



# VCU

Virginia Commonwealth University  
VCU Scholars Compass

---

Theses and Dissertations

Graduate School


---

2024

## Energy Efficient Spintronic Devices for Non-volatile Memory and Hardware AI

Walid Al Misba  
*Virginia Commonwealth University*

Follow this and additional works at: <https://scholarscompass.vcu.edu/etd>

 Part of the [Other Electrical and Computer Engineering Commons](#), and the [Other Mechanical Engineering Commons](#)

© The Author

---

Downloaded from

<https://scholarscompass.vcu.edu/etd/7810>

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact [libcompass@vcu.edu](mailto:libcompass@vcu.edu).

# **Energy Efficient Spintronic Devices for Non-volatile Memory and Hardware AI**

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Engineering at Virginia Commonwealth University.

by

**Walid Al Misba**

Master of Science, Tuskegee University, Electrical Engineering, 2018

**Advisor:** Jayasimha Atulasimha, Ph.D.

Professor, Department of Mechanical and Nuclear Engineering Virginia Commonwealth University

Virginia Commonwealth University

Richmond, Virginia

August 2024

*This dissertation is dedicated to my parents for their love and support.*

## **Article I: Abstract:**

Nanomagnetic devices have emerged as a promising alternative to conventional complementary metal-oxide-semiconductor (CMOS) devices due to their low energy dissipation and inherent non-volatility. However, the widespread adoption of these devices requires high-density, high-speed, reliable, scalable, and energy-efficient technologies. This thesis investigates the use of nanomagnetic memory devices as both conventional Boolean memory and multistate memory for hardware AI applications.

Magnetic tunnel junctions (MTJs) are nanomagnetic memory devices that can be switched reliably and energy-efficiently using stress-mediated switching. However, realistic material inhomogeneity and scalability pose challenges for stress-mediated switching of MTJs scaled to lateral dimensions below 50 nm. We demonstrate that resonant excitation of MTJs using surface acoustic waves can effectively address the scaling challenges associated with the high anisotropies required for sub-50 nm lateral dimensions.

For hardware AI applications, in-memory computing using non-volatile devices offers energy-efficient alternatives to traditional von Neumann computing by bridging the gap between computation and memory units. This approach addresses the significant energy inefficiency caused by the constant shuttling of data between these units. We implemented classification hardware based on deep neural networks (DNNs) using in-memory computing with domain wall (DW)-based non-volatile synaptic devices. Our research shows these DW devices can achieve multistate memories using energy efficient voltage control, however, with low resolution and stochasticity. By devising a strategy to address device stochasticity and limited precision during DNN training, we demonstrated that competitive accuracy can be achieved compared to 32-bit precision synapses, with at least two orders of magnitude energy savings. Additionally, spintronic reservoirs leveraging the rich magnetization dynamics of confined skyrmions are explored for autonomous time series prediction, offering an efficient alternative to recurrent neural networks (RNNs) and long short-term memory (LSTM) networks for edge AI applications.

Moreover, we investigate multiferroic structures and rare earth iron garnets (REIGs) for their potential in electric field control of magnetization, offering the possibility of ultra-low energy dissipation in high-density magnetic data storage applications. Our research includes the characterization of novel materials, such as bismuth-substituted yttrium iron garnet (Bi-YIG) on piezoelectric substrates, demonstrating voltage-induced control of magnetic properties. We also explore the static exchange coupling of magnetic multilayers consisting of thulium iron garnet (TmIG) with perpendicular

magnetic anisotropy and cobalt iron boron (CoFeB). Incorporating REIGs into MTJs is challenging due to their insulating nature. However, coupling REIGs to a magnetic metal could enable readout of the REIG magnetization if the metal forms the free layer of an MTJ. Additionally, we explore the dynamic coupling between REIGs and CoFeB driven by spin current exchange, excited by ferromagnetic resonance, revealing that the magnetic damping parameter can be controlled by non-local relaxation and coupling.

In summary, this thesis contributes to the development of energy-efficient spintronic devices for non-volatile memory and hardware AI applications, addressing challenges in scalability, stochasticity, and material characterization, and paving the way for future advancements in these cutting-edge technologies.

## **Article II: Acknowledgement**

First and foremost, I would like to express my deepest gratitude to my thesis advisor, Dr. Jayasimha Atulasimha, for his unwavering support, guidance, and encouragement throughout my research journey. His invaluable insights, expertise, and mentorship have been instrumental in shaping my work and helping me grow as a researcher.

I extend my sincere thanks to the members of my thesis committee, Dr. Supriyo Bandyopadhyay, Dr. Milos Manic, Dr. Ravi L. Hadimani, Dr. Reza Mohammadi, and Dr. Daniel B. Gopman, for their valuable feedback, constructive criticism, and discussions, which have greatly contributed to the improvement of my thesis. Special thanks to Dr. Gopman, with whom I had the opportunity to work on several research projects and use the facilities in his lab at NIST, Gaithersburg, MD.

My heartfelt appreciation goes to my colleagues and fellow researchers at M3 Lab for creating a stimulating and supportive research environment. Special thanks to my colleagues Dr. Dhritiman Bhattacharya and Md. Mahadi Rajib for their friendship, encouragement, and the countless engaging discussions we had. I would like to offer my gratitude to the scientists and staff at the Nano-Characterization Center (NCC) and Virginia Microelectronics Center for their support and guidance during the experiments I performed there, especially to Dr. Carl Mayer and Dr. Dmitry Pestov. I would also like to thank my collaborators Dr. Caroline A. Ross, Dr. Damien Querlioz, Dr. Brajesh K. Kaushik, Dr. Joseph Friedman, and Dr. Kai Liu.

I acknowledge the financial support provided by the VCU College of Engineering, the Virginia Commonwealth Cyber Initiative (CCI), and the US National Science Foundation.

I am deeply indebted to my family, especially my father, Khalid Al Azam, my mother, Khurshida Khatun, my sister, Tasnim Tabassum, and my wife, Dr. Marzia Khanam, for their unconditional love, unwavering support, and the sacrifices they have made to help me pursue my dreams. Their belief in me has been a constant source of motivation throughout this journey.

Finally, I would like to thank all the individuals who have directly or indirectly contributed to the successful completion of this thesis. Your support and encouragement have been invaluable.

## Table of Contents

Abstract.....	ii
Acknowledgement.....	iv
List of figures.....	ix
List of Tables.....	xix
Chapter 1: Introduction.....	1
1.1 Background.....	4
1.1.1 Magnetic tunnel junctions.....	4
1.1.2. Magnetic domain wall racetrack.....	4
1.1.3 Deep neural networks.....	5
1.1.4 Reservoir computing.....	8
1.1.5 Imaging techniques for magnetic structure characterization .....	9
1.1.6 Microstructure and nanostructure patterning .....	12
1.1.7 Micromagnetic modeling of magnetization dynamics.....	13
1.2 Organization of this dissertation proposal.....	16
Chapter 2: Acoustic Wave Induced FMR Assisted Spin-Torque Switching of Perpendicular MTJs with Anisotropy Variation.....	18
2.1 Model .....	20
2.2 Results and discussion .....	21
2.3 Energy dissipation.....	26
2.4 Conclusions.....	30
Chapter 3: Voltage Controlled Energy Efficient Domain Wall Synapses with Stochastic Distribution of Quantized Weights in the Presence of Thermal Noise and Edge Roughness.....	35
3.1 Device architecture and simulation.....	36
3.2 Results and Discussion.....	40
3.2.1 Effect of edge roughness on domain wall motion.....	40
3.2.2 Non-thermal statistics due to different edge roughness profiles in different racetracks.....	45
3.2.3 Thermal statistics.....	49
3.2.4 Determination of synaptic state.....	50
3.3 Energy dissipation.....	52
3.4. Additional details.....	53



3.5 Conclusion.....	55
Chapter 4: Energy Efficient Learning with Low Resolution Stochastic Domain Wall Synapse for Deep Neural Networks.....	62
4.1. Methods.....	65
4.2. Learning of fully connected DNN with domain wall nano-synapse.....	68
4.3. Results and discussions.....	76
4.4. Energy dissipation.....	81
4.5 Additional details.....	84
4.6 Conclusion.....	92
Chapter 5: Spintronic Physical Reservoir for Autonomous Prediction and Long-Term Household Energy Load Forecasting.....	99
5.1. Model.....	103
5.2. Reservoir setup.....	107
5.3. Results and discussions.....	112
5.4. Energy dissipation.....	116
5.5 Conclusion.....	117
Chapter 6: Magnetic Anisotropy Modulation in Bismuth Substituted Yttrium Iron Garnet with Voltage Controlled Strain.....	124
6.1 Sample growth and characterization.....	126
6.2 Magnetic Hysteresis Modulation with Strain.....	127
6.3 Magnetization reversal with electric field.....	131
6.4 Summary and conclusion.....	133
Chapter 7: Interfacial exchange and magnetostatic coupling in a CoFeB/perpendicular ferrimagnetic Thulium Iron Garnet heterostructure.....	139
7.1 Sample preparation and characterization.....	141
7.2. Interlayer exchange and magnetostatic coupling.....	145
7.3. Micromagnetic simulation.....	151
7.4. Summary and conclusion.....	154
Chapter 8: Dynamic Coupling in a CoFeB/Perpendicular Ferrimagnetic Thulium Iron Garnet Heterostructure.....	161
8.1 Sample preparation.....	161

8.2. Results and discussion.....	161
8.3. Conclusion.....	168
Chapter 9: Future Works and Conclusion.....	170
9.1. DW-MTJ for complex neural network architectures and tasks.....	171
9.2. Demonstration of DW-MTJ based DNN.....	171
A1: List of journals.....	173
A2: List of conferences.....	174

### Article III: List of figures

Figure 1-1 Schematics of magnetic tunnel junction with magnetizations orientations of the ferromagnets oriented along in-plane and out of plane.....	4
Figure 1-2 Domain wall racetrack memory and corresponding DW-MTJ. The resistance states of the devices can be controlled by applying different strength current pulses.....	5
Figure 1-3 Schematic of a DNN architecture showing the input, hidden and output layers neurons.....	6
Figure 1-4 Quantization of DNN weights. The corresponding equations shown at top and bottom for hidden layer neurons are for conventional DNN neurons and quantization approach respectively.....	7
Figure 1-5 Schematic of a reservoir showing the recurrent connections among neurons.....	8
Figure 1-6 Magnetic skyrmion acts as a physical reservoir. The magnetization of the patterned skyrmion oscillates due to modulation of perpendicular magnetic anisotropy (PMA).....	9
Figure 1-7 Schematics showing the a. polar b. longitudinal and c. transverse geometries of magneto-optical effects.....	10
Figure 1-8 Schematic showing the operation of a MOKE microscope.....	11
Figure 1-9 Schematic showing the operation of VSM.....	12
Figure 1-10 Schematics showing operation steps of magnetic microstructures (maskless photolithography) or nanostructures (e-beam lithography) patterning.....	13
Figure 2-1 a. MTJ arrays and SAW electrode over piezoelectric substrate b. initial magnetization state of the inhomogeneous (i.e. granular) free layer c. application of SAW induces different angle precession and the resulting incoherency reduces the net magnetization, M d. final magnetization state after application of STT current.....	20
Figure 2-2 Grain distribution of the inhomogeneous nanomagnet using Voronoi tessellation. The colormap is showing the values of the first order anisotropy constant, $K_1$ (see eq. 4) for different grains. The values of $K_1$ is chosen using gaussian distribution of mean $4.5837 \times 10^5 J/m^3$ and standard deviation of 5%.....	21
Figure 2-3 a. Evolution of magnetization deflection ( $\theta$ ) from the perpendicular z direction for 100 MPa 13.2 GHz SAW b. Average polar angle deflection ( $\theta$ ) for different excitation frequency for varying stress	

amplitude. Resonance point (highest deflection) shifts towards lower frequency with increasing stress.....22

FIG. 2-4. Evolution of magnetization states with time (shown at the left corner in nanosecond) under the excitation of SAW in a. homogeneous and b. inhomogeneous nanomagnets at T=0 K and c. homogeneous and d. inhomogeneous nanomagnets at T=300 K.....23

FIG. 2-5. Spin configuration of inhomogeneous nanomagnet at T=300 K at different snapshots in time (in nanoseconds) in left corner and average magnetization polar angle ( $\langle\theta\rangle$ ) a. polar angle ( $\theta$ ) in different regions implies incoherency for both high and low average magnetization polar angle ( $\langle\theta\rangle$ ) b. in-plane azimuth angle ( $\phi$ ) for individual spins shows that the spins are almost in phase while precessing in high polar angle ( $\langle\theta\rangle$ ) but out of phase while precess in low polar angle.....24

FIG. 2-6. Switching trajectories of SAW assisted STT a. without grains b. with grains. The inset shows the grain configuration of the nanomagnet.....25

FIG. 2-7. Switching probability vs. STT current at different resonant SAW amplitudes. Error bar shown in the inset corresponds to one of the data points (100 MPa SAW grain case at  $1.2 \times 10^{11}$  A/m<sup>2</sup> current) for 1000 simulations.....26

Figure 2-8 Nanomagnet array and SAW electrode (IDT) patterned on top of the piezoelectric substrate. The nanomagnet center to center distance (R) and IDT beamwidth (W) are shown.....27

Figure 2-9 Displacement  $u_y$  versus distance  $y$  curve (in blue) in the delay line of the piezoelectric. Maximum strain is generated when the nanomagnet center is at any zero-crossing.....28

Figure 3-1 (a) Proposed device stack where the nanoscale racetrack act as the magnetic free layer of the MTJ. DW in the racetrack moves when a current is applied to the heavy metal layer underneath the racetrack (b) Stress generation mechanism in rough edge racetrack when a voltage is applied across the piezoelectric. (c) Implementation of layers of DNN with DW based synaptic devices. The devices are arranged in crossbar to provide programmable conductance equivalent to the DNN weights.....39

Figure 3-2 Stabilized DW position distribution for 30 different nanowires for two different rms edge roughness values. The PMA of the nanowires is  $K_u = 7.5 \times 10^5$  J/m<sup>3</sup> which is modulated to  $K_u = 8.0 \times 10^5$  J/m<sup>3</sup> with a voltage pulse of 1.2 ns. SOT current pulse of  $24 \times 10^{10}$  A/m<sup>2</sup> is also applied simultaneously for 1.2 ns. All the DWs are nucleated at 450 nm as seen from the top panel. After

withdrawing the voltage and current pulse simultaneously the system is relaxed for 10 ns to determine the stabilized DW positions.....41

Figure 3-3 DW depinning current with respect to the relative distance,  $x_d$  between the pinning location and the DW starting position for four different PMA (Ku). Racetrack of dimension 500 nm x 50 nm with the DW and pinning site (triangular notch) is shown above with a sketch of demagnetization potential.....44

Figure 3-4 a. Initial pinning position of the DW in a PMA rough edge racetrack b. dependence of the DW depinning current on the anisotropy coefficient when the DW in racetrack 3-4a is in the initial pinning position c. DW positions with time in racetrack 3-4a for a fixed duration and amplitude current pulse exerting SOT and different stresses (different Ku). The SOT and stress are withdrawn at 1.2 ns. For different stresses respective DWs travel different distances and get pinned to different locations.....47

Figure 3-5 a-e. Equilibrium DW positions for 40 different racetracks at T=0 K for a fixed SOT and different stresses correspond to Ku values of 8.0, 7.8, 7.5, 7.3 and 7.0 ( $\times 10^5$ )  $J/m^3$ . For each figure in 3-5(a-e) a Gaussian distribution plot is overlaid having a mean and standard deviation identical to the data used to create the bins f. combined plot of 3-5(a-e) shows different mean positions for different  $K_u$  values.....48

Figure 3-6 a-e. Equilibrium DW positions for one racetrack at T=300K for a fixed SOT and different stresses correspond to Ku values of 8.0, 7.8, 7.5, 7.3 and 7.0 ( $\times 10^5$ )  $J/m^3$ . For each figure in 3-6(a-e) a Gaussian distribution plot is overlaid having a mean and standard deviation identical to the data used to create the bins f. combined plot of 3-6(a-e) shows different mean positions for different  $K_u$  values.....50

Figure 3-7 (a) Cumulative probability of device conductance for 5 different programming conditions (different  $K_u$ ) implementing a 5-state stochastic synapse. The black dotted lines represent the 5 target conductances for the 5-state synapse. The red dotted lines represent the boundaries of each state to ensure that no overlap happens between adjacent states. (b) Cumulative probability of device conductance for 3-state synapse. The red dotted lines represent the state boundaries. of each state. For 3-state synapse the width of state boundary is high so one state can be programmed with a smaller number of attempts.....51

Figure 3-8 A rectangular region of dimension  $60 \times 50 \times 1 \text{ nm}^3$  marked in red is centered around the DW in a racetrack. The energy densities are calculated for the red rectangular region by changing the cell sizes of the simulation. The computed energy densities are shown in table 3-2.....53

Figure 3-9 Distribution of equilibrium domain wall position in 40 different racetracks for anisotropy coefficient of  $8.0 \times 10^5 \text{ J/m}^3$  for cell size (a)  $1 \times 1 \times 1 \text{ nm}^3$  and (b)  $2 \times 2 \times 1 \text{ nm}^3$ .....54

Figure 3-10 Distribution of equilibrium domain wall position in 40 different racetracks for anisotropy coefficient of  $7.0 \times 10^5 \text{ J/m}^3$  for cell size (a)  $1 \times 1 \times 1 \text{ nm}^3$  and (b)  $2 \times 2 \times 1 \text{ nm}^3$ .....55

