neuron using NiO as the active element, while *Figure 4(b)* shows an artificial neuron using Cr_2O_3 as the active element [21]. In both these cases, Input electric current J_C is converted into a pure spin current J_S polarized $\mathbf{u}_S$ along the hard-anisotropy axis to cause precession of the sublattice moments $\mathbf{m}_A$ and $\mathbf{m}_B$, which pumps spin current back to generate an oscillating voltage signal at the output V_{out}.

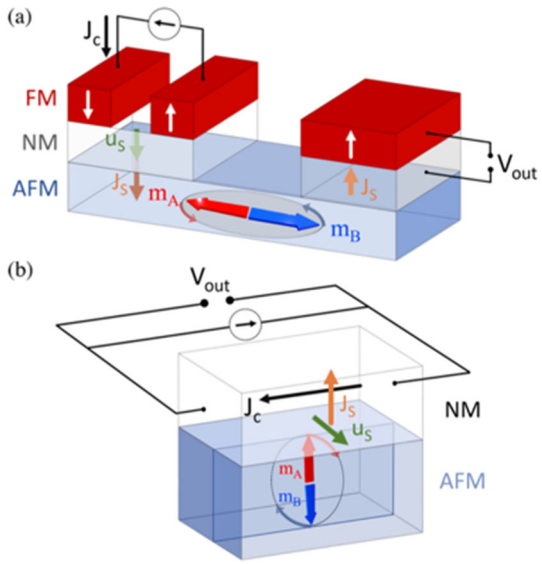

Figure 4. Hardware implementation of a neuron using antiferromagnetic insulators. (a) In-plane spin valve geometry: Lateral spin valve structure made of a perpendicular reference-layer magnetization on top of an in-plane anisotropy antiferromagnet, like NiO. (b) Perpendicular spin Hall geometry: Spin Hall structure with in-plane electric current transverse to the in-plane hard axis on top of a perpendicular anisotropy antiferromagnet, like Cr_2O_3. FM, ferromagnet; NM, normal metal; AFM, antiferromagnet.

An antiferromagnetic spiking neuron based on Mn_3Sn is shown in *Figure 5*(a) and *Figure 5(b)*. The devices differ in terms of their read mechanism. In the first case, the read is based on the anomalous Hall effect (AHE), while in the second mechanism, tunneling magnetoresistance (TMR) of an antiferromagnetic junction is used. The time domain output of the neuron is shown in *Figure 5(c)*. Both AHE and TMR have been experimentally measured in antiferromagnetic structures. Other detection methods like anisotropic magnetoresistance and tunneling anisotropic magnetoresistance could also be used, although their signal is weaker compared to TMR and AHE at 300 K.

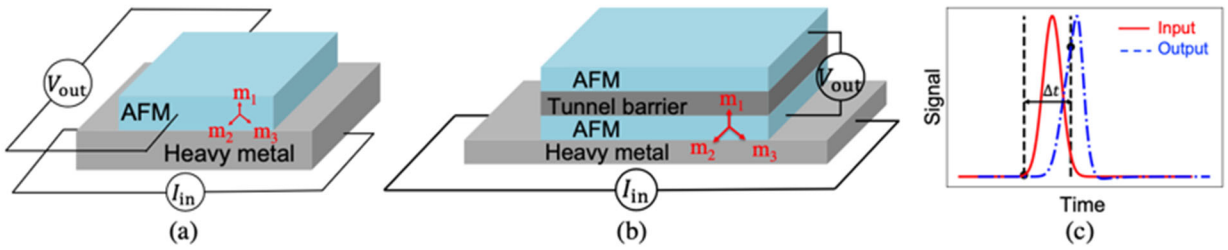

Figure 5. AFM neuron based on (a) anomalous Hall effect readout and (b) tunneling magnetoresistance readout. (c) A qualitative sketch of the time-domain output signal, V_{out} of the neuron, versus the input signal, I_{in}. In (a) and (b), $\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$ represent the magnetization vectors of the AFM. In (c), the neuron latency is the timing difference between the input signal arrival and the time the neuron's angular velocity reaches a threshold, which is chosen as 10^{10} rad/s for Mn_3Sn and $2\text{x}10^{12}$ rad/s for NiO.

i. Neuron Modeling with DC Currents

Insulating antiferromagnetic neurons (NiO and Cr_2O_3): To excite and control Néel-order precession, the spin current must exert adequate anti-damping-like torque. The threshold spin-current density that initiates precession is given as $J_{s,th} = K_e\, d_a/2$, where K_e is the easy-axis anisotropy, and d_a is the thickness of the AFM.

The corresponding electric-current density depends on the specific geometry. For the in-plane spin valve geometry, assuming low spin-memory loss in the normal metal, the conversion from electric current to spin is determined by the conductance of majority- and minority-spin electrons g_M and g_m, respectively, and the spin-mixing conductance at the interface of normal metal and antiferromagnet g_r. The threshold electric-current density is [21]

$$J_{c,th}(NiO) = \frac{2e}{\hbar}\frac{(g_r + g_M + g_m)(g_M + g_m)}{g_r(g_M - g_m)}J_{s,th.}$$

Here, e is the elementary charge, and $\hbar$ is the reduced Planck's constant.

For the perpendicular spin Hall geometry, the conversion from electric current to spin is determined by the spin Hall angle, θ_{SH}, the layer thickness, d_n, the spin diffusion length, λ, the conductivity, σ of the heavy metal, and the spin-mixing conductance at the interface of the heavy metal and the antiferromagnet g_r. The threshold electric-current density is given as [21]

$$J_{c,th}(Mn_3Sn) = \frac{e}{\hbar}\frac{\sigma}{\lambda g_r}coth\frac{\left(\frac{d_n}{2\lambda}\right)}{\theta_{SH}}J_{s,th}$$

The precessing Néel order can reciprocally pump time-varying spin current into the adjacent metal layer and experience a damping-like backaction, which virtually enhances Gilbert damping according to $\alpha = \alpha_0 + \alpha_s$, where

$$\alpha_s = \frac{\hbar^2 \gamma g_r}{2e^2 M_s d_a}.$$

The pumped spin current is converted into voltage under open-circuit conditions via spin filtering for the in-plane spin valve geometry and via the inverse spin Hall effect for perpendicular spin Hall geometry. The voltage signal generated at the output of each geometry is

$$V_{out}^{ip-valv}(t) = \frac{\hbar\gamma\sqrt{JK_e}}{2eM_s}\frac{(g_M - g_m)g_r\omega(t)}{(g_r + g_M + g_m)(g_M + g_m)}, V_{out}^{pe-Hall}(t)$$
$$= \frac{\hbar\gamma\sqrt{JK_e}}{2eM_s}\frac{\lambda g_r}{\sigma}\theta_{SH} tanh\frac{d_n}{2\lambda}\omega(t),$$

where $\omega(t) = \varphi'$ is the dimensionless angular frequency. This frequency is obtained by solving the second order equation of motion for NiO and Cr_2O_3 antiferromagnets, given as

$$\varphi'' + \beta\varphi' + sin\varphi = \Gamma,$$

where $\beta = \alpha\sqrt{J/K_e}$ is the viscous damping, and $\Gamma = 2J_s/K_e\, d_a$ represents the constant tangential force acting on the Néel vector.