Figure 4-1 a. Micromagnetic configuration of a  $\sim 2 \text{ nm}$  rms rough edge racetrack with perpendicular magnetic anisotropy (PMA). Engineered notches are placed regularly at  $75 \text{ nm}$  intervals. A DW is initialized at a notch  $60 \text{ nm}$  from the left of the racetrack. b. DW based nano-synapse device: racetrack ferromagnet/insulator/reference ferromagnet (MTJ) on top of a heavy metal layer on a piezoelectric substrate. A fixed amplitude current pulse,  $J$  in the heavy metal layer along with different amplitude voltage pulse,  $V$  across the piezoelectric changes the perpendicular anisotropy (PMA or  $K_u$  constant) of the racetrack and translates the DW (shown in red rectangle) into different longitudinal positions along the racetrack. c. Distribution of equilibrium DW positions in the racetrack (shown in Fig. 4-1a) at room temperature  $T=300\text{K}$  for a fixed SOT generating current pulse of  $J = 35 \times 10^{10} \text{ A/m}^2$  applied for  $1 \text{ ns}$  and five different PMA coefficients,  $K_u$  (corresponds to five different programming voltages). Different mean positions for different  $K_u$  implies that 5-state, 3-state or 2-state stochastic synapses can be implemented by choosing 5,3 or 2 different programming voltages. d. distribution of average perpendicular magnetization,  $\langle m_z \rangle$  (which is equivalent to DNN weights according to Eq. 4) derived directly from DW positions.....67

Figure 4-2 a. Architecture of a fully connected deep neural network (DNN). Any neuron  $i$  in layer  $l$  is connected to neuron  $j$  in layer  $l + 1$  with synaptic weight  $W_{ij}$ . At forward propagation, inputs to neuron  $j$  are summed and passed through an activation function  $f$  to generate its output,  $x_j^{l+1}$ . At backward propagation, errors of layer  $l + 1$  neurons are back propagated to calculate the error  $\delta_i^l$  of neuron  $i$  in layer  $l$ . b. Implementation of the DNN in crossbar with DW devices. The peripheral circuit and the crossbar shown here implements DNN functionalities of only one layer (“ $l$ ”) and the next layer (“ $l + 1$ ”) and the number of rows in the crossbar are determined by the number of neurons in layer,  $l$  and the number of columns by the number of neurons in layer,  $l + 1$  (shaded region in Fig. 4-2a). At each cross point of the crossbar there is a DW device with conductance  $G_{ij}^{synapse}$  and a parallel conductance  $G_p$ . The effective conductance at each cross point is equivalent to the DNN weights  $W_{ij}$

such that  $G_{ij} = \mu W_{ij}$ . Inputs and errors of neurons are scaled to voltages before feeding them into the crossbar. The flow of the training algorithm is shown at the right-hand side of the crossbar. For each of the DW devices there is a corresponding high precision weight (real weight) that is stored in a separate high precision memory unit. These high precision weights are updated after a forward and backward pass before passing it through a quantizer (i.e., 2, 3 or 5-level quantization, depending on the number of states of the device). The DW device conductances,  $G_{ij}$  (or the corresponding device weights,  $W_{ij} = \frac{G_{ij}}{\mu}$ ) are updated when they fall outside the prescribed range of the target quantized weights,  $W_q$ .....70

Figure 4-3 Online training accuracy and online testing accuracy for DNNs with different state DW devices for two different noise tolerance margins. These accuracies are compared with a DNN trained and tested with full precision weights and no stochasticity (baseline accuracy) a. and b. show the online training accuracies with the numbers of epochs for  $\alpha = 0.15$  and  $0.25$  respectively. c. and d. show online testing accuracies with numbers of epochs for  $\alpha = 0.15$  and  $0.25$  respectively.....77

Figure 4-4 Comparison of the total number of programmed weights with the number of training epochs for different networks. A significantly lower number of weights are updated during the proposed online training compared to the floating precision weight network of the same architecture.....78

Figure 4-5 Weight evolution of high precision weight, quantized weight and the DW device weight during the first few training images for two different noise tolerance margin a.  $\alpha = 0.15$  b.  $\alpha = 0.25$ . The synaptic weight shown here is connected between the neurons located in hidden layers 2 and 3.....79

Figure 4-6 Testing accuracy comparison of 3-state and 5-state DW device based DNNs for different ex-situ training algorithms with a programming noise tolerance margin of a.  $\alpha = 0.15$  b.  $\alpha = 0.25$ . The networks are trained offline with floating precision weights, quantized weights, and stochastic quantized weights derived from micromagnetic simulation. Each of the networks is trained with a total number of 10 epochs. Once training is done, the 3-state and 5-state DW devices are programmed based on the quantized value of trained weights prior to testing. For different training algorithms and for each of the test accuracies of DNNs built from 3- and 5-state hardware, a corresponding software test accuracy (no programming noise is considered, and exact trained weights are used for testing the DNN) is plotted side by side with green and yellow bar. The error bar seen in the figure is calculated from 10 different test trials. For both noise tolerance margins, the test accuracy is highest when the DNNs are trained with proposed training algorithm (quantized + stochastic) .....81

Figure 4-7 Cumulative probability of normalized DW device weights for 5-state device under different programming conditions denoted by different  $K_u$ . Black solid line represents the target quantized weights and the adjacent dotted red lines represent the programming noise tolerance margin of  $\alpha = 0.15$ .....83

Figure 4-8 Height maps show the a. training b. testing and c. average of training and testing accuracies of a 3-hidden layer DNN for varying topography of the network. The topographic features that are varied are the learning rate and the ratio of the number of neurons in a layer to the number of neurons in the previous layer or “layer size ratio”. The highlighted data point is the final topography chosen for our DW device based DNN as it gives high accuracies for a small number of synapses.....86

Figure 4-9 a. Racetrack of dimension  $600 \text{ nm} \times 60 \text{ nm} \times 1 \text{ nm}$  with rms edge roughness of  $\sim 2 \text{ nm}$  hosting a DW at an initial position of 60 nm from the left end. Engineered notches starting from 60 nm to the left of the racetrack are placed at a regular interval of 75 nm. b.-f. Distribution of equilibrium DW positions along the racetrack shown in Fig. 4-9a for different programming conditions represented by different PMA coefficient,  $K_u$  . The DWs are primarily pinned at or around the notches, thus the distribution become dominated by different notch locations for different programming conditions.....87

Figure 4-10 Offline testing accuracies for DNNs of different state DW devices with two different noise tolerance margins, a.  $\alpha = 0.15$  and b.  $\alpha = 0.25$  used during the training and programming of the devices. The accuracies are compared with a DNN trained and tested with 32-bit (floating) precision weights and no stochasticity (baseline accuracy). Error bar is calculated for a total of 10 different test trials.....88

Figure 4-11 a. Training accuracies b. Online testing accuracies and c. Offline testing accuracies for a 5-state DW device based DNN for two different noise tolerance margins of  $\alpha$ . The training accuracy and online testing accuracy does not change appreciably for different noise tolerance margins. Offline testing accuracy decreases with high noise tolerance margin due to the higher deviation of device weights during the programming of the devices.....89

Figure 4-12 Proposed in-situ training of two successive DNN layers (green shadowed region). The scope of operations that are performed in analog and digital domain during the training is shown in two different colored boxes.....90



Figure 5-1 a. A conventional reservoir computing system with input layer, reservoir block with recurrent connections among nodes and the output layer. b. The reservoir block is replaced by a set of patterned skyrmion devices where each of the ferromagnetic films with PMA host a single skyrmion. c. Stacks of a skyrmion device with metallic electrode and MTJ. d. Training of a skyrmion reservoir for autonomous prediction task. The temporal input data is mapped into voltage values which are applied to each of the skyrmion devices and the responses are collected. The responses are read at regular interval and the read-out values act as the virtual node as represented by  $r_i^j$ . The states of the nodes (or reservoir responses) are used to predict the next time step value of the input time series. The weights are trained by computing the error of the predicted and target values and accomplished with simple pseudoinverse operation. e. During testing, the predicted output value is directly fed as input to the reservoir in order to perform multi-step autonomous prediction.....105

Figure 5-2 a. Three ferromagnetic thin films each hosting a magnetic skyrmion worked as the RC. The responses of the respective skyrmion devices are shown side by side when the thin films are perturbed by the inputs of MG time series from time-step 231 to 235 (inputs are mapped into voltage pulse amplitudes). The PMA modulation by the input voltage pulses is shown in orange color. The virtual nodes are marked in red diamond.....109

Figure 5-3 a. Long term autonomous prediction of chaotic MG time series with skyrmion reservoir. The dataset is trained with 31-400 time-step data. The reservoir is tasked to predict the next 30-time step data from 402-431. The overlapping of the predicted test data with actual label suggests accurate prediction b. prediction trend for MG time series with 2-layer deep sequence to sequence LSTM architecture. The LSTM is able to accurately predict the trend. Although, RMSE magnitude of the LSTM is lower than the reservoir, the prediction errors for both of the predictions remain extremely small. c. Phase diagram of the chaotic MG attractor during the training with reservoir. The predicted training data overlapped with the actual label implying the efficacy of the ridge regression training. d. Phase diagram of the reservoir for autonomous prediction. The superimposed plots suggest good prediction accuracy of the reservoir on test data.....113

Figure 5-4 a. Long-term autonomous forecasting of individual household active power load with proposed reservoir. The reservoir is trained with 21-261 hours of data and tasked with predicting the next 23 hours of data. The predicted trend closely follows the actual load level suggesting good prediction accuracy with the skyrmion reservoir. b. The same task is performed with 2-layer sequence to sequence LSTM architecture. Although the LSTM is able to capture the trend, the prediction accuracy is less than the proposed reservoir for the first several hours of prediction.....114

Figure 5-5 a. Hourly RMSE of the prediction accuracy for individual household load forecasting task for both of the proposed reservoir and LSTM. RMSE plots for two different long-term autonomous predictions, 262-284 hours and 286-308 hours are shown. The inset shows the prediction trend of the reservoir for prediction from 286 hour to 308 hours. The RMSE plots indicate higher prediction accuracy of the proposed reservoir compared to LSTM, even for much stochastic trend such as in 286-308 hours of data.....117

Figure 6-1 Sketch of resonance curves of an oscillator for different damping coefficients and increasing Q factors.....124

Figure 6-2 a. GIXD diffraction image shows Bi-YIG growth on SiO<sub>2</sub>/PMN-PT substrate. Data has been shifted vertically for clarity. b. Hysteresis loops taken via vibrating sample magnetometry of the BiYIG/SiO<sub>2</sub>/PMN-PT sample. The curves were measured out of plane (OP) and in plane (IP) to the sample surface.....127

Figure 6-3 Top surface SEM images of a) BiYIG/Si and b) BiYIG/SiO<sub>2</sub>/PMN-PT.....127

Figure 6-4 a,b. Hysteresis loops obtained from MOKE magnetometry for different voltages applied along the thickness of the heterostructure, PMN-PT/SiO<sub>2</sub>/BiYIG when the magnetic field is applied along the in-plane direction a.  $\hat{x}$ . b.  $\hat{y}$ . Black arrows indicate the trend for increased voltage. The inset in (a) shows a schematic of the heterostructure with the direction of the applied voltage, principal axes and the polar angle,  $\theta$  and azimuthal angle,  $\varphi$  of the BiYIG film magnetization,  $M$ . c. ratio of remanent and saturation magnetization vs the applied voltage for both in-plane directions,  $\hat{x}$  and  $\hat{y}$ . d. hysteresis loops as a function of voltage obtained from polar MOKE for the out of plane direction,  $\hat{z}$ .....129

Figure 6-5 a. Hysteresis curves with external fields applied along the in-plane direction x when the heterostructure is subjected to an applied voltage of 0 V and 450 V. b. longitudinal MOKE images showing magnetization reversal process. The corresponding field values for which the images are taken are also marked in the hysteresis loops. The upward (downward) arrows represent domain magnetizations that are pointed along the  $+\hat{x}$  ( $-\hat{x}$ ) directions. c. hysteresis loops for in-plane direction y for 0 V and 450 V and d. corresponding magnetization reversal images. The upward (downward) arrows represent domain magnetizations oriented along the or  $+\hat{y}$  ( $-\hat{y}$ ) directions.....130

Figure 6-6 MOKE images showing domain reversal along in-plane direction  $\hat{x} // [100]$  for varying voltages. The sample is first saturated with a -70 mT field and the reversal field is set to + 27 mT, then a voltage of 450 V was applied and reduced in steps of 50 V.....131

Figure 6-7 MOKE images showing domain reversal along in-plane direction  $\hat{y} // [01\bar{1}]$  for different voltages. The sample is first saturated with a -70 mT field and the reversal field is set to + 27 mT, the voltage is increased in 50 V steps from 0 V to 450 V.....132

Figure 6-8 Magnetolectric coefficient with respect to the electric field applied across the Bi-YIG/SiO2/PMN-PT heterostructure.....133

Figure 7-1 a. The prepared FM stack CoFeB(x)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5) are deposited on FI stack GGG/TmIG(40) to investigate FI/FM coupling. All the numbers in the parenthesis are thickness in nm. b. The FM stacks with CoFeB of x=1 nm and lower thickness have magnetization predominantly oriented along the out of plane direction whereas in stacks with x=3 nm and higher thickness the magnetizations are predominantly along in-plane. The resulting magnetizations in FI/FM stacks are canted with respect to out of plane and in-plane due to interfacial exchange coupling and magnetostatic interaction.....141

Figure 7-2 a. Out of plane and b. in-plane hysteresis loops for control samples, Si/CoFeB(x)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5) for variable thickness, x of CoFeB.....142

Figure 7-3 a. Out of plane and c. in-plane hysteresis loops of FI/FM heterostructure samples, GGG/TmIG(40)/CoFeB(x)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5), with CoFeB thickness, x= 1 nm or lower. d. out of plane and e. in-plane hysteresis loops with CoFeB thickness x=3 nm or higher. The out of plane and in-plane hysteresis loops of the pristine GGG/TmIG(40) sample are presented for comparison.....144

Figure 7-4 Polar MOKE magnetometry for FI only stack and FI/FM samples where the CoFeB thickness is varied to x= 1 nm, 3 nm and 4 nm. All the samples are first saturated to +50 mT and then the out of plane external fields are decreased. The snapshots are taken at fields of 0 mT, -1 mT, -3 mT, -5 mT and -13 mT. Labyrinthine domains with regular periodicity are observed for FI only stack, GGG/TmIG (40). Labyrinthine domains with dendritic patterns are observed with CoFeB overlayers. The periodicity of the domain wall periodicity decreases with increased CoFeB overlayer thickness implying higher magnetostatic coupling.....146

Figure 7-5 a. Domain wall pattern for TmIG, TmIG/CoFeB heterostructures with 1 nm and 4 nm variable CoFeB layer thickness at the corresponding fields. b. Radial FFT intensity to determine the periodicity of the domain wall for the respective samples.....147

Figure 7-6 FORC distributions for samples with a. FI-only b. control FM stack with x= 1 nm CoFeB and FI/FM heterostructures with variable CoFeB layer thickness of c. x= 1 nm d. x= 3 nm e. x= 4 nm for out of plane external fields,  $H_{\perp}$  . f. Family of FORC curves determined for different reversal fields,  $H_R$  for FI/FM sample with x= 1 nm CoFeB.....149

Figure 7-7 a. FORC distributions and b. family of FORC curves determined for different reversal fields for sample TmIG/CoFeB (3nm) with in-plane field direction. c. Schematic showing flux closure in heterostructure with CoFeB thickness, x=4 nm.....151

Figure 7-8 Hysteresis loops along the out of plane directions derived from micromagnetic simulations for the total heterostructure, top layer CoFeB and bottom layer TmIG are shown for samples with x=4 nm and x=1 nm. The micromagnetic snapshots for both of the samples for total magnetizations, top CoFeB layer and bottom TmIG layers magnetizations are shown. For the bottom TmIG layer, two snapshots taken at different depths of TmIG (contact to FM denotes the TmIG layer that is directly in contact with CoFeB and contact to sub. shows the TmIG layer that is in direct contact with the GGG substrate) .....153

Figure 7-9 Micromagnetic spin orientation in FI/FM heterostructures with x=4 nm and x= 1 nm along the cross-section of domain walls for out of plane external field,  $\mu_0 H_{\perp}=0$  mT and  $\mu_0 H_{\perp}=-3$  mT respectively. The number of spins is down sampled from actual micromagnetic simulations for better visibility.....154

Figure 8-1 a. FI/FM heterostructure, GGG/TmIG(40)/CoFeB(x)/W(0.4)/CoFeB(0.0)/MgO(1)/W(5). All the numbers in parenthesis are in nanometers. b. The schematic of the FI/FM stack under the application of microwave power, the applied field, H and the angle of the field,  $\theta$  are indicated in the figure. c. The FMR absorption data and the corresponding fitting with two Lorentzian for TmIG and CoFeB modes for FI/FM heterostructure with x= 1 nm CoFeB. The applied field angle for the FMR measurements is shown.....163

Figure 8-2 The FMR linewidths of the TmIG modes for pristine TmIG sample and the heterostructure samples with different thickness CoFeB are plotted as a function of the rf field frequencies. The corresponding line fits are shown where the slopes indicate the values of the damping coefficients.

Damping enhancement of the TmIG modes in heterostructures compared to the pristine TmIG sample are evident. The enhancement is highest for samples with $x=3$ nm CoFeB.....	165
Figure 8-3 Schematics of conditions where a. FM layer magnetization is resonantly excited b. FI layer magnetization is resonantly excited and c. both of the FM and FI layers' magnetizations are resonantly excited.....	166
Figure 8-4 a. Resonant fields vs the filed angle for the CoFeB and the TmIG modes in FM/FI heterostructure with $x=3$ nm. b. linewidths vs the filed angles of the two modes for the same heterostructures. The FMR measurements are carried out for rf field frequency of 14 GHz.....	167
Figure 8-5 a. Resonant fields vs the filed angle for the CoFeB and the TmIG modes in FM/FI heterostructure with $x=1$ nm. b. linewidths vs the filed angles of the two modes for the same heterostructures. The FMR measurements are carried out for rf field frequency of 11 GHz.....	168
Figure 9-1 Polar MOKE images during the DW nucleation and propagation along the racetrack stacks of Pt (3 nm)/Ta(3nm)/Co (0.3 nm)/Ni(0.7 nm)/Co(0.3 nm)/Ta(0.3 nm) exhibiting perpendicular magnetic anisotropy (PMA).....	172

**Article IV: List of Tables**

Table 2-1: FeGa material properties.....	21
Table 3-1: Material parameters used for the CoFe soft layer in the Pt/ CoFe/MgO heterostructure....	40
Table 3-2: Energy densities for simulations with different cell sizes.....	54
Table 4-1: Simulation parameter.....	66
Table 5-1: Simulation parameter.....	106
Table 7-1: Magnetic properties of control samples.....	142
Table 6-2. Magnetic properties of the FI-only stack and FI/FM heterostructures.....	144

## Chapter 1: Introduction

The exponential growth in computing power and the rapid advancement of artificial intelligence (AI) have led to unprecedented demands on energy consumption in the technology sector. Considering that the world's computing needs are increasing exponentially, and "computing" encompasses several aspects of our daily lives from a Google search, to using an Apple map while driving, or making a query on ChatGPT, it is vitally important to consider the energy and carbon footprint of computing. To put that into perspective, training a large language-based transformer model is responsible for 626,155 lbs of CO<sub>2</sub> emission, whereas a car's lifetime emission is 126,000 lbs [1]. Furthermore, the annual CO<sub>2</sub> emission due to data centers and high-performance computing is estimated to be 100 billion kilograms [2]. In summary, even if the energy cost per computation is small, the enormous volume of activity makes the energy needed and the environmental impact significant.

Another significant source of energy consumption is data processing, which has been centralized traditionally, with learning and inference performed in computers, servers, or high-performance data centers. However, an increasing number of Internet of Things (IoT) devices, from smart homes to industrial controls, processors in biomedical devices, self-driving cars, sensor networks, etc., need to process complex information at the edge. This necessitates inference at the edge, although learning can take place in servers given the large amount of data required for training and the relatively large computing power needed. Furthermore, some edge AI devices operate in constantly changing environments and need to learn in real-time. While prior training can be applied to the new environment (transfer learning), a few layers would need to be retrained. Thus, performing all inference and some learning at the edge can be extremely beneficial as it is energy efficient since no communication is needed with a server, has low latency, enables real-time learning and inference (needed in self-driving cars), and is more cyber-secure as the computation is performed at the edge. Furthermore, edge devices are constrained by limitations in hardware resources and available energy. Thus, the need of the hour is to enable inference and learning on edge devices where available energy and hardware resources are severely constrained.

As we approach the physical limits of traditional CMOS-based computing architectures, there is an urgent need for innovative solutions that can address the mounting challenges of energy efficiency, particularly in the domains of non-volatile memory and hardware AI implementations. This thesis explores the potential of energy-efficient spintronic devices to meet these critical needs, offering a promising pathway towards sustainable and high-performance computing systems.

The current computing landscape is dominated by the von Neumann architecture, which suffers from a fundamental bottleneck: the constant shuttling of data between processor and memory units. This inefficiency results in significant energy dissipation, a problem that is particularly acute in data-intensive tasks such as those required by deep neural networks (DNNs). Our research investigates various spintronic technologies that offer compelling alternatives to traditional volatile CMOS-based memory devices, aiming to bridge the gap between computation and memory storage.

A primary focus of our work is on magnetic tunnel junctions (MTJs), which have emerged as promising candidates for non-volatile nanomagnetic memory devices. While spin transfer torque (STT) has been the predominant mechanism for magnetization switching in MTJs, our research explores alternative strategies such as strain-mediated and voltage-mediated switching. These novel approaches have the potential to substantially reduce energy consumption, pushing the boundaries of energy efficiency in memory devices. Our work includes detailed micromagnetic simulations that incorporate realistic factors such as thermal noise, material inhomogeneities, and edge roughness, providing a comprehensive understanding of magnetization dynamics in these devices under various conditions.

Another key area of our research is the investigation of domain wall (DW) based magnetic structures for neuromorphic computing applications particularly for edge AI. By combining spin-orbit torque (SOT) with stress-induced modulation of perpendicular magnetic anisotropy (PMA), we demonstrate an innovative approach to controlling DW position in racetracks. This technique shows great promise for implementing energy-efficient synaptic devices that can be programmed in real-time, opening new avenues for hardware AI implementations.

A critical aspect of our work focuses on bridging the gap between the theoretical capabilities of DNNs and the practical limitations of hardware implementations using spintronic devices. We address this challenge through the development of quantized DNNs specifically designed for implementation with low-resolution non-volatile magnetic domain wall racetrack-based memory. Our approach allows for weights to be represented with as few as two states (binary or 1-bit) or a limited number of states (e.g., 3-state or 5-state), aligning well with the characteristics of DW devices. We employ a mixed-precision framework where synaptic weights are maintained at low resolution in the DW devices, while gradient calculations and weight updates are performed in high precision. This strategy enables us to achieve classification accuracies comparable to full-precision networks while leveraging the energy efficiency and non-volatility of DW-based memory. We also explore the scalability of our approach for complex data sets using convolutional neural networks (CNNs). We should mention that such device-aware training is equally applicable to other non-volatile technologies such as RRAM and PCRAM technologies and could be adopted for those equally well if future advances demand it.

Our research extends into the realm of recurrent neural networks (RNNs) and reservoir computing (RC), where we explore the use of spintronic devices, particularly those based on skyrmions, for temporal data processing tasks. We demonstrate that these systems can perform complex operations such as long-term prediction of chaotic time series and household power demand forecasting with high efficiency and low training cost.

Furthermore, we delved into the study of multiferroic structures and rare earth iron garnets (REIGs) for their potential in electric field control of magnetization. These materials offer the exciting possibility of ultra-low energy dissipation in high-density magnetic data storage applications, potentially outperforming current-based technologies by several orders of magnitude. Our research in this area includes the characterization of novel materials such as bismuth-substituted yttrium iron garnet (Bi-YIG) on piezoelectric substrates, demonstrating voltage-induced control of magnetic properties. In addition, we explore static exchange coupling and dynamic coupling phenomena of magnetic multilayers consisting of thulium iron garnet (TmIG) having perpendicular magnetic anisotropy of magnetoelastic origin and cobalt iron boron (CoFeB) which has been a standalone choice for MTJ. Coupling the REIG to a magnetic metal would enable readout of the magnetization of the REIG, which is otherwise challenging due to the insulating nature of REIGs, if the metal formed the free layer of an MTJ.

Throughout this thesis, we emphasize the critical importance of energy efficiency in spintronic devices for non-volatile memory and hardware AI applications. By leveraging novel materials, device architectures, and control mechanisms, we aim to push the boundaries of what is possible in energy-efficient computing. Our work contributes to the ongoing effort to overcome the limitations of traditional computing paradigms and pave the way for next-generation technologies that can meet the growing demands of AI and big data processing while minimizing energy consumption.

The implications of this research extend beyond academic interest, offering potential solutions to some of the most pressing challenges in modern computing. As we move towards an increasingly data-driven and AI-centric world, the need for energy-efficient computing solutions becomes ever more critical. Our work on spintronic devices aims to address this need, potentially enabling new classes of low-power, high-performance computing systems that could revolutionize fields ranging from edge computing and Internet of Things (IoT) devices to large-scale data centers and supercomputers.

In the following chapters, we will present detailed analyses of our various research directions, including device design, experimental results, theoretical modeling, and potential applications. Through this comprehensive exploration of energy-efficient spintronic devices, we hope to contribute significantly to the



ongoing transformation of computing technology and pave the way for a more sustainable and capable technological future.

## 1.1 Background:

### 1.1.1 Magnetic tunnel junctions:

Magnetic tunnel junction (MTJ) is the fundamental building block of spintronic memory and computing. It consists of a ferromagnetic layer (free layer) whose magnetizations can be switched, one ferromagnetic layer (fixed layer) whose magnetization orientation is fixed. The layers are separated by a thin insulator layer (typically MgO). A synthetic antiferromagnet (SAF) layer to cancel dipole coupling from this fixed layer can be stacked on top of the fixed layer, not shown in Fig. 1-1 for the sake of simplicity. When a voltage is applied across the MTJ, electrons can tunnel through the insulator layer where the tunneling probability is largely dependent on the relative orientation of the two ferromagnetic layers. When the ferromagnetic layers are in parallel orientation, the tunneling probability is high and current passing through the tunnel barrier experiences less resistance (denoted as memory bit, 0). Conversely, with anti-parallel configuration, the current experience highest resistance (memory bit, 1). This difference in resistance, known as tunneling magnetoresistance (TMR), is crucial for applications regarding magnetic memory and spintronic applications. Two types of MTJ geometries are common where the magnetic layers magnetizations are either in-plane or out-of-plane.

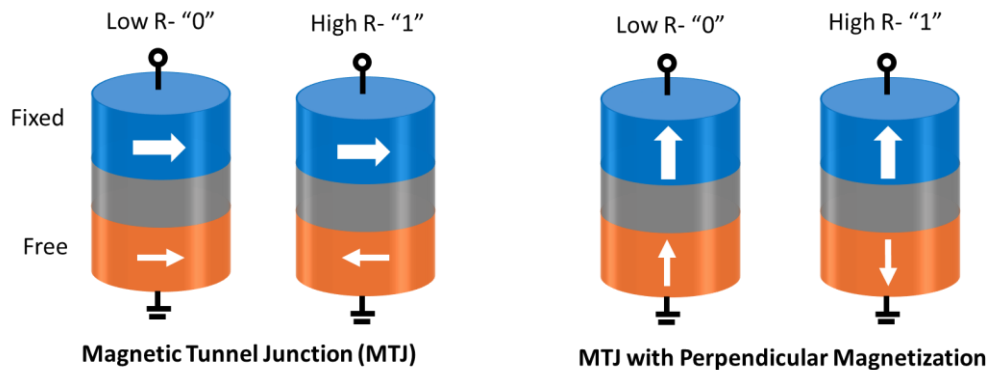


Figure 1-1 Schematics of magnetic tunnel junction with magnetizations orientations of the ferromagnets oriented along in-plane and out of plane.

### 1.1.2. Magnetic domain wall racetrack:

A magnetic domain wall racetrack is a memory technology that uses the movement of magnetic domain walls within a nanowire to store and manipulate information. In this system, bits of data are represented by the position of domain walls, which separate regions of opposite magnetic orientation. By applying current

pulses or magnetic fields, these domain walls can be moved along the racetrack, allowing for high-speed, high-density, and non-volatile data storage.

A domain wall magnetic tunnel junction (DW-MTJ) is a hybrid device where the racetrack hosting the magnetic domain wall acts as the free layer of the MTJ. An insulating tunnel barrier such as MgO and a fixed magnetic layer completes the DW-MTJ stack. As the domain walls are translated along the racetrack, the resistance of the devices varies resulting in multi-state memory, Fig. 1-2.

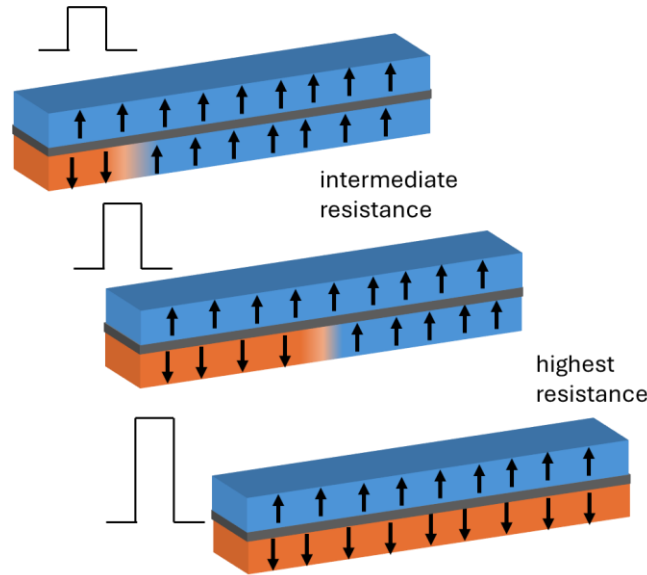


Figure 1-2 Domain wall racetrack memory and corresponding DW-MTJ. The resistance states of the devices can be controlled by applying different strength current pulses.

### 1.1.3 Deep neural networks:

A deep neural network (DNN) is a neural network architecture that is characterized by one or several hidden layers in between the input and output layers as shown in Fig. 1-3. Each layer of a DNN consists of several interconnected neurons that apply nonlinear transformations to the input. During the learning, the DNN's weights are adjusted based on the error between the predicted output and actual (true) output using backpropagation of errors and gradient decent methods.

The sets of equation explaining the forward propagation of the DNN are shown below:

$$\begin{aligned}
 a_j &= f \left( \sum_{i=1}^N w_{ij} x_i + b_j \right) \\
 o_k &= g \left( \sum_{j=1}^M w_{jk} a_j + b_k \right)
 \end{aligned} \tag{1}$$

$$f(x) = g(x) = \frac{1}{1 + e^{-x}}$$

Here,  $x$  is the inputs of neurons,  $a$  is the activations,  $w_{ij}$  are the weights and  $b$  is the biases. In Eq. 1, the activation function is sigmoid. Other functions such as hyperbolic tan, SoftMax and rectified linear unit (ReLU) can be used as activation functions depending on the architecture and model for achieving high accuracy.

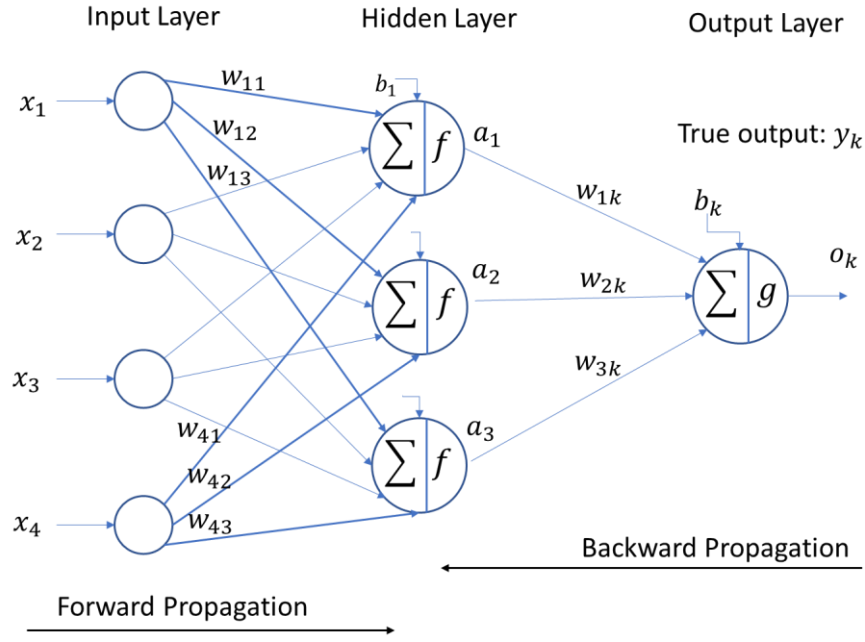


Figure 1-3 Schematic of a DNN architecture showing the input, hidden and output layers neurons.

The sets of equation explaining the backward propagation where the actual learning of the DNN is accomplished are the following:

$$\text{Cost: } C = \frac{1}{2} \sum_{k=1}^P (y_k - o_k)^2$$

$$\delta = \sum_{j=1}^M w_{jk} a_j + b_k$$

$$\frac{\partial C}{\partial w_{1k}} = \frac{\partial C}{\partial o_k} \cdot \frac{\partial o_k}{\partial \delta} \cdot \frac{\partial \delta}{\partial w_{1k}} \quad (2)$$

$$w_{1k}^{iter+1} = w_{1k}^{iter} - \eta \cdot \frac{\partial C}{\partial w_{1k}}$$

Where  $C$  is the cost function,  $y_k$  and  $o_k$  are the actual target value and the predicted value of the neural network. Stochastic gradient decent (SGD) is used to optimize the neural network weights where  $\eta$  represents the learning rate. We note that SGD is one of the popular choices for hardware DNN learning due to its simplicity where the DNN weights are updated at each of the input sample of the DNN.

### Quantization of the DNN:

A quantized neural network (QNN) is a type of neural network where the precision of the weights or activations or both are reduced from floating-point to lower-bit representations, such as 8-bit integers to even binary values. This quantization process significantly reduces the computational and memory requirements, enabling the deployment of neural networks on resource-constrained devices and systems. Despite the reduced precision, QNNs can maintain high accuracy through careful training and optimization techniques. They offer a practical solution for implementing deep learning models in real-time applications, where efficiency and power consumption are critical considerations.