Metallic antiferromagnetic neuron (Mn_3Sn): Mn_3Sn is a hexagonal antiferromagnet with ABAB stacking sequence of the (0001) plane consisting of a Kagome lattice of Mn atoms with a small magnetic anisotropy in the Kagome plane. Mn_3Sn can only be stabilized in excess of Mn atoms (which replaces some of the Sn atoms), whereas with lower Mn concentration, the system gets contaminated with Mn_2Sn. The combination of exchange coupling and DM interaction between the Mn moment stabilizes an anti-chiral 120 degrees spin structure, as shown in *Figure 6*, below the Neel temperature of 410-420K. At lower temperatures we get other phases including the spin glass like texture below 50K. In the presence of Mn deficiency, a helix phase is also possible where Mn moments form a spiral structure propagating along the c-axis. The combination of three sublattices having two-fold single-ion anisotropy yields six-fold magnetic anisotropy. Mn_3Sn is a Weyl semi-metal which leads to large magneto-transport responses even though it has a vanishingly small magnetization about $0.002\mu_B$ per Mn atom.

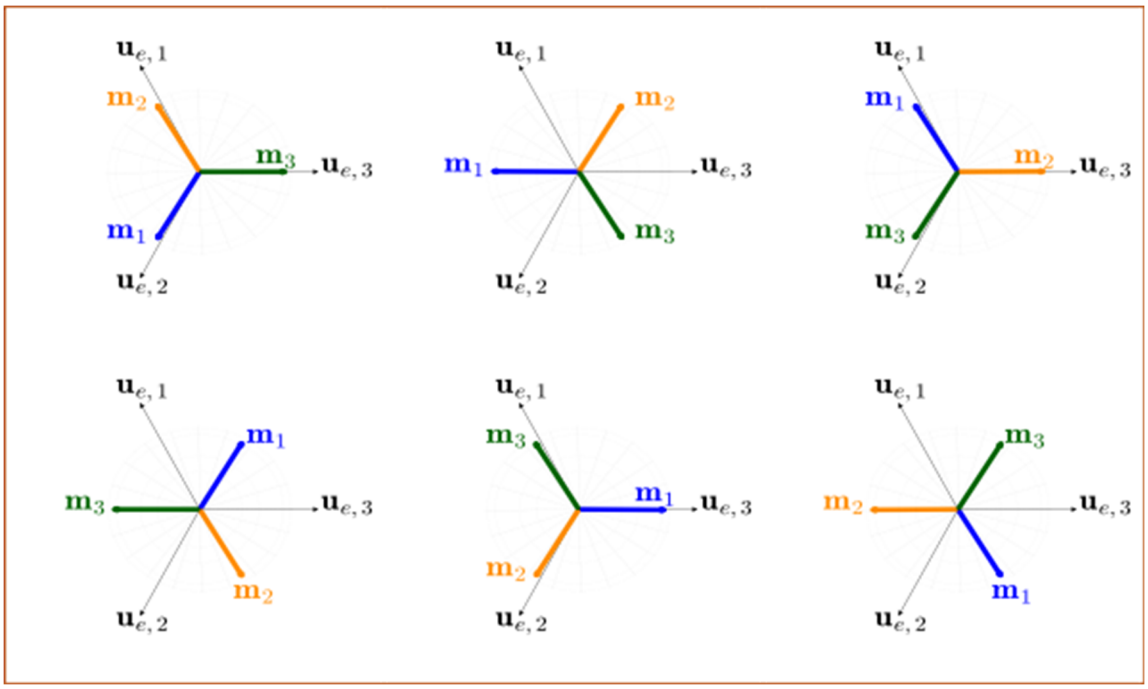

Figure 6. Six equivalent stable states in equilibrium for antiferromagnetic Mn_3Sn.

When Mn3Sn is perturbed using spin orbit torque from an adjacent heavy metal, then the Néel vector undergoes oscillations. The equation of motion of the azimuthal angle of Néel vector, measured from *x*-axis, is expressed as [22]

$$\varphi'' + \alpha\omega_E\varphi' + y\omega_K\omega_E sin(6\varphi) + \omega_E\omega_s = 0,$$

where α is the Gilbert damping, $\omega_E = 3\gamma J/M_s$, $\omega_K = 2\gamma K_e/M_s$, $\omega_s = \frac{\hbar}{2e}\frac{\gamma J_s}{M_s d_a}$. Here, M_s is the saturation magnetization and γ is the gyromagnetic ratio. The factor y in the equation of motion is of the order of $\left(\frac{\omega_K}{\omega_E}\right)^2 \ll 1$, which means that the frequency of oscillations of the Néel vector can be tuned readily over a very broad range, from the gigahertz to the terahertz.

As shown in *Figure 7*, the frequency of oscillations for currents that are close to the threshold value is 191 MHz. However, as the input excitation magnitude increases by 4x, the frequency of oscillations increases to 1.14 GHz.

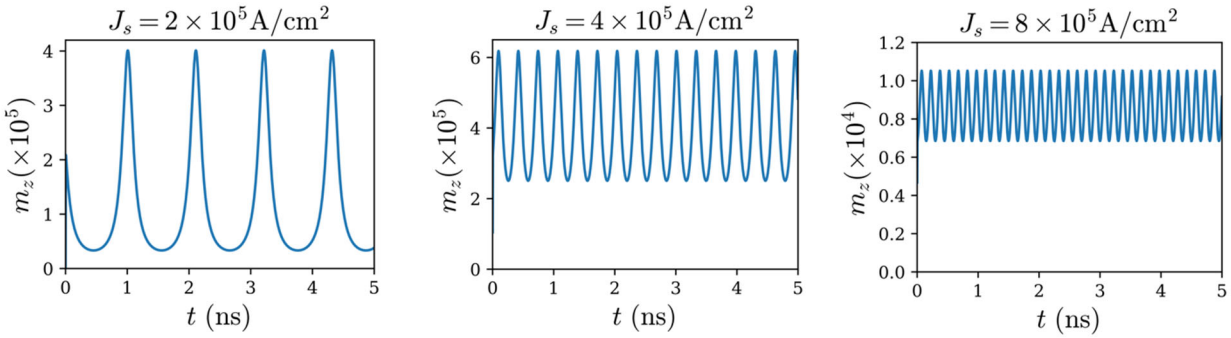

Figure 7. Oscillation of the z-component of the average magnetization of Mn3Sn. The critical spin current is $1.7 \times 10^5 \frac{A}{cm^2}$. *As shown in the left graph, the output resembles the spikes of a neuron because the input excitation is close to the threshold. As the input excitation increases, the response of the average magnetization begins to oscillate coherently and at a much higher frequency.*

From the equation of motion, we derived the scaling of frequency of oscillations with input excitation [22, 23]:

$$f = \frac{1}{2\pi}\frac{\omega_s}{\alpha} = \frac{1}{2\pi}\frac{\hbar}{2e}\frac{\gamma}{M_s}\frac{J_s}{d_a}\frac{1}{\alpha}.$$

The validation of the model of frequency against numerical simulations is shown in *Figure 8*.

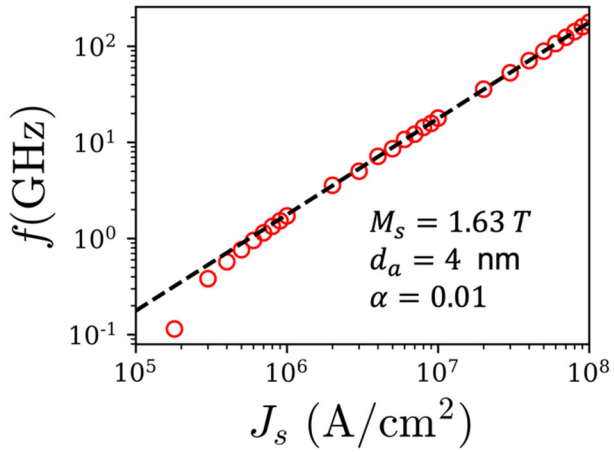

Figure 8. Scaling of frequency with input spin current in Mn₃Sn. Symbols are numerical solutions of the equation of motion, while dashed line represents the analytic model.

ii. Neuron Modeling with Pulsed Currents

For a practical AI hardware, it is preferred to study the dynamics of the neurons when they are perturbed by a pulsed input, rather than a DC input as we have discussed previously. In this case, the equation of motion is the same as described previously, but the solutions of the average magnetization and the angular velocity are obtained for a Gaussian input signal as shown in *Figure*

5(c). The latency of the neuron is defined as the time it takes from when the input pulse arrives to the time the output of the neuron reaches a critical value. *Table 1* shows the latency of Mn_3Sn neuron for different pulse widths and peak values of the input Gaussian excitation.