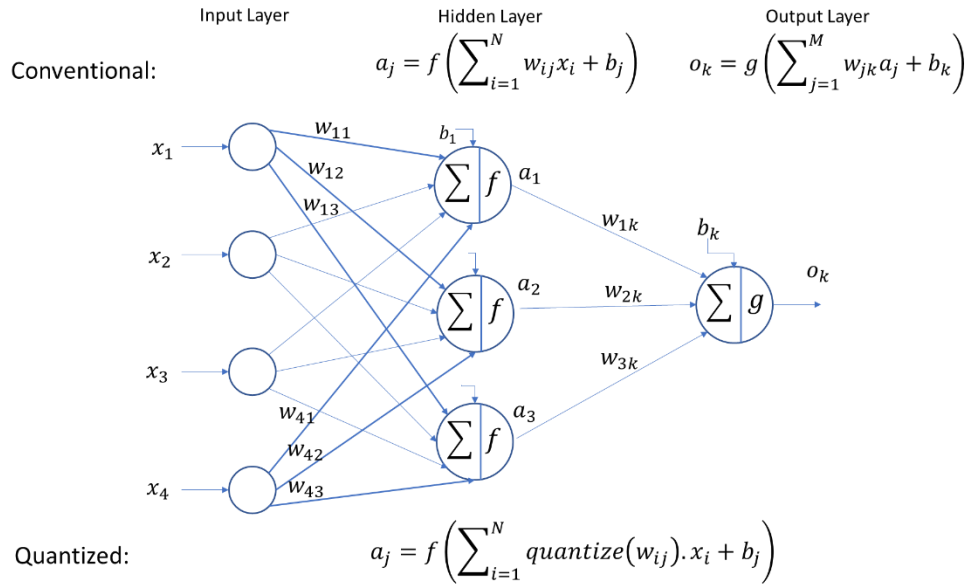


Figure 1-4 Quantization of DNN weights. The corresponding equations shown at top and bottom for hidden layer neurons are for conventional DNN neurons and quantization approach respectively.

Weight Quantization: The number of states of the DNN weights can be selected arbitrarily with the interval,

$\Delta = \frac{w_{ij}^{max} - w_{ij}^{min}}{L}$ . If we assume the quantization function:

$$\text{quantize}(w_{ij}) = w_{ij}^q \tag{3}$$

During the learning of the neural network, the derivative of quantized weight with respect to real weights are required. However, due to the discontinuity in quantization function,  $\frac{\partial w_{ij}^q}{\partial w_{ij}}$  is indeterminate. Thus, backpropagation suffers.

$$\frac{\partial C}{\partial w_{1k}} = \frac{\partial C}{\partial o_k} \cdot \frac{\partial o_k}{\partial \delta} \cdot \frac{\partial \delta}{\partial w_{1k}^q} \cdot \frac{\partial w_{1k}^q}{\partial w_{1k}} \quad (4)$$

We use Straight-through Estimator (STE) approach [3] to address this issue.:

$$\frac{\partial w_{1k}^q}{\partial w_{1k}} = 1 \quad (5)$$

#### 1.1.4. Reservoir computing:

Reservoir computing is a neural network architecture that consists of an input layer, an output layer, and an intermediate reservoir composed of a large number of interconnected neurons with recurrent connections (Fig. 1-5). It leverages the dynamical properties of the reservoir, where the interconnected nodes provide high-dimensional representations of the input data due to their complex and recurrent connections. The connection strengths (or weights) among the constituent nodes (or neurons) of the reservoir are fixed and do not need to be adjusted during learning, which greatly simplifies the training process. Only the weights of the output layer are trained. Due to the reservoir's ability to capture temporal dependencies within the data, it can serve as a powerful alternative to recurrent neural networks (RNNs) or even long short-term memory (LSTM) networks.

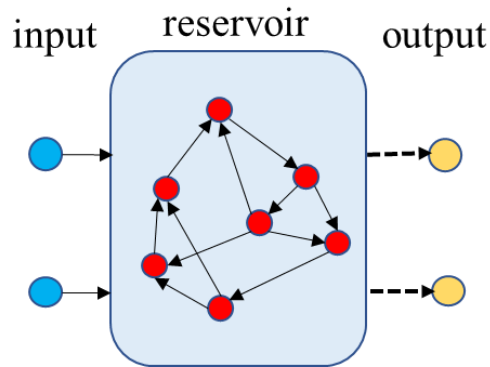


Figure 1-5 Schematic of a reservoir showing the recurrent connections among neurons.

Spintronic devices show rich magnetization dynamics which are highly non-linear. In addition, they possess short term memory properties. These properties make them ideal candidates for reservoir computing. The non-linear dynamic properties of magnetic skyrmion are illustrated in Fig. 1-6.

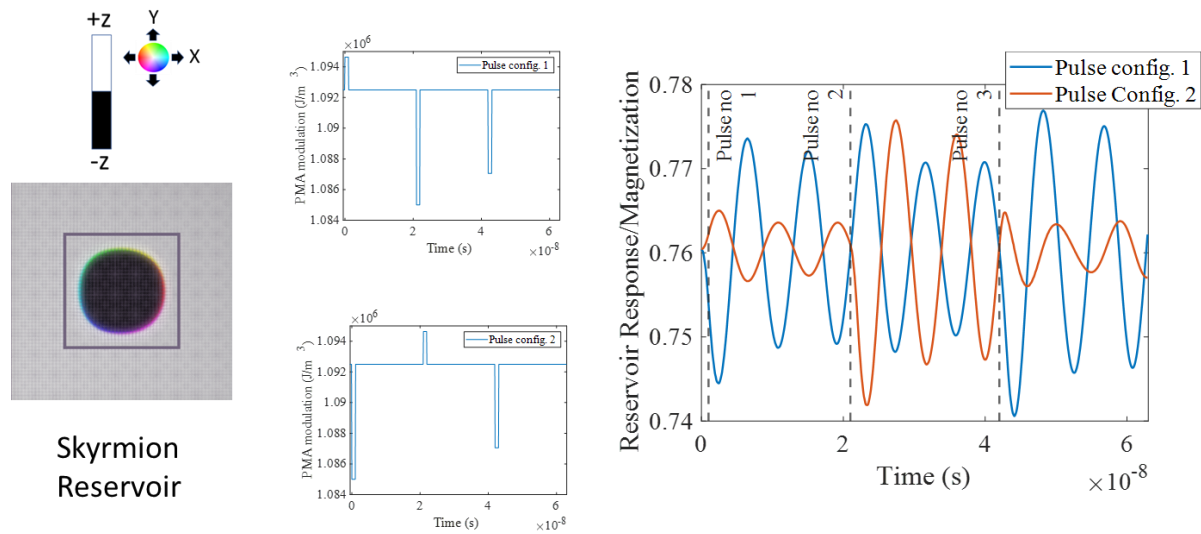


Figure 1-6 Magnetic skyrmion acts as a physical reservoir. The magnetization of the patterned skyrmion oscillates due to modulation of perpendicular magnetic anisotropy (PMA).

### 1.1.5 Imaging techniques for magnetic structure characterization:

To characterize magnetic structures ranging from nanoscale to micrometer scale and magnetic films we use magneto-optical kerr microscopy (MOKE), vibrating sample microscopy (VSM) and magnetic force microscopy (MFM).

#### Magneto optical Kerr Microscopy:

Magneto-optical effects refer to the influence of magnetization direction on the optical constants of a material. Information about magnetic domains can be obtained through the reflected light, due to the Kerr effect, or the transmitted light, through the Faraday effect. Both effects involve a slight rotation of the polarization plane of incident light and can be observed with a polarization microscope. However, since most magnetic materials are not transparent, the Kerr microscope is typically used to observe domain images. The Kerr effect can be applied to any metal or light-absorbing material, while the Faraday effect is used for transparent media.

In the following, we describe the Kerr effect based on the magnetization direction of the sample. Consider a sample with magnetization oriented perpendicular to the film surface, as shown in Fig. 1-7a. When linearly polarized light is incident on the sample, the electrons oscillate parallel to the plane of polarization, corresponding to the plane of the electric field,  $\mathbf{E}$ . Normally reflected lights have the same polarization

plane as the incident light, we denote it as  $\mathbf{E}_{R,N}$ . Lorentz force, acting on the electrons at the same time exert a small motion,  $v_L$  which is proportional to,  $v_L = -\mathbf{m} \times \mathbf{E}$ , where  $\mathbf{m}$  is the magnetization vector. This motion generates the Kerr amplitude,  $\mathbf{E}_{R,K}$  for reflections and the superposition of normal reflection component  $\mathbf{E}_{R,N}$  with  $\mathbf{E}_{R,K}$  creates the magnetization dependent rotation of the polarized light. Polar Kerr effects are shown in Fig. 1-7a where the sample magnetization is perpendicular to the plane. The effect is strongest when the angle,  $\Theta=0^\circ$ . The effect will be similar for incident lights polarized parallel or perpendicular to the incident plane due to the symmetry with respect to,  $\Theta=0^\circ$ .

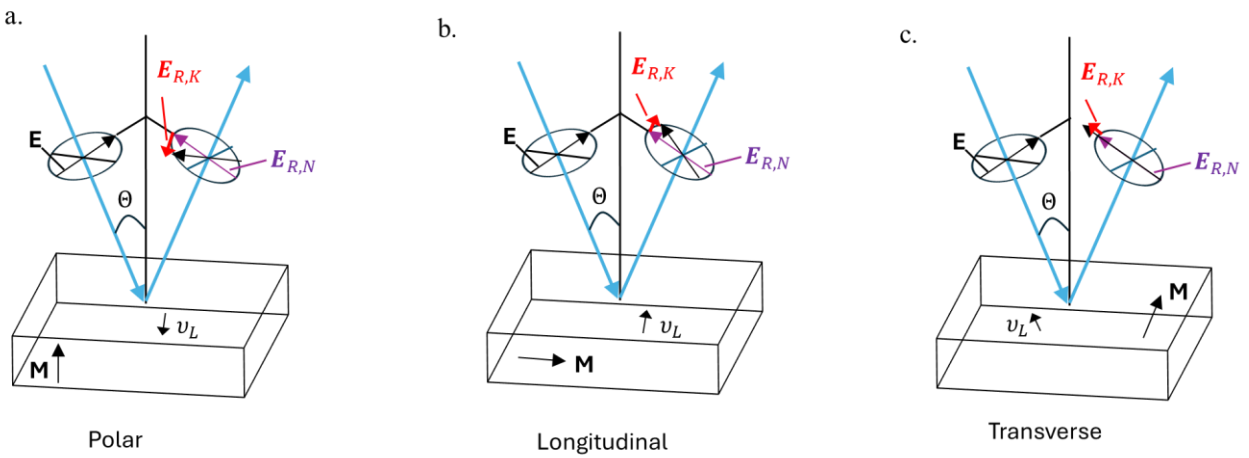


Figure 1-7 Schematics showing the a. polar b. longitudinal and c. transverse geometries of magneto-optical effects.

For longitudinal effect, the plane of incidence is parallel to the direction of the sample magnetization and the magnetization is along the sample surface. Fig. 1-7b shows the longitudinal effects for incident light polarized parallel to the incident plane. The angle of the light beam cannot be zero, as the Lorentz motion will vanish for zero angle incident light. The longitudinal effect is proportional to the angle of incidence following  $\sin\theta$ .

For transverse effect, the plane of incidence is perpendicular to the magnetization direction where the magnetization is parallel to the sample plane, Fig. 1-7c. In reflection, the parallel polarized light (parallel to the plane of incidence) experiences Kerr rotation but along the same direction of the normal reflection. Thus, the transverse effect causes amplitude variation of the light, however, does not produce any appreciable contrast in the magnetic domain images. Similar to the longitudinal, the transverse effect is proportional to  $\sin\theta$ .

The schematic of a high-resolution MOKE microscope is shown in Fig. 1-8. Light passing through a polarizer is reflected from the sample, experiencing normal amplitude reflection coefficients for the incident light. Additionally, Kerr amplitudes are excited depending on the polar, longitudinal, or transverse effect. The reflected light then passes through an analyzer of a specific configuration, which determines the total signal amplitude relative to the incident light. By using an oil immersion lens and shorter wavelength blue light, a minimum resolution of  $0.3 \mu\text{m}$  can be achieved [4].

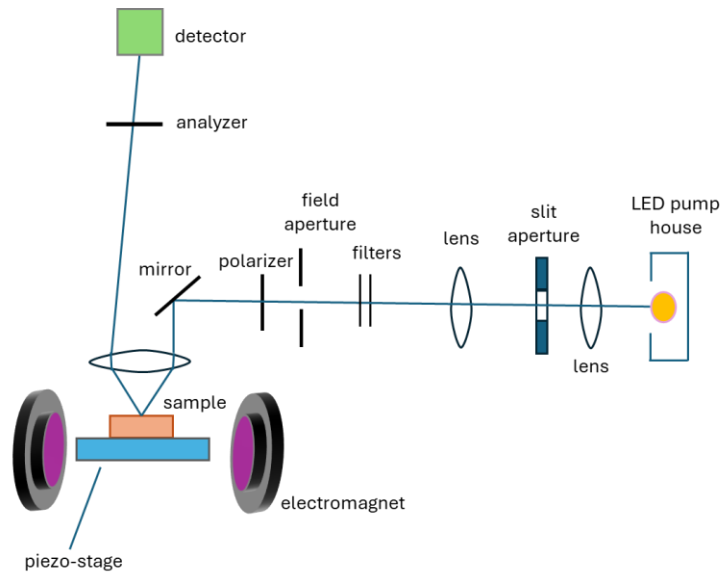


Figure 1-8 Schematic showing the operation of a MOKE microscope.

### **Magnetic Force Microscopy:**

Magnetic Force Microscopy (MFM) is an imaging technique used to measure the magnetic information of the surface at nanoscale. It is a special form of atomic force microscopy (AFM) used to image and measure the magnetic properties of surfaces at the nanoscale. In MFM, a magnetic tip in a cantilever scan over the sample surface, and the sample-tip interactions causes deflections of the cantilever which is detected by a photodetector enabling the visualization of magnetic domains, domain walls, and other magnetic structures.

### **Vibrating sample microscopy:**

Vibrating sample microscopy (VSM) works by vibrating a magnetized sample in a uniform magnetic field, inducing a voltage in pickup coils that is proportional to the sample's magnetic moment, allowing the magnetic properties of the sample to be determined. A simplified sketch of the VSM is shown in Fig. 1-9.



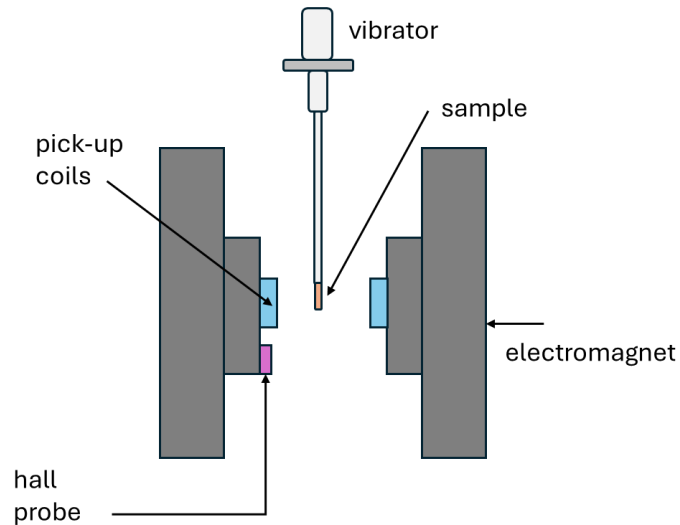


Figure 1-9 Schematic showing the operation of VSM.

### 1.1.6. Microstructure and nanostructure patterning:

Micro and nanoscale structures are patterned using maskless photolithography and electron-beam lithography. Fig. 1-10 shows the steps of patterning the magnetic structures using positive photoresist (or positive electron beam resist).

#### Maskless Photolithography:

Maskless photolithography, also known as direct-write lithography, is used to create patterns on substrates without the need for a physical mask. First, a substrate is coated with a light-sensitive photoresist layer. Then a digital pattern is created and subsequently projected onto the photoresist using a modulated light source such as a laser. This exposure step selectively alters the photoresist according to the digital pattern. After exposure, the substrate is developed in a chemical solution that removes either the exposed or unexposed regions of the photoresist, depending on whether a positive or negative photoresist is used. For positive photoresist, the exposed region will be removed after development. Then cohesive metals (such as Ti or Pt) and magnetic materials are deposited using ion-beam sputtering or electron-beam deposition method to transfer the pattern from the photoresist to the substrate itself. Finally, the remaining photoresist is stripped away, leaving behind the patterned structures.

#### Electron-beam lithography:

The process steps are similar to maskless photolithography except the resist coated substrate is radiated with e-beam in a vacuum. Here, electron-sensitive resist, such as PMMA is used. A digital pattern is

designed and the substrate is then exposed to a focused electron beam in an EBL system, which scans the resist according to the pattern. After development and material deposition, the remaining resist is removed using lift-off, leaving the desired nanoscale features on the substrate.

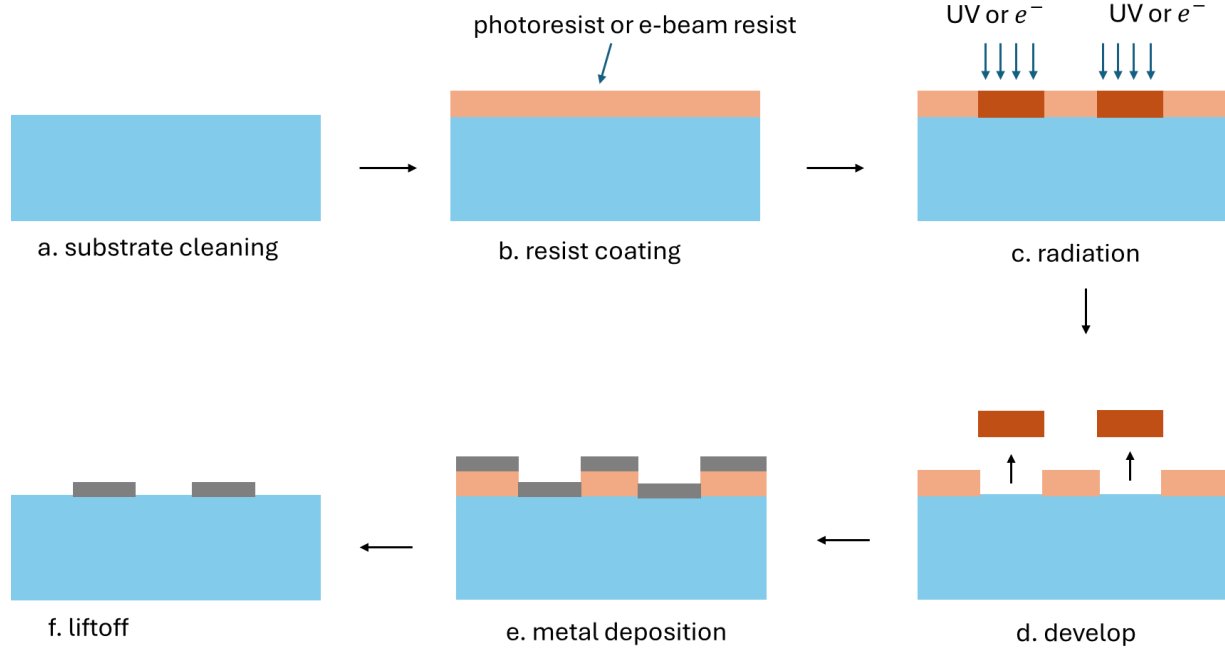


Figure 1-10 Schematics showing operation steps of magnetic microstructures (maskless photolithography) or nanostructures (e-beam lithography) patterning.

### 1.1.7 Micromagnetic modeling of magnetization dynamics:

Magnetization dynamics was simulated by solving the Landu-Lifshitz-Gilbert (LLG) equation in micromagnetic framework [5].

$$(1 + \alpha^2) \frac{d\vec{m}}{dt} = -\gamma \vec{m} \times \vec{H}_{eff} - \alpha \gamma (\vec{m} \times (\vec{m} \times \vec{H}_{eff})) \quad (6)$$

Here  $\gamma$  is the gyromagnetic ratio,  $\vec{H}_{eff}$  is the effective field,  $\alpha$  is the damping,  $\vec{m}$  is the reduced or unit magnetization,  $M_s$  is the saturation magnetization.

$\vec{H}_{eff}$  can be expressed as the contribution from several fields:

$$\vec{H}_{eff} = \vec{H}_{anis} + \vec{H}_{demag} + \vec{H}_{stress} + \vec{H}_{exch} + \vec{H}_{thermal} \quad (7)$$

We assume first order uniaxial anisotropy field,  $\vec{H}_{anis}$  :

$$\vec{H}_{anis} = \frac{2K_u}{\mu_0 M_s} (\vec{v} \cdot \vec{m}) \vec{v} \quad (8)$$

$K_u$  is the first order anisotropy constant and  $\vec{v}$  represents the uniaxial anisotropy direction (i.e. perpendicular to plane).

If a uniaxial stress is applied in plane (perpendicular to the out of plane anisotropy direction), the stress induced field due to the inverse magnetostriction (Villari) effect,  $\vec{H}_{stress}$  can be expressed as:

$$\vec{H}_{stress} = \frac{2K_t}{\mu_0 M_s} (\vec{s} \cdot \vec{m}) \vec{s} \quad (9)$$

The effective stress anisotropy constant can be represented by  $K_t = \frac{3}{2} \lambda_s \sigma$ , where  $\lambda_s$  is the saturation magnetostriction and  $\sigma$  is the stress amplitude and  $\vec{s}$  represents the stress direction. We note that stress induced in the in-plane direction perpendicular to the  $\vec{s}$  direction (which is opposite in sign) can add to the  $\vec{H}_{stress}$ . We do not include this term; thus, our stress estimation is conservative (though the qualitative dynamics remains unchanged as in chapter 2).

Exchange field  $\vec{H}_{exch}$  has two contributions, one from Heisenberg exchange and another from the Dzyaloshinskii–Moriya interaction (DMI). DMI contribution to the exchange field is expressed in micro-magnetic configuration as:

$$\vec{H}_{DMI} = \frac{2D}{\mu_0 M_s} [(\vec{v} \cdot \vec{m}) \hat{z} - \vec{v} m_z] \quad (10)$$

Here,  $m_z$  is the z-component of the unit magnetization vector  $\vec{m}$  and D is the effective DMI constant.

Thermal fluctuation generates  $\vec{H}_{thermal}$  in the following manner:

$$\vec{H}_{thermal} = \vec{\eta} \sqrt{\frac{2\alpha kT}{\mu_0 M_s \gamma \Omega \Delta}} \quad (11)$$

$\vec{\eta}$  is a randomly generated normal Gaussian distributed vector generated at each time step,  $k$  is Boltzmann constant,  $\Omega$  is the finite difference cell volume,  $\Delta$  is the step size.

The effect of the STT torque in micromagnetics is included by means of augmented LLG equations which is called Landu-Lifshitz-Gilbert-Slonczewski equation:

$$(1 + \alpha^2) \frac{d\vec{m}}{dt} = -\gamma\vec{m} \times \vec{H}_{eff} - \alpha\gamma \left( \vec{m} \times (\vec{m} \times \vec{H}_{eff}) \right) - \beta\gamma(\varepsilon - \alpha\varepsilon')(\vec{m} \times (\vec{m}_p \times \vec{m})) \\ + \beta\gamma(\varepsilon' - \alpha\varepsilon)(\vec{m} \times \vec{m}_p) \quad (12)$$

$$\beta = \frac{\hbar J}{\mu_0 e d M_s}, \quad \varepsilon = \frac{P\Lambda^2}{(\Lambda^2 + 1) + (\Lambda^2 - 1)(\vec{m} \cdot \vec{m}_p)} \quad (13)$$

Here  $\hbar$  is the reduced Planck constant,  $J$  is the current density along z direction,  $d$  is the free layer thickness,  $e$  is the electron charge,  $\mu_0$  is the permeability of free space,  $\vec{m}_p$  is the fixed layer unit magnetization. We assume Slonczewski parameter,  $\Lambda = 1$ , secondary spin-torque parameter  $\varepsilon' = 0$  and spin polarization  $P = 0.5669$ .

The effect of the spin orbit torque (SOT) on magnetization dynamics can be addressed by adding two spin orbit torques, damping like torques also called Slonczewski torque and a field like torque expressed as follows:

$$(1 + \alpha^2) \frac{d\vec{m}}{dt} = -\gamma\vec{m} \times \vec{H}_{eff} - \alpha\gamma \left( \vec{m} \times (\vec{m} \times \vec{H}_{eff}) \right) - \tau_D\gamma(\vec{m} \times (\vec{m} \times \vec{\sigma})) \\ + \tau_F\gamma(\vec{m} \times \vec{\sigma}) \quad (14)$$

Where  $\vec{\sigma}$  is the spin polarization direction for the electrons that are accumulated at the interface of the ferromagnet and heavy metal. The direction of  $\vec{\sigma}$  can be found from the direction of the SOT current. If,  $\vec{J}_x$  is the unit vector defining the direction of current flow and  $\vec{z}$  is the direction of inversion asymmetry then  $\vec{\sigma} = \vec{J}_x \times \vec{z}$ .  $\tau_D$  and  $\tau_F$  are the constants associated with the damping like torque (Slonczewski torque) and the field like torque. We neglect the field like torque as it is typically small. The constant  $\tau_D$  can be expressed as:

$$\tau_D = \frac{\hbar J \theta}{2\mu_0 e d M_s} \quad (15)$$

Here,  $J$  is the value of current flowing through the heavy metal layer and  $\theta$  is the spin Hall angle which is 0.1 for Platinum.

As there is no built-in function to address SOT in micromagnetic software MUMAX3 [5] we incorporate SOT by equating the SOT (Eq. 14-15) with the STT as described in Eq. 12-13. For that, we assume spin polarization to be  $P = 1$ , Slonczewski parameter to be  $\Lambda = 1$  and  $\vec{m}_p = -\vec{\sigma} = -\vec{J}_x \times \vec{z}$ . In addition, we consider secondary spin torque parameter to be  $\varepsilon' = \alpha\varepsilon$  and neglect the field like torque.

## **1.2 Organization of this dissertation proposal:**

Chapter 2 discuss the use of acoustically induced ferromagnetic resonance (FMR) for surface acoustic wave (SAW) assisted spin transfer torque (STT) switching of MTJs that can potentially scale to lateral dimensions  $< 50$  nm.

Chapter 3 presents energy efficient multistate non-volatile synapse based on translating domain walls (DWs) in racetracks with a combination of spin orbit torque (SOT) and voltage induced strain. The corresponding stochastic distribution of quantized weights in the presence of thermal noise and edge roughness are explained in detail.

Chapter 4 presents energy efficient learning with low resolution stochastic DW synapse for deep neural networks (DNNs). The algorithm to train the DNNs implemented with such synapses and the corresponding results are explained in detail.

Chapter 5 presents physical reservoir implemented with magnetic skyrmions for autonomous prediction and long-term household Energy load forecasting.

Chapter 6 presents the details of experimentally achieved modulation of magnetic anisotropy in bismuth substituted yttrium iron garnet with voltage-controlled strain. The modifications in domain formation pattern due to the change in voltages observed are explained.

Chapter 7 presents the details of interfacial exchange and magnetostatic coupling observed experimentally in the heterostructures consisting of REIG TmIG having perpendicular magnetic anisotropy of magnetoelastic origin and different thickness of magnetic metal CoFeB.

Chapter 8 presents the details of the experimentally observed dynamic coupling found in the TmIG/CoFeB heterostructures due to exchange of non-equilibrium exchange of spin current when the magnetic layers are driven into resonance.

Chapter 9 presents the conclusion and the future works.

## **References:**

- [1] E. Strubell, A. Ganesh, and A. McCallum, Energy and Policy Considerations for Deep Learning in NLP, arXiv:1906.02243 (2019)
- [2] L. Lanelongue, J. Grealey, M. Inouye, Green Algorithms: Quantifying the Carbon Footprint of Computation, Advanced Science, vol. 8, 12 (2021)

- [3] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations", *The Journal of Machine Learning Research* vol. 18, no. 187, pp. 1-30, Apr. 2017.
- [4] Alex Hubert, and Rudolf Schäfer. *Magnetic domains: the analysis of magnetic microstructures*. Springer Science & Business Media, 2008.
- [5] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. G. -Sanchez, and B. V. Waeyenberge, The design and verification of MuMax3, *AIP Advances* 4, 107133 (2014).

## Chapter 2: Acoustic Wave Induced FMR Assisted Spin-Torque Switching of Perpendicular MTJs with Anisotropy Variation

We have investigated Surface Acoustic Wave (SAW) induced ferromagnetic resonance (FMR) assisted Spin Transfer Torque (STT) switching of perpendicular MTJ (p-MTJ) with inhomogeneities using micromagnetic simulations that include the effect of thermal noise. With suitable frequency excitation, the SAW can induce ferromagnetic resonance in magnetostrictive materials, and the magnetization can precesses in a cone with high deflection from the perpendicular direction. With incorporation of inhomogeneity via lateral anisotropy variation as well as room temperature thermal noise, the magnetization precession in different grains can be significantly incoherent. Interestingly, the precession in different grains is found to be in phase, even though the precession amplitude (angle of deflection from the perpendicular direction) varies across grains of different anisotropy. Nevertheless, the high mean deflection angle due to acoustically induced FMR can complement the STT switching by reducing the STT current significantly; even though the applied stress induced change in anisotropy is much lower than the total anisotropy barrier. This work indicates that SAW induced FMR assisted switching can improve energy efficiency while being scalable to very small dimensions, which is technologically important for STT-RAM and elucidates the physical mechanism for the potential robustness of this paradigm in realistic scenarios with thermal noise and material inhomogeneity.

Magnetic tunnel junctions (MTJ) are finding increasing application as non-volatile nanomagnetic memory devices, which are an alternative to volatile CMOS based memory devices. The most prevalent scheme to accomplish magnetization switching of the free layer of an MTJ (i.e. writing bits) utilizes spin transfer torque (STT) [1, 2]. Although STT memory can be scaled to  $\sim 10$  nm, the energy requirement has not decreased below 100 fJ/bit [3]. Therefore, alternative strategies such as strain mediated [4, 5] and voltage mediated MTJ [6, 7] switching have been investigated. However, scaling strain-based devices is challenging. As the volume ( $V$ ) shrinks, to maintain an energy barrier with sufficient thermal stability ( $E_b = K_u V \sim 1eV$ ), large perpendicular anisotropy ( $K_u$ ) is required. In such a scenario, the static stress ( $\sigma$ ) required to erode the energy barrier, which is determined by  $E_{stress} = 3/2 \lambda_s \sigma V = E_b$ , would also be very large. For example, to erode an energy barrier of  $E_b \sim 1eV$  in a circular nanostructure with lateral dimension of 20 nm and thickness 1 nm (i.e. volume,  $V \sim 314$  nm<sup>3</sup>), and a saturation magnetostriction of  $\lambda_s = 200$  ppm the stress amplitude could be as high as 1.7 GPa. This is possibly an order of magnitude higher than the stress that can be generated dynamically or applied for many cycles in a practical device. While using

material with higher magnetostriction such as Terfenol-D [8] may address this issue partly, there could be other concerns due to bidirectional coupling between magnetization and strain [9], and thus the stress requirement could still be high.

In contrast to static stress, time varying stress can drive the magnetization to acoustically induced ferromagnetic resonance (A-FMR) [10]. This can lead to large amplitude magnetization precession even when the stress induced anisotropy change is substantially smaller than the total energy barrier as the amplitude of this precession grows due to the energy added over many cycles. Physically, time varying strain can be generated by surface acoustic waves (SAW) via interdigital SAW electrode deposited and patterned over a piezoelectric substrate. Previously, SAW driven FMR on Ni film [11] and magnetization switching for magnetostrictive Co nanomagnets has been experimentally reported [12]. Precessional magnetization switching [13] with SAW on (Ga, Mn) (As, P) film and field free switching with SAW for (Ga, As) P [14] has also been experimentally reported at low temperature. Laser pump induced SAW and their magnetization dynamics has been studied for single nanomagnet [15] and on patterned periodic nanodots [16] and their magnetization reversal has been numerically investigated in nanomagnets [17].

SAW assisted STT induced magnetization reversal for in-plane and perpendicular MTJ based on simulations using macro-spin assumption has recently been reported [18]. The main purpose of the scheme is to induce magnetization rotation with SAW so that when the STT is applied, the magnetization experiences higher torque. However, macrospin assumption precludes modeling of incoherence in the magnetization dynamics, which could arise due to inhomogeneity in material properties. This inhomogeneity can be an intrinsic material property [19], or due to edge modifications [20], roughness variation [21], thickness variation over an extended area [22], etc. While the incoherence in magnetization mainly stems from the competition between the long ranged weak magnetostatic energy and short ranged strong exchange energy and the balance is ultimately decided by the size and shape of the nanomagnetic structure; the above-mentioned inhomogeneities can increase the incoherency even in smaller size nanomagnet by varying the local anisotropy field.

In this study, we perform micromagnetic simulations that incorporate the incoherent magnetization dynamics in the presence of room temperature thermal noise as well as lateral variation in the uniaxial (perpendicular) anisotropy. Furthermore, the cell size of lateral dimensions  $1.56 \text{ nm} \times 1.56 \text{ nm}$  naturally includes an edge roughness  $\sim 1 \text{ nm}$  that is consistent with lithography and fabrication limitations. It is expected that all these realistic variations could lead to incoherent magnetization precession. For example, the magnetization in different grains (Voronoi tessellation was used to create regions with an average lateral dimension  $\sim 10 \text{ nm}$ ) can precess at different cones when driven by a SAW as shown in Fig. 2-1c. This incoherent (non-uniform) magnetization precession reduces the net magnetization of the nanomagnet (as in



Fig. 2-1c although some regions precess in high cone, the net magnetization remains very low) and in extreme cases reduces it to zero. Therefore, we study this incoherent precession, particularly in the SAW induced FMR regime, to understand the underlying magnetization dynamics as well as its ramifications on the ability to significantly reduce the STT switching current. This has important implications towards implementation of energy efficient STTRAM that are scalable to small lateral dimensions.

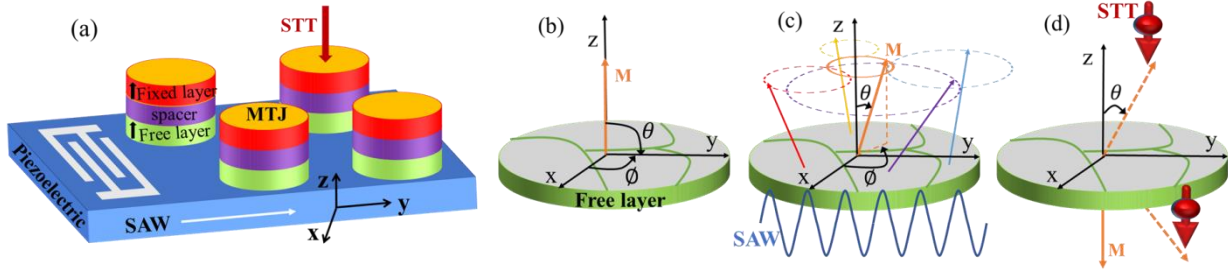


Figure 2-1 a. MTJ arrays and SAW electrode over piezoelectric substrate b. initial magnetization state of the inhomogeneous (i.e. granular) free layer c. application of SAW induces different angle precession and the resulting incoherency reduces the net magnetization,  $M$  d. final magnetization state after application of STT current.