Table 1. Mn_3Sn neuron latency for various Gaussian input excitations. For the calculation of the energy dissipation, we use 2 MA/cm² as the input current and its pulse width is set as 5 ps. Thus, the latency of the neuron is 12 ps.

Δt (ps)	1 MA/cm²	2 MA/cm²	3 MA/cm²	4 MA/cm²	5 MA/cm²
5	N/A	12 ps	9 ps	8 ps	7 ps
10	32 ps	21 ps	18 ps	15 ps	14 ps
20	55 ps	40 ps	34 ps	30 ps	28 ps

To calculate the energy consumption of the neuron, we consider that the spin current is generated via the spin Hall effect in an adjacent heavy metal layer in a structure, similar to the one depicted in *Figure 5*(a) and (b). Assuming that the antiferromagnetic layer has a cross-section of 40 nm × 120 nm × 4 nm and for the spin current density of 2 MA/cm², the required spin current becomes $I_s = 2 \times 10^6 \times (40 \times 120 \times 10^{-14}) = 96\mu A$. The required charge current is $I_c = I_s/\theta_{SH}$. Assuming $\theta_{SH} = 0.056$, we get $I_c = 1.7mA$.The energy consumption occurs due to Joule heating associated with passing the charge current via the spin Hall layer. Thus for a Gaussian signal,

$$E_{neuron} = I_c^2 R \Delta t \sqrt{\pi},$$

where $R = \rho L/A$ is the resistance of the heavy metal layer. Assuming $\rho = 10\mu\Omega$.cm, $L = 120$ nm, and $A = (40 \times 4)nm^2$, we get $R = 10 \times 10^{-8}(\Omega.m) \times 120 \times \frac{10^9}{40\times4} = 75\Omega$. Substituting the values in the equation for energy dissipation we get, $E_{neuron} = (1.7 \times 10^{-3})^2 \times 75 \times 5 \times 10^{-12} \times \sqrt{\pi} = 2 \times 10^{-15} Joules$. The latency of the neuron is 12 ps.

To read out the value encoded in the magnetization of the neuron, we can exploit either the AHE or the TMR read-out process. Theoretically, it is predicted that the minimum and maximum resistance-area product of an antiferromagnetic tunnel junction is $RA_{min} = 0.04\mu\Omega.\mu m^2$ and $RA_{max} = 0.2\mu\Omega.\mu m^2$, which gives $\Delta RA = 0.16\mu\Omega.\mu m^2$. For a cross-sectional area of $(40 \times 120)nm^2 \Rightarrow \Delta R = 33\Omega$. To get a detectable read voltage, a read current on the order of a milli-Amp is required. Hence, we can conclude that for $I_{read} = 1mA, V_{out} = 33mV$,which can be detected via microelectronics compatible circuitry (for example, by amplifying the signal via sense amplifier). Note that $I_{read} = 1mA$ translates to a read current density of $J_{read} = I_{read}/A_{read}$, where $A_{read} = (40 \times 120)nm^2$. This gives us $J_{read} = \frac{10^{-3}}{40\times120} \times 10^{18} \frac{A}{m^2} = 2.1 \times 10^{11} \frac{A}{m^2} = 2.1 \times 10^7 \frac{A}{cm^2}$. Although the required current is large, we only need to turn it on during the read process, which can be quite fast, in the range of picoseconds.

For readout, we could use the anomalous hall effect in Mn_3Sn. Here, the polarization of the magnetic octupole determines the polarity of the anomalous Hall signal. When a heavy metal layer contacts the Mn_3Sn, the AHE potential from Mn3Sn drives a transverse current in the HM (as shown in). Thus, for the transverse direction, the bilayer can be considered as a closed circuit of

two resistors connected in series with this potential V_y(Mn_3Sn). The total Hall voltage in the bilayer is then given as

$$V_y = I_{read}\left[1 + \frac{\rho^{Mn_3Sn}}{\rho^{HM}}\frac{t^{Mn_3Sn}}{t^{HM}}\right]^{-2} R_{xy}^{Mn_3Sn}$$

Here, ρ represents the resistivity of various layers, while t represents the thickness. The resistance $R_{xy}^{Mn_3Sn}$ is the Hall resistance of the layer, which is typically measured to be in the range of (10 – 350 mΩ). Using $I_{read} = 4mA$, $\frac{\rho^{Mn_3Sn}}{\rho^{HM}} = 1$ and $\frac{t^{Mn_3Sn}}{t^{HM}} = 1$, we get V_y in the range of (0.04 mV to 1.6 mV). In this case, the sense amplifier design to amplify the signal would be rather complex. To improve the detected signal, either I_{read} or the Hall resistance must be increased.

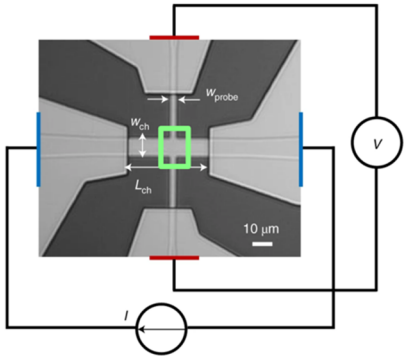

Figure 9. A Hall cross structure fabricated on Mn_3Sn to measure the anomalous Hall voltage. Figure from [24].

The summary of the results of NiO and Mn_3Sn neurons' performance is given below. Results are reported for a pulsed input with a duration of 5 ps. It can be seen that Mn_3Sn is more energy efficient compared to NiO. This is because the critical spin current needed to excite oscillations in Mn3Sn is an order of magnitude lower than that in the case of NiO.

Table 2. For both antiferromagnets, the cross-sectional area is (120×40) nm^2. The thickness of both antiferromagnets is 4 nm. The read mechanism in NiO is spin pumping in a lateral spin valve structure, while in the case of Mn_3Sn, the read mechanism could be TMR or AHE, although TMR is expected to result in higher output voltage. The resistance of the spin-Hall layer for the write process in both cases is 75 Ω.

Neuron type	Write current (A/cm^2)	Write energy (Joules)	Write latency (ps)	Read voltage (Volts)
NiO	3.57× 10^8 (spin current = 2× 10^7 A/cm^2 and $\theta_{SH} = 0.056$)	2×10^{-13}	15	400 µV.
Mn_3Sn	3.57× 10^7 (spin current = 2× 10^6 A/cm^2 and $\theta_{SH} = 0.056$	2×10^{-15}	12	33 mV (TMR, assuming $I_{read} = 1mA, \Delta RA = 0.16\mu\Omega.cm^2$)

b. Ferromagnetic synapses

Memristive dynamics based on domain wall (DW) movement [24] can be easily produced in a stripe-shaped ferromagnetic structure as shown in *Figure 10*(a). The device response is shown in *Figure 10*(b), which highlights the ability of the device to store real-valued weights with plasticity. Thus, ferromagnetic materials can act as compact and non-volatile hardware emulators of synapses. During the training phase, current flows between terminals T2 and T3. The synapse's conductance is set by the magnitude and duration of the input current, I . *Figure 10*(c) shows the movement of the domain-wall motion with increasing input pulse durations. During inferencing, the output current between terminals T1 and T3 is measured. The output current is given as the product of the voltage across T1 and T3 and the memristor's conductance. The memristors can be electrically connected in an analog crossbar fashion such that the net current flowing through the bit line is the weighted sum of the memristors' conductance multiplied by the input voltage.

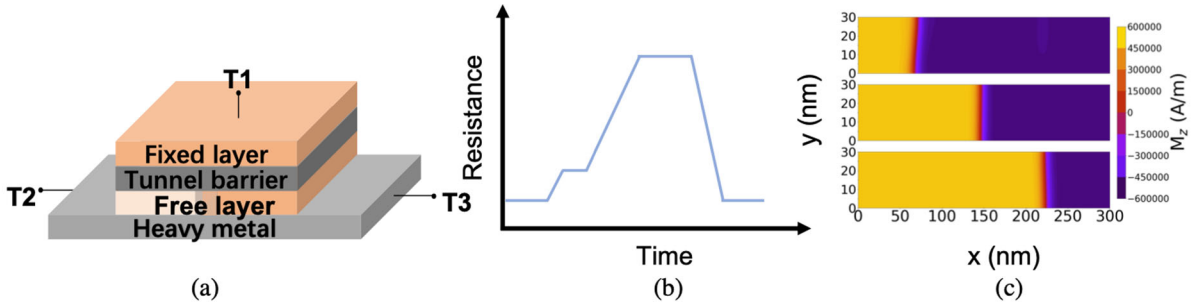

Figure 10. Schematic of a spintronic synapse where the free layer and the fixed layer are made of ferromagnetic materials. Input current is applied across the terminals, T2 and T3, while the output is measured across T1 and T3. (b) A representative response of the ferromagnetic synapse over time due to applied input current. (c) A representative result showing the movement of domain wall under applied training current using MuMax3c.

i. Synapse Modeling

To model the memristive domain wall-based ferromagnetic synapses considered here, a 30 nm x 300 nm x 2 nm device that can host up to 64 distinct resistance levels is considered. Ignoring any effects of device-level non-ideality, the synapse's resistance-area (RA) product ranges from $RA_{min} = 8 \times 10^{-12}\Omega \cdot m^{-2}$to $RA_{max} = 27 \times 10^{-12}\Omega \cdot m^{-2}$, which gives an average value of resistance as $R_{avg} = \frac{RA_{min}+RA_{max}}{2A} = \frac{8\times10^{-12}+27\times10^{-12}}{2(30\times300)} \times 10^{18} = 1.944\text{k}\Omega$. Assuming that the current of the synapse is the same as that required for neuron to fire, the energy dissipation of the synapse is given as

$$E_{syn} = I_{neu}^2 R_{avg} \Delta t \sqrt{\pi},$$

where Δt = 5ps and $I_{neu} = 1.7$ mA (Mn_3Sn) or 17 mA (NiO). Using appropriate material values, the energy of the synapse is 5×10^{-14} Joules for feeding into Mn_3Sn neuron, while it is 5×10^{-12} Joules for feeding into NiO neuron.

c. Interconnects

In addition to the neurons and synapses, we consider the interconnects in the performance evaluation. The interconnects are assumed to be implemented using Copper/low-k technology. The resistivity of copper is assumed to be affected by both sidewall scatterings and grain boundary scatterings [26, 27].

$$\rho = \rho_0 + \rho_0\lambda\frac{3}{4W}(1-p) + \rho_0\lambda\frac{3R}{2D(1-R)},$$

where $\rho_0 = 1.67 \times 10^{-8}$ Ω.m (bulk resistivity), $\lambda = 40$nm, $p = 0.5$, $R = 0.3$, D is the grain size and assumed to be the same as the height of the interconnect (D=h); for an aspect ratio of 2, $D = 2W$ (W is the interconnect width). With a liner thickness of L_b = 3 nm, the effective width reduces to $(W - 2L_b)$ while the effective height reduces to $(D - L_b)$.

The capacitance per unit length of the interconnect is given as [28]

$$C_l = \epsilon_{ox}\left[1.15\left(\frac{W}{H_{di}}\right) + 2.8\left(\frac{h}{H_{di}}\right)^{0.222}\right] + 2\epsilon_{ox}\left[0.03\left(\frac{W}{H_{di}}\right) + 0.83\left(\frac{h}{H_{di}}\right) - 0.07\left(\frac{h}{H_{di}}\right)^{0.222}\right]\left(\frac{S}{H_{di}}\right)^{-1.34}$$

Here, ϵ_{ox} is the oxide permittivity, H_{di} is the dielectric thickness, and $S = W$ is the spacing between the interconnects.

Figure 11 show the impact of dimensional scaling on the interconnect resistivity and capacitance per length. It can be seen that scaling the width of the interconnect adversely impacts the interconnect resistance, which is associated with the delay of the interconnect. In the evaluation of neuromorphic architectures, we consider interconnect widths of 10 nm, 15 nm, 20 nm, 25 nm, and 30 nm.

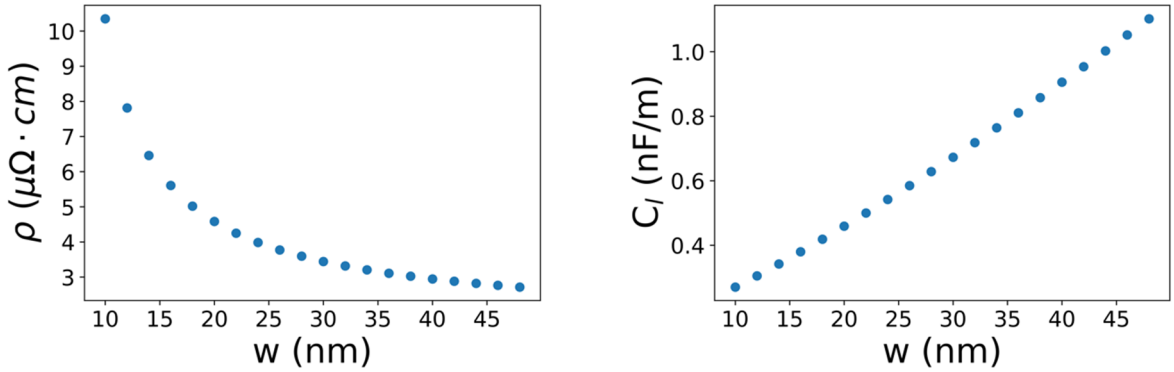

Figure 11. Impact of dimensional scaling of interconnect on its resistivity and per-unit-length capacitance.

d. Evaluation of neuromorphic architectures

In the evaluation of neuromorphic architectures, we consider that the network is composed of multiple layers, each layer includes several identical cores, and each core is made up of a crossbar array of neurons and synapses as shown below.

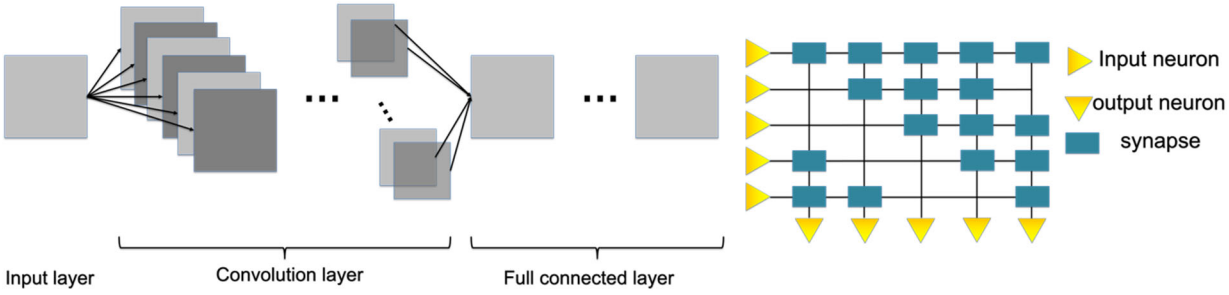

Figure 12. Setup for the evaluation of neuromorphic architectures using magnetic neurons and synapses.