## 2.1 Model:

We assumed the piezoelectric substrate to be Lithium Niobate and the SAW wave launched by the Interdigital transducer (IDT) patterned on top the piezoelectric is Rayleigh wave which is propagating along y-axis (Fig. 2-1a). Such a Rayleigh wave has three dominant strain components, normal strain along propagation direction y, normal strain along direction z and a shear strain in x-z plane. We only consider the strain component along the SAW propagation direction and neglect the normal strain along z-direction as the nanomagnet is not clamped on top and the shear strain is weak near the surface [23]. For simplicity, we do not consider Einstein-de Hass effect in the manner of Ref. [24].

We performed micromagnetic simulation using mumax3 [25] where we divided our circular MTJ free layer with 50 nm diameter and 1.5 nm thickness into  $32 \times 32 \times 1$  cells. The cell size is well within the ferromagnetic exchange length  $\sqrt{2A_{ex}/\mu_0 M_s^2} \sim 6$  nm. We simulate an inhomogeneous nanomagnet shown in Fig. 2-2 with 10 nm average lateral dimension grains (regions with different anisotropies created by the Voronoi tessellation). The anisotropy direction (easy axis) for all the grains is assumed to be the same and perpendicular to the plane (z-axis). The anisotropy constant,  $K_1$  (for details see eq. 4) is varied within the grains and the values of  $K_1$  have Gaussian distribution with mean value of  $4.5837 \times 10^5$  J/m<sup>3</sup> and standard deviation of 5%. The exchange stiffness constant,  $A_{ex}$  which defines the strength of the interaction between

neighboring spins could be reduced at the grain boundaries but assumed to be constant for the sake of simplicity. For homogeneous nanomagnet the mean value of  $K_1$  is considered to be the anisotropy constant.

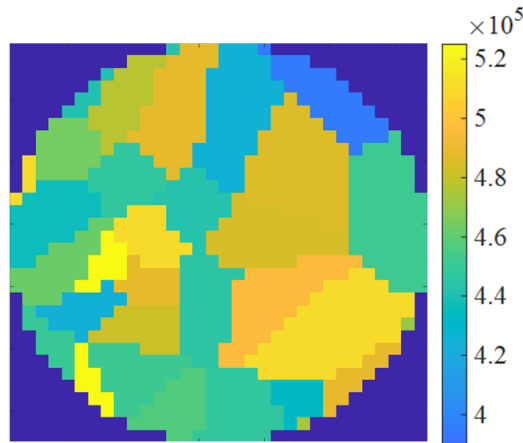


Figure 2-2 Grain distribution of the inhomogeneous nanomagnet using Voronoi tessellation. The colormap is showing the values of the first order anisotropy constant,  $K_1$  (see eq. 4) for different grains. The values of  $K_1$  is chosen using gaussian distribution of mean  $4.5837 \times 10^5 J/m^3$  and standard deviation of 5%.

Magnetization dynamics in the presence of SAW induced stress and STT current was simulated by solving the Landu-Lifshitz-Gilbert-Slonczewski equation in MuMax3 [25] as explained in section 1.2.7. The material parameters for simulation are shown in Table 2-1.

Table 2-1: FeGa material properties [26, 27]

Parameters	$\text{Fe}_{81}\text{Ga}_{19}$
Saturation magnetostriction ( $\lambda_s$ )	350 ppm
Gilbert damping ( $\alpha$ )	0.015
Saturation magnetization ( $M_s$ )	$0.8 \times 10^6$ A/m
Gyromagnetic ratio ( $\gamma$ )	$2.21 \times 10^5$ m/A.s
Exchange stiffness ( $A_{ex}$ )	18 pJ/m

## 2.2 Results and discussion:

We first simulate a perpendicular homogeneous nanomagnet with an energy barrier  $\sim 70$  kT with uniform initial magnetization tilted ( $\sim 2^\circ$ ) from the perpendicular z-axis. This assumption is conservative as thermal noise was found to tilt the mean equilibrium magnetization by  $\sim 20^\circ$ . To investigate the magnetization precession behavior with time varying strain, we excite the nanomagnet with SAW of different frequencies at  $T = 0$  K. In response, the magnetization starts to precess in a cone around the perpendicular axis as seen from Fig. 2-3a. The precession slowly settles to a mean deflection of  $\sim 35^\circ$  from the perpendicular direction

when 100 MPa SAW excitation is applied for a sufficiently long time. However, we note that even a static 200 MPa stress cannot induce any reasonable equilibrium deflection ( $< 1^\circ$ ), proving that the large deflection from the perpendicular direction is specifically a resonance effect.

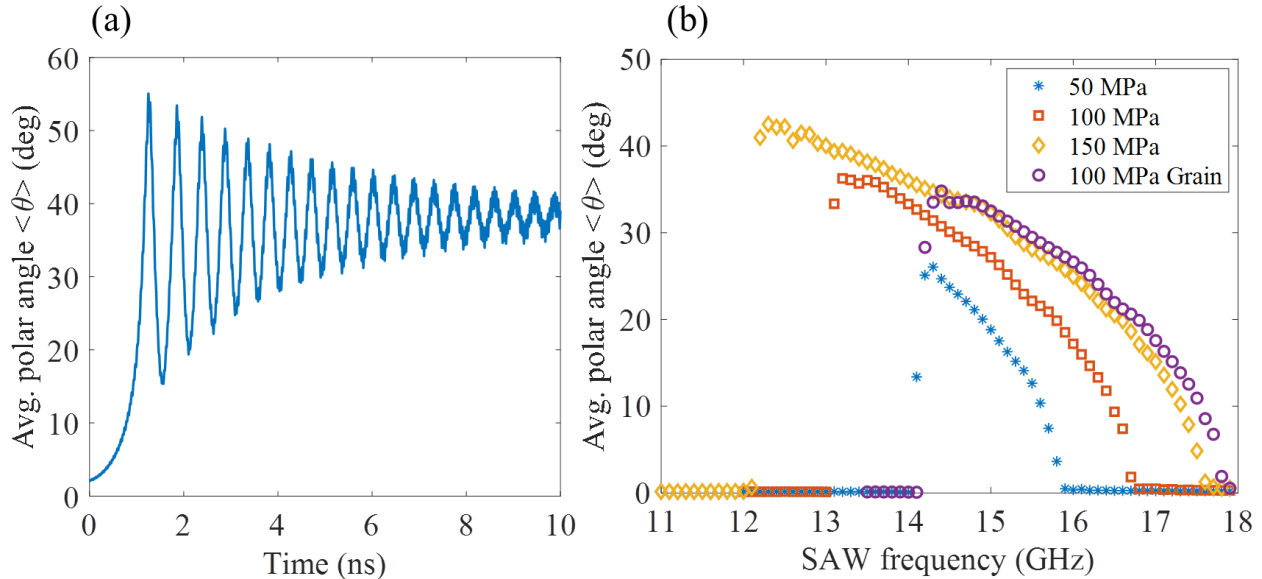


Figure 2-3 a. Evolution of magnetization deflection ( $\theta$ ) from the perpendicular  $z$  direction for 100 MPa 13.2 GHz SAW b. Average polar angle deflection ( $\theta$ ) for different excitation frequency for varying stress amplitude. Resonance point (highest deflection) shifts towards lower frequency with increasing stress.

The average precession cone deflection angle (polar angle,  $\theta$  see Fig. 2-1) is plotted for different excitation frequency in Fig. 2-3b for varying stress amplitude. As we can see from Fig. 2-3b, the highest deflection angle (resonant point) shifts towards the low frequency with increasing stress amplitude. As we increase the stress amplitude, the magnetization deflects more from the perpendicular direction and the effective field in the perpendicular  $z$ -direction decreases. This low effective field in the  $z$ -direction causes slower magnetization precession at resonance.

Introducing inhomogeneity in nanomagnets (i.e. anisotropy variation between grain) could modify the overall (mean effective) anisotropic field strength along the  $z$ -direction and consequently alter the resonance frequency compared to the homogeneous nanomagnets. As different grains have different resonance frequencies, one may expect that the deflection angle vs. frequency is less steep (resonance is not sharp). However, it is likely that the strong exchange interaction forces the individual regions' magnetization to precess nearly in phase. Therefore, the deflection angle vs. frequency (and resonance characteristics) was found to be similar to the homogeneous structure with a mere frequency shift because of mean anisotropy change. Similarly, in the presence of room temperature thermal noise (at  $T=300$  K) the equilibrium magnetization fluctuates randomly producing a higher mean deflection angle. In such a situation, the

effective field in the z-direction is lower compared to the T=0 K case. Therefore, the resonant point was found to shift towards low SAW excitation frequency (not shown in Fig. 2-3b).

The micromagnetic configurations of SAW induced magnetization dynamics are presented in Fig. 2-4. For the no grain case at T=0 K (Fig. 2-4a), the spins rotate coherently and eventually settle to an equilibrium cone. Similar behavior is observed for the case with inhomogeneous grains except the spin dynamics is incoherent as spins in different regions precess with different polar angle ( $\theta$ ) with respect to the z-axis (Fig. 2-4b, 0.99 ns and 1.61 ns). However, at room temperature both homogeneous (Fig. 2-4c) and inhomogeneous case (Fig. 2-4d) become incoherent (details in Fig 2-5). Notably, for the incoherent cases, the average magnetization deflection can still be high. We next investigate the incoherency and how it affects the magnetization dynamics in SAW driven FMR. Local variation in anisotropy due to material inhomogeneity and thermal perturbation introduces significant incoherency in the nanomagnet as evident from different deflection angles of magnetization precession for different regions with SAW excitation. Notwithstanding the incoherency due to the different deflection angles (polar angle,  $\theta$ ), when driven by resonant SAW that produces a sufficiently large  $\theta$ , magnetization in the different regions precess almost in phase (in-plane azimuth angle,  $\phi$ ). This phase matching of the precessions can produce high net deflection from the perpendicular axis.

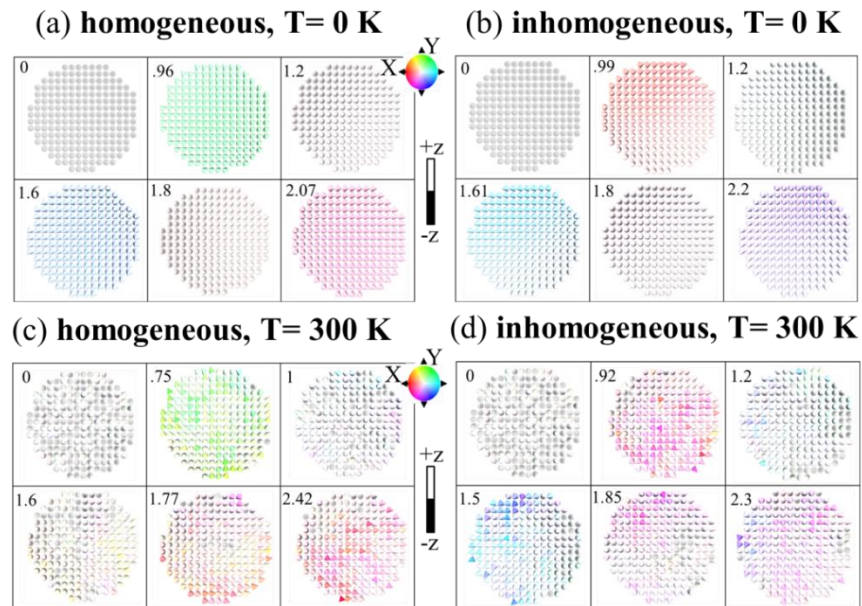


FIG. 2-4. Evolution of magnetization states with time (shown at the left corner in nanosecond) under the excitation of SAW in a. homogeneous and b. inhomogeneous nanomagnets at T=0 K and c. homogeneous and d. inhomogeneous nanomagnets at T=300 K.

Fig. 2-5a and 2-5b plots the spin configuration for inhomogeneous nanomagnet at T=300K, which show the evolution of polar angle ( $\theta$ ) and the in-plane phase angle/azimuth angle ( $\phi$ ) for individual spins respectively. Mean polar angle deflection ( $\langle\theta\rangle$ ) of the magnetization at every time reference is also presented. From Fig 2-5a and 2-5b it is evident that spins in different regions are incoherent as they have non-uniform polar and phase angle. However, from Fig 2-5b it can be seen that spins are repeatedly precessing almost in phase ( $\phi$ ) while producing high mean polar angle ( $\langle\theta\rangle$ ) (snapshot at 0.61 ns, 1.35 ns, 2.12 ns in Fig. 2-5b) but out of phase ( $\phi$ ) for low mean polar angle (snapshot at 0.25 ns, 0.94 ns, 1.71 ns in Fig. 2-5b). This suggests that high polar deflection between the regions at resonance makes our SAW assisted scheme robust to inhomogeneities.

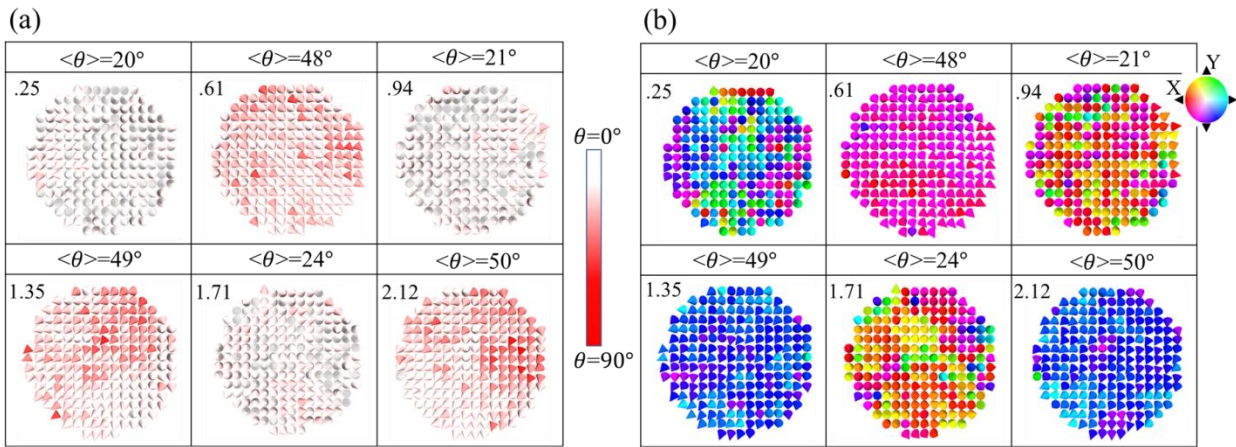


FIG. 2-5. Spin configuration of inhomogeneous nanomagnet at T=300 K at different snapshots in time (in nanoseconds) in left corner and average magnetization polar angle ( $\langle\theta\rangle$ ) a. polar angle ( $\theta$ ) in different regions implies incoherency for both high and low average magnetization polar angle ( $\langle\theta\rangle$ ) b. in-plane azimuth angle ( $\phi$ ) for individual spins shows that the spins are almost in phase while precessing in high polar angle ( $\langle\theta\rangle$ ) but out of phase while precess in low polar angle.

This large mean deflection in the presence of thermal noise and inhomogeneity improves the efficacy of SAW assisted STT devices. The STT effective field can be expressed as  $\vec{H}_{STT} = \beta \epsilon (\vec{m}_p \times \vec{m})$ , which shows that the field magnitude is a function of  $\sin\theta$ , where  $\theta$  is the angle between fixed layer and free layer magnetization. Thus, the SAW induced high magnetization precession of the free layer can assist in building large STT torque compared to the no SAW case. We investigate the performance of SAW assisted STT switching scheme in the presence of room temperature thermal noise. Here, we assume the SAW simultaneously excites arrays of several MTJs and the STT current writes bits (Fig. 2-1a) and thus we do not consider a precise synchronization between the SAW and STT application. This is simulated as follows: we excite the nanomagnets with SAW from t=0 to t=5 ns while STT current is applied for 1ns from t=3ns to t=4ns. Therefore, after the withdrawal of STT current, the SAW is still applied for 1ns (t=4 ns to t=5 ns) as the MTJs can be exposed to SAW even after the STT current pulse is withdrawn. We analyze the final

magnetization states of the MTJs after 2 ns of SAW withdrawal ( $t = 7$  ns). We assume the threshold for switching to be  $\sim 130^\circ$ , which is conservative. Several switching trajectories for both the case with no grain (no inhomogeneity) and grain (inhomogeneity) are presented in Fig. 2-6, where the granular nanostructure is sketched in the inset of Fig. 2-6b. From Fig. 2-6a and 2-6b, we can see that the precession cone (polar angle,  $\theta$ ) oscillates between highest and lowest peaks at  $\sim 1$  GHz, which motivates the STT application window of 1 ns, so it coincides with at least one precession peak.

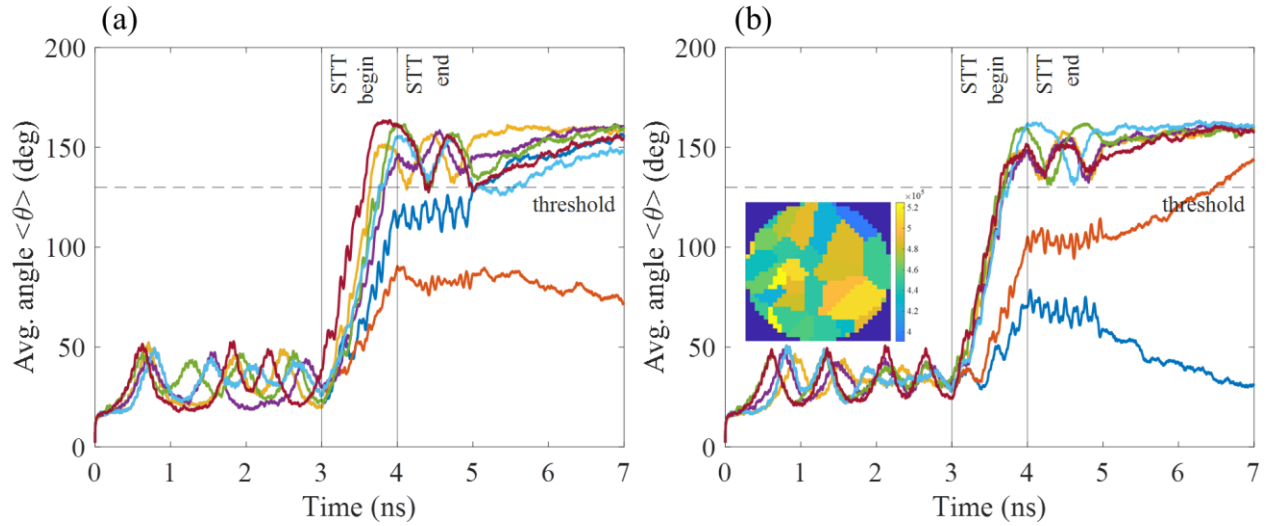


FIG. 2-6. Switching trajectories of SAW assisted STT a. without grains b. with grains. The inset shows the grain configuration of the nanomagnet.

Finally, to see the effectiveness of our SAW assisted STT switching scheme we simulate 100 switching trajectories for different stress amplitudes and varying STT current for both homogeneous and inhomogeneous nanomagnets (Fig. 2-7). For the no SAW case, STT current is applied for 1 ns and final magnetization state was evaluated 2 ns after the STT is withdrawn with timing details and reasons for their choice described earlier. From the switching probability curve, it is seen that SAW assisted STT scheme requires less STT current than the no SAW case. While the no SAW case requires a current density of  $2.0 \times 10^{11} \text{ A/m}^2$  for switching with 100% probability, in the presence of SAW the STT current can be reduced to at least  $1.5 \times 10^{11} \text{ A/m}^2$ . The energy dissipation due to SAW excitation is very low (see next section 2.3). Moreover, energy pumped from one SAW source is amortized over several MTJs. Therefore, energy dissipation is dominated by STT current as discussed in the next section where we show that the SAW energy at 50 MPa stress amplitude is less than 0.3% of the STT energy. Thus, our scheme provides  $\sim 1.8$  times improvement in energy efficiency by reducing the STT current with 50 MPa stress amplitude and this could be higher if we further optimize the design. If we can increase the stress to 100 MPa then at least two-times energy reduction is possible (section 2.3). The interesting result to be noted here is that

regardless of homogeneous or inhomogeneous nanomagnet studied the switching current is decreased by approximately the same amount for the same SAW amplitude. Therefore, the incoherent magnetization precession in granular nanomagnets does not degrade the performance of the SAW assisted STT switching. Moreover, for stress induced change in anisotropy is 4 times less than the total anisotropy barrier for 50 MPa SAW that demonstrates that resonant SAW (acoustic FMR) allows extreme scalability not possible with static stress.

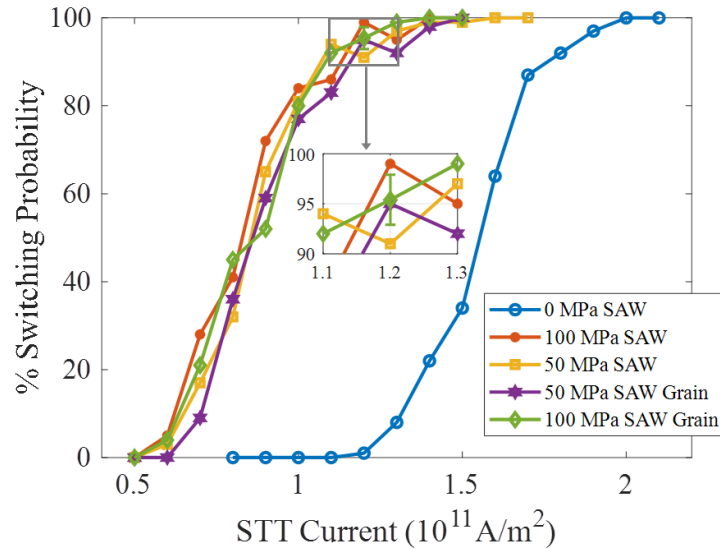


FIG. 2-7. Switching probability vs. STT current at different resonant SAW amplitudes. Error bar shown in the inset corresponds to one of the data points (100 MPa SAW grain case at  $1.2 \times 10^{11} \text{ A/m}^2$  current) for 1000 simulations.

### 2.3 Energy dissipation:

Our proposed device structure is shown in Fig. 2-8. The free layer (i.e. magnetostrictive layer) of the magnetic tunnel junction (MTJ) array is in contact with the piezoelectric substrate. A Rayleigh wave (SAW) is launched by applying a voltage across the interdigital transducer (IDT) electrodes delineated on top of the piezoelectric. As seen from Fig. 2-8, the Rayleigh wave propagates along in-plane y-direction. This propagating wave can generate periodic strain (tensile strain in one-half cycle and compressive strain in the other half) in the piezoelectric substrate, which is then mechanically transferred to the free layer. For materials with positive magnetostriction, such as FeGa, tensile (compressive) strain creates easy (hard) magnetization axes along the direction of the strain (in plane y-direction). This strain (stress) induced anisotropy competes with perpendicular (out of plane z-direction) magnetic anisotropy and consequently the magnetization deflects towards the in-plane y-direction. Resonant SAW can produce higher magnetization deflection than static strain which is then leveraged by a STT current to accomplish the magnetization reversal in our SAW assisted STT switching scheme.

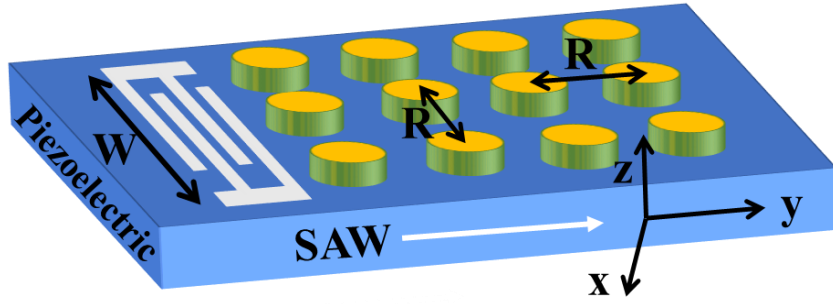


Figure 2-8 Nanomagnet array and SAW electrode (IDT) patterned on top of the piezoelectric substrate. The nanomagnet center to center distance ( $R$ ) and IDT beamwidth ( $W$ ) are shown.

The energy required to strain the magnetostrictive free layer of the nanomagnet is related to the frequency and beamwidth of the Rayleigh wave and the acoustic parameters of the piezoelectric media such as piezoelectric coefficient  $d_{33}$  and Rayleigh wave propagation speed. We assume Lithium Niobate as our piezoelectric substrate. Piezoelectric coefficient  $d_{33}$ , defined by the ratio of the induced strain to the applied electric field where both strain and electric field are in the same direction, is reported to be  $\sim 34.45$  pm/V for commercially available Lithium Niobate wafer [28]. However, some studies report much higher  $d_{33}$  values  $\sim 400$  pC/N (pm/V) [29, 30] for alkaline based Niobates. The SAW (Rayleigh wave) propagation speed in Lithium Niobate is reported to be 3488 - 4750 m/s [31].

To determine the total energy dissipation in nanomagnet for our resonant SAW assisted STT switching, we assume that all the strain generated in the piezoelectric is transferred to the nanomagnet on top of it. Note that the depth of penetration of the SAW in the Lithium Niobate is about one wavelength, which is a few 100 nm, whereas any adhesion layer and MTJ layers are each a few nm, enabling close to 100% strain transfer. If the required stress for successful switching of the nanomagnet is  $\sigma$ , then the strain that needs to be produced in the piezoelectric,

$$\varepsilon = \sigma/Y \quad (11)$$

Where  $Y \sim 75$  GPa is the Young's modulus of FeGa [32]. For  $\sigma = 50$  MPa stress, the strain needs to be generated is,  $\varepsilon \sim 667 \times 10^{-6}$ .

The surface potential,  $V$ , needed to produce the required strain in the piezoelectric can be found from the  $d_{33}$  coefficient and SAW wavelength. For the inhomogeneous nanomagnet we found the resonant frequency to be  $\sim 13$  GHz at room temperature. For SAW propagation speed of 3488 m/s, the wavelength is found to be,  $\lambda \sim 268$  nm.



Once the wavelength and the piezoelectric coefficients are known, we can find the strain from the particle displacement of the Rayleigh wave at the piezoelectric surface. We have only considered the strain due to particle displacement in the propagation direction,  $y$  and neglect the  $z$ -direction strain as the nanomagnet is not clamped from the top. In addition, the shear strain in the  $x$ - $z$  plane is weak near the surface [23]. At a particular snapshot, the displacement along the acoustic wave propagation direction (i.e.  $y$ -direction) can be expressed as,

$$u_y = d_{33}V \sin \frac{2\pi y}{\lambda} \quad (12)$$

As the normal strain is expressed by the slope of the displacement,  $\varepsilon_{yy} = \frac{\partial u_y}{\partial y}$ , the maximum strain can be found around  $y=0$ . If the nanomagnet diameter is,  $D=50$  nm then the displacement for  $y = \pm 25$  nm can be determined from the displacement Eq. 12 (see Fig. 2-9).

The total displacement between the two extremes of the nanomagnet can be determined from the following,

$$u_y|_{y=+25nm} - u_y|_{y=-25nm} = 2d_{33}V \sin \frac{50\pi}{\lambda} \quad (13)$$

Where  $\lambda$  is expressed in nm. Therefore, the maximum strain generated in nanomagnet is,

$$\varepsilon = \frac{u_y|_{y=+25nm} - u_y|_{y=-25nm}}{D} = \frac{2d_{33}V \sin \frac{50\pi}{\lambda}}{D} \quad (14)$$

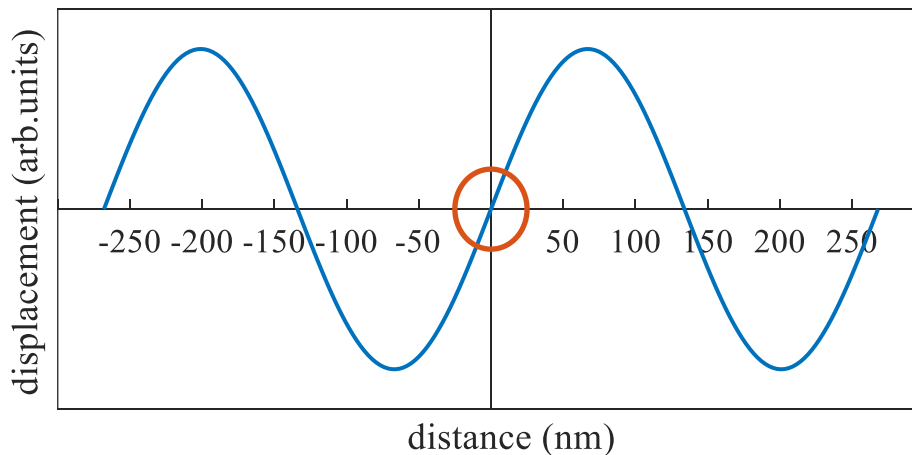


Figure 2-9 Displacement  $u_y$  versus distance  $y$  curve (in blue) in the delay line of the piezoelectric. Maximum strain is generated when the nanomagnet center is at any zero-crossing.

Combining the strain Eqs. 11 and 14 the required surface potential can be determined,

$$V = \frac{\sigma D}{2 Y d_{33} \sin \frac{50\pi}{\lambda}} \quad (15)$$

For  $\sigma=50$  MPa and  $d_{33}= 34.45$  pm/V, the maximum surface potential required is,  $V= 0.875$  Volts.

The power launched by the SAW IDT electrode can be determined from the IDT beamwidth  $W$  (Fig. 2-8), the surface potential and the admittance of the piezoelectric medium. The ac power dissipation can be calculated from the following equation [33],

$$\frac{P}{W} = \frac{1}{2} |V|^2 \frac{y_a}{\lambda} = 299.96 \text{ w/m} \quad (16)$$

Where,  $y_a = 0.21 \times 10^{-3}$  (S) is the admittance of Lithium Niobate [33].

Attenuation of the acoustic wave inside the piezoelectric medium is closely related to the frequency and should be considered for the energy calculation purpose. Acoustic attenuation in Lithium Niobate can be determined from the following equation [34],

$$\alpha = 0.88 f^{1.9} + 0.19 f \quad (17)$$

Where  $\alpha$  is expressed in  $dB/\mu s$  and  $f$  is in GHz. For  $f=13$  GHz attenuation is found to be  $\alpha =117.54$   $dB/\mu s$ . If 10% attenuation of the input power is allowed then the distance travelled by the SAW with speed  $v_r=3488$  m/s before 10% attenuation is,

$$L = \frac{-10 \log_{10}(0.9 P/p)}{\alpha} v_r = 13.58 \mu m \quad (18)$$

If the IDT beamwidth is taken to be  $W= 600$  nm ( $\sim 4$  times of the finger pitch which is  $\lambda/2 \sim 134$  nm) and nanomagnet center to center distance in the array is taken to be,  $R=75$  nm (Fig. 2-8), the number of nanomagnets that can be accommodated within the area of  $W \times L$  is  $N \sim 1448$ .

The energy dissipated per nanomagnet to produce a stress of 50 MPa can be calculated from the following,

$$E_{SAW} = \frac{P}{W} W t_{SAW} \frac{1}{N} = 1.51 \times 10^5 \text{ kT} \quad (19)$$

Where SAW application interval is  $t_{SAW} =5$  ns. The STT current density required for 50 MPa SAW assisted case is,  $J_{STT-SAW} = 1.5 \times 10^{11}$   $A/m^2$ . If the MTJ resistance is taken to be,  $R_{MTJ} = 3.5$   $k\Omega$  [3], then the energy dissipation for the STT current is,

$$E_{STT-SAW} = (J_{STT-SAW} A)^2 R_{MTJ} t_{STT} = 7.39 \times 10^7 \text{ kT} \quad (20)$$

Where  $A$  is the nanomagnet area,  $A=\pi r^2$  where  $r$  is the radius of circular nanomagnet and STT application period is,  $t_{STT} = 1\text{ns}$ .

The energy dissipated due to the material damping can be found from the following,

$$E_{damp} = \int_0^{t_{Total}} \frac{\alpha \gamma}{(1 + \alpha^2) M_s \Psi} |\tau_{eff}(t)|^2 dt \quad (21)$$

Where  $\gamma$  is the gyromagnetic ratio (rad/Ts),  $\Psi$  is the nanomagnet volume and  $\tau_{eff}$  is the effective torque due to shape anisotropy, stress anisotropy, spin transfer torque and  $t_{Total} = 7 \text{ ns}$  is the total time period. The damping energy dissipation is found to be  $\sim 115 \text{ kT}$  which is negligible. Finally, the total energy dissipated in a single nanomagnet for 50 MPa SAW assisted STT switching (dissipation for magnetic damping is neglected) is found to be

$$E_{SAW} + E_{STT-SAW} = 7.4 \times 10^7 \text{ kT} \quad (22)$$

It is evident that energy dissipation is dominated by STT current dissipation and hence the SAW energy dissipation can be neglected.

Without the SAW assistance the required STT current density is,  $J_{STT} = 2.0 \times 10^{11} \text{ A/m}^2$ . In that case, the total energy dissipation in a single nanomagnet is found to be,

$$E_{STT} = (J_{STT} A)^2 R_{MTJ} t_{STT} = 1.31 \times 10^8 \text{ kT} \quad (23)$$

The energy dissipation in STT-only case is  $\sim 1.8$  times higher than the 50 MPa SAW assisted STT case.

For higher stress amplitude such as 100 MPa stress the energy dissipation for SAW excitation is found to be,  $E_{SAW} = 6.05 \times 10^5 \text{ kT}$  which is 4 times higher than the 50 MPa stress case. The STT current density required for 100 MPa SAW assisted case is,  $J_{STT-SAW} = 1.4 \times 10^{11} \text{ A/m}^2$  which translates to an energy dissipation of  $E_{STT-SAW} = 6.43 \times 10^7 \text{ kT}$ . In addition, the damping dissipation is found to be  $\sim 150 \text{ kT}$ . Therefore, the total energy dissipation is,  $E_{SAW} + E_{STT-SAW} = 6.49 \times 10^7 \text{ kT}$  neglecting the damping dissipation,  $\sim 2$  times less than the STT only case.