Instead of thoroughly designing each part of the chip, we introduce empirical factors to approximate the area associated with peripherals and to accommodate the design rules for component spacing. All cores in our work are connected using a crossbar architecture. The core area is given as

$$a_{core} = \left(a_{neu} n_{core} F_{neu} + a_{syn} n \;\; n_{out} F_{syn}\right) F_{core},$$

where $F_{syn} = F_{neu} = F_{core} = 2$ represent the empirical area factor for synapses, neurons and cores, respectively, while n and n_{out} are the number of input and output neurons for the core. The area of the neuron is represented as a_{neu} and the area of the synapse is represented as a_{syn}.

In case the number of input neurons is smaller than the synapses per neuron, the core area is estimated using the convolution core architecture. Replacing the input neuron numbers by the synapse numbers per neuron (s_{neu}), we obtain

$$a_{core} = \left(a_{neu} n_{core} F_{neu} + a_{syn} s_{neu} n_{out} F_{syn}\right) F_{core},$$

The fan-in of AFM neurons is considered to be infinite because the input current sums up in the interconnects naturally. All the cores in the same layer operate simultaneously and each core operation delay is determined by one synaptic operation and one neuron operation. The delay per core is given as

$$\tau_{core} = \tau_{neu} + \tau_{syn} + \tau_{neu,ic} + \tau_{syn,ic}.$$

Here, $\tau_{neu,ic}$ refers to the delay of global interconnects that connect neurons across the layers, while $\tau_{syn,ic}$ refers to the delay of local interconnects that connect synapses to neurons within a single layer.

The energy consumption per core is the summation of energy dissipated in all the synapses and neurons. The energy consumption of a core is workload-dependent because both the active synapse rate and the neuron fire rate depend on the workload-related parameters presented to the neurons and synapses. The active synapses per neuron and the active neurons per inference (per core) are

$$s_{act,neu} = s_{neu}s_{act}, n_{act,core} = n_{core}n_{act}.$$

Here, s_{act} and n_{act} are the fraction of active synapses and neurons per layer, respectively. Adding all up, the energy consumption per core is

$$E_{core} = E_{syn}s_{act,neu}n_{core} + E_{neu}n_{act,core}.$$

At the chip level, the area and energy consumption are the summation of all cores:

$$a_{chip} = \Sigma_i\Sigma_j a_{core,ij},$$

$$E_{chip} = \Sigma_i\Sigma_j E_{core,ij},$$

where the index *i* is the layer index, while *j* is the core index in a specific layer. It is worth mentioning that all cores of the same layer operate in parallel and thus the chip latency is given as

$$\tau_{chip} = \Sigma_i \underset{layer,i}{max}\tau_{core,j}.$$

4. Results and Discussion

Following Nikonov and Young's energy estimation strategy [29], we defined a crossbar architecture for the task of image recognition on MNIST. Then, given a spike workload, we calculated the area, latency and energy of the crossbar architecture for the classification of a single image. In their framework, Nikonov and Young assume an approximate layout for each component of the architecture in the chip, an approximate placement for each neuron, synapse and interconnect in each crossbar. Our crossbar architectures, like theirs, follow the LeNet architecture, as presented in (ours only differing on the size of the last layer). For a more accurate performance estimation, we simulate the network in Doryta and gather the usage of each component (leak, fire and integrate) to create a specific workload for LeNet using MNIST and Fashion-MNIST. The performance estimation process takes all equations from as described previously, and computes them given the workloads produced by Doryta.

All workloads were obtained from Doryta's output. First the trained LeNet model on MNIST (or Fashion-MNIST) was loaded into Doryta, and then a total of 10,000 test 28 × 28 black and white images as spikes were injected into the simulation. From processing these images, Doryta calculated the usage of key operations (fire, integrate and leak) for each layer of the network. The first workload was based on the LeNet model with the MNIST dataset, we nicknamed it the Small LeNet or "SL" workload. The second workload was constructed by training LeNet on the Fashion-MNIST dataset, which is a dataset made to be a "hot swapping" replacement of MNIST where

each image belongs to one of ten categories of clothing (the Small LeNet Fashion or "SLF" workload). To improve accuracy, we increased the number of filters for the four intermediate layers of LeNet from 6 and 16, to 32 and 48, respectively. We found a 2% improvement in accuracy with this larger network on MNIST (Large LeNet "LL"), and a 3% improvement on Fashion-MNIST (Large LeNet Fashion or "LLF").

Table 3. Cross-bar configurations for LeNet

#	Type	Input Lines	Filters	Number of Neurons	Synapses per Neuron
1	Conv	784	1	784	1
2	Conv	784	6	784	22.90
3	Conv	784	6	196	4
4	Conv	1176	16	100	150
5	Conv	100	16	25	4
6	Full	400	-	120	400
7	Full	120	-	84	120
8	Full	84	-	100	84

It is worth mentioning that, according to the output voltage of AFM neurons we adopted, sense amplifiers are required as part of the neuron to generate an appropriate signal. The absence of sense amplifier in this benchmark overestimates the performance of spintronics based networks. *Figure 13* shows the dependence of the latency and energy consumption on the interconnect width. The network latency increases as the wire width shrinks because of higher surface and grain boundary scatterings of the interconnect, which increases the interconnects' RC delay. The capacitance per unit length has a weak dependence on the wire width and it contributes to the energy consumption drop for smaller interconnect dimensions.

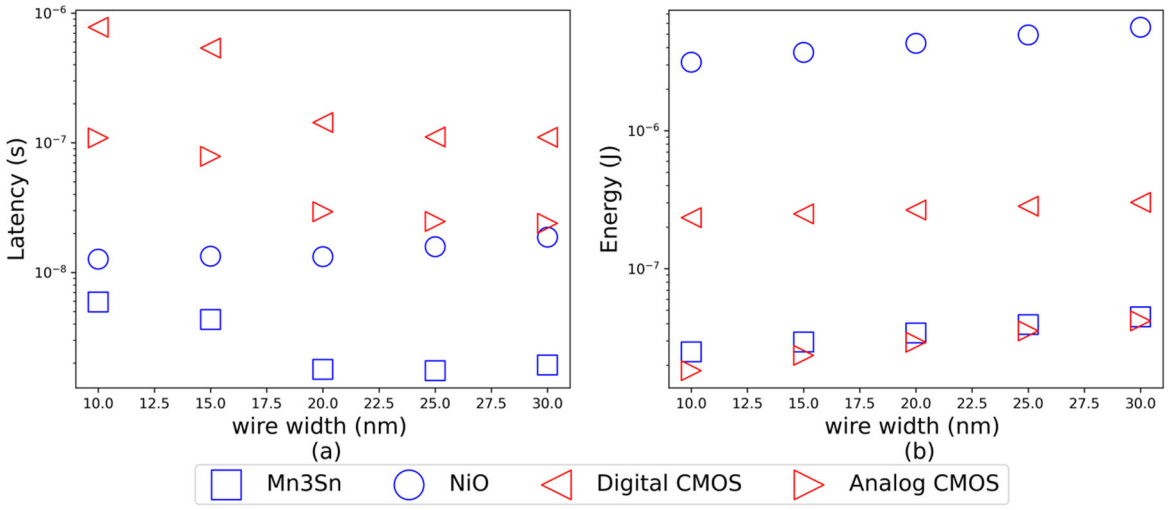

Figure 13. (a) Latency of Small LeNet workload versus interconnect width. (b) Energy consumption of Small LeNet workload versus interconnect width for one inference.

The energy consumption is composed of four different parts, as illustrated in *Figure 14*: neurons, synapses, interconnects connected to the synapses, and interconnects connected to the neurons. The energy of the neuron firing is negligible, while the energy consumed in interconnects, especially the short interconnects with synapses, is dominant. In addition, we find that synapses themselves are comparable in energy dissipation with short local interconnects that connect synapses to wires. On the latency side, the longer interconnects dominate the chip performance. *Table 4* shows a comparison of energy, delay, and energy-delay product of various neuromorphic device options. Benchmarks corresponding to analog and digital CMOS neural networks are also included as a comparison. We can see that magnetic devices occupy significantly smaller footprint compared to their silicon counterparts. Moreover, magnetic devices also operate at a much higher speed, 3-100×, compared to analog and digital CMOS implementations. Between Mn3Sn and NiO, the performance of all networks is much superior if the network is implemented using Mn3Sn. This is because Mn_3Sn being a Weyl semi-metal with a weak in-plane anisotropy can be excited at much lower currents, compared to NiO.

Comparing our results to real world data is fundamental to check their validity. For this, we look at the energy performance of a neuromorphic chip implementation, IBM's TrueNorth chip. According to Cheng et al. [30], TrueNorth could inference images from the MNIST dataset at a rate of 1249 frames per second with an energy performance of 6122.44 frames per second per Watt. Given that 1 Watt equals 1 J/s, 6122.44 frames/s/W is the same as 6122.44 frames/J or 163.334 μJ/frame (163, 334 nJ/frame). One might be tempted to compare this energy estimate, 163, 334 nJ/frame, with the estimates of *Table 4*, but one must be careful because the size and shape of the architectures studied on both works are different. Cheng et al. based their work on the CIFAR network architectures, while ours is based on LeNet. Additionally, TrueNorth was conceptualized as a collection of 256×256 crossbars (called cores), while we assume variable size crossbars up to a size of 784x784. We approximate that the number of TrueNorth cores necessary to implement our largest network (LLF) is at least 1400. Given that Cheng et al. used 4064 cores for their network, an appropriate factor to compare their estimation to ours is 3 (divide 4064 by 1400). This means that inferencing an imagine using TrueNorth chip with an architecture based on LeNet would take around 54.4 µJ, while our estimation is 4.65 µJ for digital CMOS. This one order of magnitude discrepancy can be explained by the cost associated with peripheral circuitry and transducers or amplifiers that will be needed, which we do not account for in our calculations.

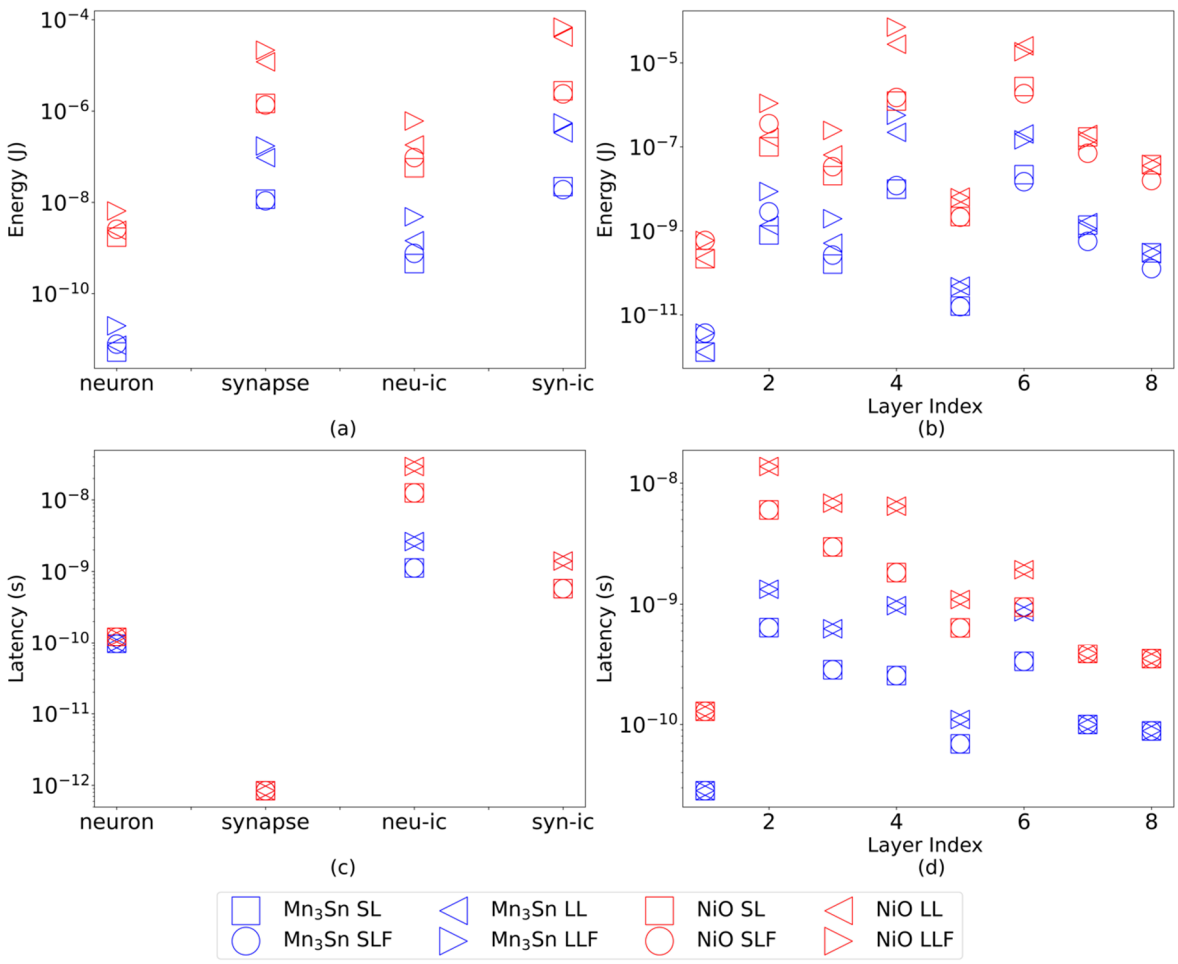

Figure 14. (a)Energy consumed by different components: 'neu-ic' ('syn-ic')ischip-level(core-level)interconnect.(b)Energyconsumption reported layer by layer. (c) Latency of devices and interconnects. (d) Latency of each layer.

Figure 14(b) shows the energy dissipated in each layer, which is dependent on workload allocation. The SLF workload has approximately 40% more integration and firing operations but the energy consumption is similar to the SL workload. Also, the LLF's integration and firing are 2.5 times that of LL's but less than twice that of energy consumption. The fourth and sixth layer are dominant in energy consumption and the 'Fashion' workload does not consume much more in these two layers so they are more energy efficient.

The latency of each layer is the collaborative contribution of one synaptic operation and one neuron operation, and it is dominated by the number of layers and the interconnect delay per core. The area of each layer, dominated at the chip level by the interconnect, is the largest contribution to the delay difference between different workloads implemented using the same neuron device. *Figure 14*(d) shows the latency of processing single layers. Note that only the network size, the number of cores and their size, influence the latency of each layer. There is no difference latency-wise between workloads running on the same network architecture (e.g., no difference in latency between SL and SLF).

Table 4. Performance of various technologies on different workloads. Note that the iso-latency power dissipation of various networks will be directly proportional to their energy consumption and is therefore not reported specifically in the table.

Neuron	Workload	Area(mm^2)	Latency (ns)	Energy (nJ)	E·τ (10^{-18} s·J)
Mn3Sn	SL	0.18	1.79	34	61
	SLF	0.18	1.79	31	55
	LL	1.04	4.11	428	1759
	LLF	1.04	4.11	725	2980
NiO	SL	0.18	13	429	5577
	SLF	0.18	13	383	4979
	LL	1.04	31	5364	166284
	LLF	1.04	31	9092	281852
Analog CMOS	SL	3.5	29	29	862
	SLF	3.5	29	26	762
	LL	20	49	414	20400
	LLF	20	49	693	34100
Digital CMOS	SL	38	143	266	38100
	SLF	38	143	243	34800
	LL	204	328	2670	875000
	LLF	204	328	4650	1520000

Even though the performance results for magnetic neural networks are exciting, the spintronic devices have some limitations compared to ideal LIF neurons. First, the weights that synapses store can only be non-negative. We found that enforcing non-negative constraints on the synapses weights resulted in trickier to train SNN models, yet, once trained, the workloads were not substantially different from those that we used. Secondly, neuron leak and threshold cannot be tweaked as they are values intrinsic to the materials. Fortunately, when only considering positive weights for synapses, the neuron threshold can be fixed, so that the synapses weights need only to be scaled accordingly to fit the fixed threshold. Lastly, the spintronic synapses have about 64 different levels (6 bits), which—compared to the SNNs trained for this work—is a fraction of what single floating-point numbers can store. We found that restricting the network to unsigned 8-bit numbers did not affect accuracy by more than half a percent, but when restricting the network to 6 bits there was significant reductions in accuracy. However, 4 bits have been shown to be enough for NNs to learn. Thus, with the right technique to discretize the network, 6 bits could behave as 8 bits, i.e., our workloads are a good enough approximation of what the architecture would encounter when constrained to 6 bits. This is without considering different input image encoding. We found positive improvements on accuracy (up to 10%) when spikes are temporally encoded.

a. Key Accomplishments

The following major accomplishments were achieved during this effort:

- Development of analytical models to describe the spiking dynamics in antiferromagnetic neurons
- Device proposals for the electrical detection of spiking dynamics in both insulating and metallic antiferromagnets
- Development of analytical models for ferromagnetic synapses
- Benchmarking interconnect latency as a function of scaling and its impact on neuromorphic cores and chips
- End to end modeling and simulation of neuromorphic cores using various technologies and understanding the impact of workload on hardware costs
- Quantitative estimates of energy, delay, energy-delay product, and area of magnetic neuromorphic chips, analog CMOS chips, and digital CMOS chips

The results of this work will be used to inform ongoing research efforts at UIUC that are focused on the development of magnetics-based neuromorphic hardware for AFRL applications. More broadly, our end-to-end evaluation of a new type of AFM neuron and thermodynamics limits of computing have applications for a variety of machine learning, neuromorphic computing, and AI efforts.

5. Conclusions

We estimated the hardware costs associated with neuromorphic tasks. To do this, we used Dortya, a parallel discrete-event-based, chip-agnostic simulator for neuromorphic applications. Our framework allowed us to estimate the energy costs associated with magnetic and CMOS neuromorphic building blocks. For magnetic devices, we considered antiferromagnets to implement spiking neurons, while ferromagnets to implement non-volatile analog memristors that act as synapses. Both metallic and insulating antiferromagnets were considered. We showed that Mn3Sn (metallic antiferromagnet) outperforms NiO (insulating antiferromagnet) in terms of neuromorphic performance. We also showed that when the neurons are optimized, then the bulk of the energy consumption of the chips is attributed to the connections or interconnects between neurons. Compared to CMOS architectures, our analysis indicates that Mn3Sn-based chips have an energy delay product that is four orders of magnitude smaller at inferencing a single image using the LeNet architecture.

For future work, we intend to implement further non-ML applications using SNNs, such as digital circuits, RAM, and a fully-fledged digital computer. On the hardware performance estimation, we plan to incorporate the cost associated with peripheral circuitry and transducers or amplifiers that are needed in a working chip. We also plan to look into ferromagnetic based spiking neurons due to their ease of fabrication and characterization in the GHz regime. Finally, since we found that interconnects are the bottlenecks, we see an interesting avenue for research in the improvement of interconnects. We believe that building on Doryta and the techniques in this work will be beneficial to deepening our understanding of the neuromorphic computing devices of the future.

6. Recommendations

In the future, our goal is to enable hardware for brain-inspired computing, rather than just neuromorphic computing, which only mimics the brain but does not capture the true learning and cognitive abilities of the brain.

The human brain is widely regarded as the ultimate computing engine with extremely high energy efficiency, reliability, and learning and cognitive capabilities. Although the field of neuromorphic computing has made tremendous strides in the last decade, current neuromorphic systems have yet to demonstrate the cognitive functionality of mainstream artificial intelligence (AI) methods (e.g., deep nets) or the energy efficiency approaching that of the brain. The key reasons for this are (1) mismatch between the properties of the mainstream devices and architectures and those of the brain and (2) lack of heterogeneous integration strategies to deliver neuromorphic systems at scale. In the future, we will develop a unified "materials-to-systems" methodology with a strong emphasis on hardware-software co-design to realize brain-inspired computing systems with 1000x improvements in speed, energy, and cognitive abilities compared to state-of-the-art approaches. Such hardware is expected to support advanced data analytics capabilities with higher processing performance and better system scalability for many advanced AI applications of the future, which are expected to profoundly impact human lives in areas such as energy, advanced manufacturing, public health, security, and business.

We plan to develop a diverse library of spatially compact functional units, referred to as neuroprimitives, that display neuro-inspired functionality and generate simple yet complete functional behaviors, e.g., finite state machine. Toward this, physical phenomena of ferreoelectricity, spin dynamics in magnetically ordered materials, and topological states in matter, will be investigated and modeled. These material platforms and devices will be designed to display intelligent neuro-mimetic dynamics including stochastic oscillations, memristive switching, high-frequency signal modulation, and diffusion-dominated transport of information to approach the thermodynamic limits on energy efficiency. These neuromprimitives can be employed by to construct circuit macros, also known as neuromacros. As examples of the neuromacros we will implement

- **Self-organizing logic gates using dipolarly coupled magnets** → time nonlocality to simultaneously process and store data for hard combinatorial optimization problems
- **Neuronal networks for population coding using stochastic magnetic oscillators and spike generators** → higher reliability at ultra-low energy cost.

On the architecture side, our goal is to exploit the properties of intelligent materials and devices to minimize communication energy and latency. We plan to explore several architectural configurations, such as dataflow-based architectures with minimal processing elements, data-parallel architectures that capture parallel neurons in our brain, and analog communication that can more efficiently encode synapse firings (closer to the brain's analog firings). We aim to identify the correct architectural design that can best propel the development of hardware at scale, and that can be used as an underpinning for a wide range of neuromorphic applications (e.g., real-time visual identification, image reconstruction, decision making, replication/replacement of human senses).

7. References

[1] Mead, Carver. "Neuromorphic electronic systems." *Proceedings of the IEEE* 78.10 (1990): 1629-1636.

[2] Zhou, Jing, and Jingsheng Chen. "Prospect of spintronics in neuromorphic computing." *Advanced Electronic Materials* 7.9 (2021): 2100465.

[3] Krzysteczko, Patryk, et al. "The memristive magnetic tunnel junction as a nanoscopic synapse-neuron system." *Advanced Materials* 24.6 (2012): 762-766.

[4] Furuta, Taishi, et al. "Macromagnetic simulation for reservoir computing utilizing spin dynamics in magnetic tunnel junctions." *Physical Review Applied* 10.3 (2018): 034063.

[5] Jungwirth, Tomas, et al. "Antiferromagnetic spintronics." *Nature nanotechnology* 11.3 (2016): 231-241.

[6] Baltz, Vincent, et al. "Antiferromagnetic spintronics." *Reviews of Modern Physics* 90.1 (2018): 015005.

[7] Fukami, Shunsuke, Virginia O. Lorenz, and Olena Gomonay. "Antiferromagnetic spintronics." *Journal of Applied Physics* 128.7 (2020): 070401.

[8] Kurenkov, Aleksandr, Shunsuke Fukami, and Hideo Ohno. "Neuromorphic computing with antiferromagnetic spintronics." *Journal of Applied Physics* 128.1 (2020): 010902.

[9] Locatelli, Nicolas, et al. "Spintronic devices as key elements for energy-efficient neuroinspired architectures." *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2015.

[10] Mizrahi, Alice, et al. "Magnetic stochastic oscillators: Noise-induced synchronization to underthreshold excitation and comprehensive compact model." *IEEE Transactions on Magnetics* 51.11 (2015): 1-4.

[11] Mehonic, Adnan, and Anthony J. Kenyon. "Brain-inspired computing needs a master plan." *Nature* 604.7905 (2022): 255-260.

[12] STEINBUCH, Karl. "DIE LERNMATRIX-THE BEGINNING OF ASSOCIATIVE MEMORIES." *Advanced Neural Computers*. North-Holland, 1990. 21-29.

[13] MANCINI, LORENZO. "The Hopfield model: study of the basins of attractions and the matrix factorization problem."

[14] Rai, Sadhana, and Basavaraj Talawar. "Nonvolatile Memory Technologies: Characteristics, Deployment, and Research Challenges." *Frontiers of Quality Electronic Design (QED) AI, IoT and Hardware Security* (2023): 137-173.

[15] Stiles, Mark D., and Jacques Miltat. "Spin-transfer torque and dynamics." *Spin dynamics in confined magnetic structures III* (2006): 225-308.

[16] Gambardella, Pietro, and Ioan Mihai Miron. "Current-induced spin–orbit torques." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 369.1948 (2011): 3175-3197.

[17] Fiebig, Manfred. "Revival of the magnetoelectric effect." *Journal of physics D: applied physics* 38.8 (2005): R123.

[18] Moriyama, Takahiro, et al. "Spin torque control of antiferromagnetic moments in NiO." *Scientific reports* 8.1 (2018): 14167.

[19] Shiratsuchi, Yu, Kentaro Toyoki, and Ryoichi Nakatani. "Magnetoelectric control of antiferromagnetic domain state in Cr2O3 thin film." *Journal of Physics: Condensed Matter* 33.24 (2021): 243001.

[20] Markou, A., et al. "Noncollinear antiferromagnetic Mn 3 Sn films." *Physical Review Materials* 2.5 (2018): 051001.
[21] Parthasarathy, Arun, et al. "Precessional spin-torque dynamics in biaxial antiferromagnets." *Physical Review B* 103.2 (2021): 024450.
[22] Shukla, Ankit, and Shaloo Rakheja. "Spin-Torque-Driven Terahertz Auto-Oscillations in Noncollinear Coplanar Antiferromagnets." *Physical Review Applied* 17.3 (2022): 034037.
[23] Takeuchi, Yutaro, et al. "Chiral-spin rotation of non-collinear antiferromagnet by spin–orbit torque." *Nature Materials* 20.10 (2021): 1364-1370.
[24] Higo, Tomoya, et al. "Anomalous Hall effect in thin films of the Weyl antiferromagnet Mn3Sn." *Applied Physics Letters* 113.20 (2018): 202402.
[25] Wang, Chao, et al. "Compact model of Dzyaloshinskii domain wall motion-based MTJ for spin neural networks." *IEEE Transactions on Electron Devices* 67.6 (2020): 2621-2626.
[26] Steinhoegl, W., et al. "Scaling laws for the resistivity increase of sub-100 nm interconnects." *International Conference on Simulation of Semiconductor Processes and Devices, 2003. SISPAD 2003.*. IEEE, 2003.
[27] Dutta, Shibesh, et al. "Finite size effects in highly scaled ruthenium interconnects." *IEEE Electron Device Letters* 39.2 (2018): 268-271.
[28] Wong, Shyh-Chyi, Gwo-Yann Lee, and Dye-Jyun Ma. "Modeling of interconnect capacitance, delay, and crosstalk in VLSI." *IEEE Transactions on semiconductor manufacturing* 13.1 (2000): 108-111.
[29] Nikonov, Dmitri E., and Ian A. Young. "Benchmarking delay and energy of neural inference circuits." *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* 5.2 (2019): 75-84.
[30] Cheng, Hsin-Pai, et al. "Understanding the design of IBM neurosynaptic system and its tradeoffs: A user perspective." *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*. IEEE, 2017.

8. Appendix A: Publications resulting from this research

Papers that resulted from this work:

- Shukla, Ankit, and Shaloo Rakheja. "Spin-Torque-Driven Terahertz Auto-Oscillations in Noncollinear Coplanar Antiferromagnets." *Physical Review Applied* 17.3 (2022): 034037.
- Cruz-Camacho, E., Qian, S., Shukla, A., McGlohon, N., Rakheja, S., & Carothers, C. D. (2022, June). Evaluating Performance of Spintronics-Based Spiking Neural Network Chips using Parallel Discrete Event Simulation. In *Proceedings of the 2022 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation* (pp. 69-80).
- Shukla, Ankit, and Shaloo Rakheja. "Terahertz auto oscillations in non-collinear coplanar metallic antiferromagnets." *2021 Device Research Conference (DRC)*. IEEE, 2021.
- Shukla, A., & Rakheja, S. (2021). Spin-torque driven self-oscillations in non-collinear coplanar antiferromagnets. In *APS March Meeting Abstracts* (Vol. 2021, pp. F38-010).

Papers currently under review:

- Cruz-Camacho, E., Qian, S., Shukla, A., McGlohon, N., Rakheja, S., & Carothers, C. D. (2022, June). Evaluating Performance of Spintronics-Based Spiking Neural Network Chips using Parallel Discrete Event Simulation. Submitted to ACM Transactions on Modeling and Computer Simulation of Devices.

Papers current under preparation:

- Shukla, Ankit, Siyuan Qian, and Shaloo Rakheja. Pulsed response of noncollinear coplanar metallic antiferromagnets. APL Materials.
- Qian, Siyuan and Shaloo Rakheja. Numerical and compact modeling of spintronic oscillators. IEEE Transactions on Magnetics.

9. List of Acronyms

AFM: Antiferromagnet
AFRL: Air Force Research Laboratory
AHE: Anomalous Hall effect
ANN: Artificial neural network
CIFAR: Canadian Institute for Advanced Research
CMOS: Complementary Metal Oxide Semiconductor
FM: Ferromagnet
HRS: High resistance state
I&F: Integrate and fire
LRS: Low resistance state
ML: Machine Learning
MNIST: Modified National Institute of Standards and Technology
MTJ: Magnetic tunnel junction
NN: Neural Network
R&D: Research & Development
RAM: Random Access Memory
SNN: Spiking Neural Network
SOT: Spin-orbit torque
SPICE: Simulation Program with Integrated Circuit Emphasis
STT: Spin-transfer torque
TAMR: Tunneling anisotropic magnetoresistance
TMR: Tunneling magnetoresistance
UIUC: University of Illinois at Urbana-Champaign