## 2.4 Conclusions:

In summary, we have studied magnetization precession dynamics with time varying stress generated by SAW for nanomagnets with inhomogeneity due to lateral variations in anisotropy as well as thermal noise. When SAW induces FMR in such inhomogeneous nanomagnets, different regions' spins precess nearly in

phase at high net magnetization deflection, consequently lowering STT current required to switch the magnetization. While out of phase precession occurs at low deflection from perpendicular anisotropy axis, the precession in different regions synchronize (are in phase) for high amplitude magnetization deflection from anisotropy axis. Thus, the efficacy of the SAW that produces large resonant deflections does not degrade due to such incoherency in the presence of lateral anisotropy variations and thermal noise.

Technologically, such SAW induced FMR assisted STT memory devices have potential to scale below ~20 nm lateral dimensions even though the concomitantly high anisotropy energy density needed at such small volumes cannot be overcome by static stress. In addition, using well optimized SAW excitation frequency that can maximize the torque on the magnetization when STT is applied and choosing magnetostrictive materials with extremely low damping material one can potentially achieve over an order of magnitude energy reduction in write energy while being able to scale aggressively to very low lateral dimensions. Finally, such SAW induced FMR could have application in low power electronics beyond memory devices [35].

Apart from inhomogeneous nanomagnets there could be additional challenges in the SAW mediated STT memory devices when a large arrays of nanomagnets are being excited with the SAW electrodes. Magnetostatic interactions between the nanomagnets can alter the resonance frequency. We observed reasonable magnetization precession for a wide range of SAW frequencies, but the range became narrower for small amplitude stress. The maximum stress amplitudes that could be transferred to magnetostrictive magnets depends on piezoelectric coefficients and magnetostriction and for most material systems those values are small. Furthermore, there could be additional unwanted transverse modes generated perpendicular to the directions of main mode of SAW induced vibration. This issue can be addressed by changing the shape and geometry of the SAW electrodes using the apodization method [36,37].

## References:

- [1] J. C. Slonczewski, Current-driven excitation of magnetic multilayers, *J. Magn. Magn. Mater.* 159, L1 (1996).
- [2] H. Kubota, A. Fukushima, K. Yakushiji, T. Nagahama, S. Yuasa, K. Ando, H. Maehara, Y. Nagamine, K. Tsunekawa, D. D. Djayaprawira, N. Watanabe, and Y. Suzuki, Quantitative measurement of voltage dependence of spin-transfer torque in MgO-based magnetic tunnel junctions, *Nat. Phys.* 4, 37 (2008).
- [3] J. J. Nowak, R. P. Robertazzi, J. Z. Sun, G. Hu, J.-H. Park, J. Lee, A. J. Annunziata, G. P. Lauer, R. Kothandaraman, E. J. O'Sullivan, P. L. Trouilloud, Y. Kim, and D. C. Worledge, Dependence of

- voltage and size on write error rates in spin-transfer torque magnetic random-access memory, *IEEE Magn. Lett.* 7, 3102604 (2016).
- [4] Z. Zhao, M. Jamali, N. D'Souza, D. Zhang, S. Bandyopadhyay, J. Atulasimha, and J.-P. Wang, Giant voltage manipulation of MgO-based magnetic tunnel junctions via localized anisotropic strain: A potential pathway to ultra-energy-efficient memory technology, *Appl. Phys. Lett.* 109, 092403 (2016).
- [5] P. Li, A. Chen, D. Li, Y. Zhao, S. Zhang, L. Yang, Y. Liu, M. Zhu, H. Zhang, and X. Han, Electric field manipulation of magnetization rotation and tunneling magnetoresistance of magnetic tunnel junctions at room temperature, *Adv. Mater.* 26, 4320 (2014).
- [6] Y. Shiota, S. Miwa, T. Nozaki, F. Bonell, N. Mizuochi, T. Shinjo, H. Kubota, S. Yuasa, and Y. Suzuki, Pulse voltage-induced dynamic magnetization switching in magnetic tunneling junctions with high resistance-area product, *Appl. Phys. Lett.* 101, 102406 (2012).
- [7] D. Bhattacharya and J. Atulasimha, Skyrmion-mediated voltage-controlled switching of ferromagnets for reliable and energy-efficient two-terminal memory, *ACS Appl. Mater. Interfaces* 10, 17455 (2018).
- [8] L. Sandlund, M. Fahlander, T. Cedell, A. E. Clark, J. B. Restorff, and M. Wun-Fogle, Magnetostriction, Elastic moduli, and coupling factors of composite Terfenol-D, *J. Appl. Phys.* 75, 5656 (1994).
- [9] Z. Xiao, R. L. Conte, C. Chen, C.-Y. Liang, A. Sepulveda, J. Bokor, G. P. Carman and R. N. Candler, Bi-directional coupling in strain-mediated multiferroic heterostructures with magnetic domains and domain wall motion, *Sci. Rep.* 8, 5207 (2018).
- [10] D. Labanowski, A. Jung, and S. Salahuddin, Power absorption in acoustically driven ferromagnetic resonance, *Appl. Phys. Lett.* 108, 022905 (2016).
- [11] M. Weiler, L. Dreher, C. Heeg, H. Huebl, R. Gross, M. S. Brandt, and S. T. B. Goennenwein, Elastically driven ferromagnetic resonance in nickel thin films, *Phys. Rev. Lett.* 106, 117601 (2011).
- [12] V. Sampath, N. D'Souza, D. Bhattacharya, G. M. Atkinson, S. Bandyopadhyay, and J. Atulasimha, Acoustic-wave-induced magnetization switching of magnetostrictive nanomagnets from single-domain to nonvolatile vortex states, *Nano Lett.* 16, 5681 (2016).
- [13] L. Thevenard, I. S. Camara, S. Majrab, M. Bernard, P. Rovillain, A. Lemaître, C. Gourdon, and J.-Y. Duquesne, Precessional magnetization switching by a surface acoustic wave, *Phys. Rev. B* 93, 134430 (2016).
- [14] I.S. Camara, J.-Y. Duquesne, A. Lemaître, C. Gourdon, and L. Thevenard, Field-free magnetization switching by an acoustic wave, *Phys. Rev. Appl.* 11, 014045 (2019).
- [15] S. Mondal, M. A. Abeer, K. Dutta, A. De, S. Sahoo, A. Barman, and S. Bandyopadhyay, Hybrid magnetodynamical modes in a single magnetostrictive nanomagnet on a piezoelectric substrate arising from magnetoelastic modulation of precessional dynamics, *ACS Appl. Mater. Interfaces* 10, 43970 (2018).

- [16] Y. Yahagi, B. Harteneck, S. Cabrini, and H. Schmidt, Controlling nanomagnet magnetization dynamics via magnetoelastic coupling, *Phys. Rev. B* 90, 140405(R) (2014).
- [17] V. S. Vlasov, A. M. Lomonosov, A. V. Golov, L. N. Kotov, V. Besse, A. Alekhin, D. A. Kuzmin, I. V. Bychkov, and V. V. Temnov, Magnetization switching in bistable nanomagnets by picosecond pulses of surface acoustic waves, *Phys. Rev. B* 101, 024425 (2020).
- [18] A. Roe, D. Bhattacharya, and J. Atulasimha, Resonant acoustic wave assisted spin-transfer-torque switching of nanomagnets, *Appl. Phys. Lett.* 115, 112405 (2019).
- [19] T. Thomson, G. Hu, and B. D. Terris, Intrinsic distribution of magnetic anisotropy in thin films probed by patterned nanostructures, *Phys. Rev. Lett.* 96, 257204 (2006).
- [20] J. M. Shaw, S. E. Russek, T. Thomson, M. J. Donahue, B. D. Terris, O. Hellwig, E. Dobisz, and M. L. Schneider, Reversal mechanisms in perpendicularly magnetized nanostructures, *Phys. Rev. B* 78, 024414 (2008).
- [21] J. M. Shaw, H. T. Nembach, and T. J. Silva, Roughness induced magnetic inhomogeneity in Co/Ni multilayers: Ferromagnetic resonance and switching properties in nanostructures, *J. Appl. Phys.* 108, 093922 (2010).
- [22] D. Winters, M. A. Abeed, S. Sahoo, A. Barman, and S. Bandyopadhyay, Reliability of magnetoelastic switching of nonideal nanomagnets with defects: a case study for the viability of straintronic logic and memory, *Phys. Rev. Appl.* 12, 034010 (2019).
- [23] L. Thevenard, J.-Y. Duquesne, E. Peronne, H. J. Bardeleben, H. Jaffres, S. Ruttala, J.-M. George, A. Lemaître, and C. Gourdon, Irreversible magnetization switching using surface acoustic waves, *Phys. Rev. B* 87, 144402 (2013).
- [24] J. Tejada, E. M. Chudnovsky, R. Zarzuela, N. Statuto, J. C. Rosa, P. V. Santos, and A. Hernández-Mínguez, Switching of magnetic moments of nanoparticles by surface acoustic waves, *EPL* 118, 37005 (2017).
- [25] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. G. Sánchez, and B. V. Waeyenberge, The design and verification of MuMax3, *AIP Advances* 4, 107133 (2014).
- [26] D. B. Gopman, V. Sampath, H. Ahmad, S. Bandyopadhyay, and J. Atulasimha, Static and dynamic magnetic properties of sputtered Fe–Ga thin films, *IEEE Tran. Magn.* 53, 6101304 (2017).
- [27] A. E. Clark, M. Wun-Fogle, J. B. Restorff, and T. A. Lograsso, Magnetostrictive properties of galferol alloys under compressive stress, *Mater. Trans.* 43, 881 (2002).
- [28] <https://www.mtixtl.com/linbo3sawgradewafer128degy-xcut3x05mm2sp-2.aspx>.
- [29] Y. Saito, H. Takao, T. Tani, T. Nonoyama, K. Takatori, T. Homma, T. Nagaya, and M. Nakamura, Lead-free piezoceramics, *Nat.* 84, 432 (2004).

- [30] B. Zhang, J. Wu, X. Cheng, X. Wang, D. Xiao, J. Zhu, X. Wang, and X. Lou, Lead-free piezoelectrics based on Potassium–Sodium Niobate with giant  $d_{33}$ , *ACS Appl. Mater. Interfaces* 5, 7718 (2013).
- [31] C.C.W. Ruppel, L. Reindl, R. Weigel, SAW devices and their wireless communications applications, *IEEE Microw. Mag.* 3, 65 (2002).
- [32] S. Datta, J. Atulasimha, C. Mudivarthi, and A.B. Flatau, Stress and magnetic field-dependent Young's modulus in single crystal iron–gallium alloys, *J. Magn. Magn. Mater.* 322, 2135 (2010).
- [33] S. Datta, *Surface Acoustic Wave Devices* (Prentice-Hall, Englewood Cliffs, New Jersey, 1986).
- [34] C. Campbell, *Surface Acoustic Wave Devices for Mobile and Wireless Communications* (Academic Press, San Diego, 1998), p. 180.
- [35] D. Hunter, W. Osborn, K. Wang, N. Kazantseva, J. H. Simpers, R. Suchoski, R. Takahashi, M. L. Young, A. Mehta, L. A. Bendersky, S. E. Lofland, M. Wuttig and I. Takeuchi, Giant magnetostriction in annealed  $\text{Co}_{1-x}\text{Fe}_x$  thin-films, *Nat. Commun.* 2, 518 (2011).
- [36] D.C. Malocha, Surface acoustic wave design fundamentals, *Archives of Acoustics*, 21, 4, pp. 387-398 (1996).
- [37] M. Giovannini, S. Yazici, N.-K. Kuo, G. Piazza, Apodization technique for spurious mode suppression in AlN contour-mode resonators, *Sensors and Actuators A: Physical*, vol. 206, pp. 42-50 (2014)

### **Chapter 3: Voltage Controlled Energy Efficient Domain Wall Synapses with Stochastic Distribution of Quantized Weights in the Presence of Thermal Noise and Edge Roughness**

We propose energy efficient voltage induced strain control of domain wall (DW) in a perpendicularly magnetized nanoscale racetrack on a piezoelectric substrate that can implement a multi-state synapse to be utilized in neuromorphic computing platforms. Here strain generated in the piezoelectric is mechanically transferred to the racetrack and modulates the Perpendicular Magnetic Anisotropy (PMA) in a system that has significant interfacial Dzyaloshinskii–Moriya interaction (DMI). When different voltages are applied (i.e. different strains are generated) in conjunction with SOT due to a fixed current flowing in the heavy metal layer for a fixed time, DWs are translated to different distances and implement different synaptic weights. We have shown using micromagnetic simulations that 5-state and 3-state synapses can be implemented in a racetrack that is modeled with the inclusion of natural edge roughness and room temperature thermal noise. These simulations show interesting dynamics of DWs with roughness induced pinning sites both at the beginning and end of the SOT current pulse for different PMA modulation. Thus, notches need not be fabricated to implement multi-state nonvolatile synapses. Such a strain-controlled synapse has an energy consumption of  $\sim 1$  fJ and could thus be very attractive to implement energy-efficient quantized neural networks, which has been shown recently to achieve near equivalent classification accuracy to the full-precision neural networks.

Neuromorphic computing outperforms traditional von-Neumann type processors in data-intensive classification tasks. Moreover, their in-memory computing architecture can reduce energy dissipation [1] required to shuttle data back and forth between processor and memory unit in traditional computing architectures. Examples of hardware realization for neuromorphic computing include phase change random access memory (PCRAM) [2-4], resistive random-access memory (RRAM) [5,6] and spin transfer torque random-access memory (STTRAM) [7]. While device variability is a persistent issue for all of the above-mentioned devices, recent work in fully connected artificial neural network (ANN) [8] shows equivalent accuracy to software-based training. Unfortunately, PCRAM and RRAM based devices consume energy on the order of a few pJs per synaptic weight alteration event [9]. Hence, the future IoTs and edge-devices where power is limited will necessitate alternate neuromorphic hardware that are energy efficient and enable real time programming of synaptic weights so the networks can be trained in-situ.

Recently, nanomagnet based synaptic devices have shown potential to be energy efficient compared to PCRAM and RRAM [9, 10, 11]. Among nanomagnet based neuromorphic devices, domain wall (DW)



based magnetic tunnel junctions (MTJs) are one of the most promising. To implement these devices, domain walls (DWs) are translated to different positions by externally applied magnetic field [12], an electric current that causes spin-orbit torque (SOT) [13-15], spin transfer torque (STT) [16-18] or a strain gradient [19-20]. Strain control of magnetization consumes ultra-low energy [21-27]. Hence, manipulation of DWs with strain can be utilized to implement energy efficient neuromorphic devices. Recently, strain-mediated control of DW has been reported [28, 29]. Strain gradient in conjunction with SOT or STT [10] has also been proposed to control DW position to implement energy efficient synaptic devices that can be programmed in real time.

In this work, we propose to utilize SOT to translate the DW in a realistic nanoscale racetrack modeled with edge roughness and thermal noise where the DW position is controlled by modulating the perpendicular magnetic anisotropy (PMA) of the racetrack with the application of stress. Here, deterministic control of DW to realize different synaptic values is hard to achieve when different stress values are generated by applying voltage pulse of different amplitudes to the electrodes patterned on top of a piezoelectric. This is because equilibrium DW positions are often stochastic in nature and with the presence of defects [30], local imperfections [31] and thermal noise [32] it could be very difficult to achieve deterministic control. Nevertheless, the DW can be arrested by providing trap sites such as curved shape [33] and notch or protrusion [34], which can act as a potential well or barrier. Moreover, edge roughness [35-36] can introduce pinning sites for DW motion. In this study, we use edge roughness and obtain the statistical distribution of DW position from micromagnetic simulations which shows that the mean positions are different for different stress induced change of PMA for a fixed current induced SOT of a fixed “clock” time. Although the number of states (different DW positions) attained are limited and there are overlaps between the states, such a DW based racetrack as synapse is particularly attractive to implement quantized deep neural networks (DNN) [37-39] as these networks have been shown to reach accuracy very close to the infinite states network. The overlap between states can be addressed during the training stage of a learning network. Moreover, the stochastic variation of a state can be useful in generating stochastic weights for training the network which can work as DNN regularizer to reduce overfitting of training [40]. Studies have shown training with stochastically determined weights rather than deterministic ones can potentially increase the classification accuracies for some data sets [37].

### **3.1 Device architecture and simulation:**

The proposed device structure is illustrated in Fig. 3-1(a). The stack consists of a heavy metal layer and a magnetic tunnel junction (MTJ) containing the nanoscale racetrack as free layer, along with the tunnel

barrier and the hard layer. Such a stack is patterned on top of a piezoelectric substrate. We consider Pt/CoFe (soft or free racetrack layer)/MgO/CoFe (hard or fixed layer) as our stack materials where the heavy metal layer Pt will create perpendicular anisotropy and strong DMI at Pt/CoFe interface, which is known to favor the chiral Neel DWs [41]. We propose to arrest the DWs at different positions in the free layer of the MTJ, which will modify the resistance value of the MTJ stack. Thus, different synaptic weights, which define the strength between the neurons can be determined from the DW positions. Different layers of a DNN can be implemented by arranging the DW devices in the crossbar as shown in Fig. 3-1(c), where the DW devices provide the programmable conductances which are equivalent to the DNN weights.

To arrest the DW at various positions we apply different amplitude stress in combination to a fixed amplitude and fixed duration SOT pulse. When a voltage is applied between the electrodes on top and bottom of the piezo-substrate as shown in Fig. 3-1(b), mechanical strain is generated. This strain is then transferred to the racetrack and consequently modulates the perpendicular anisotropy due to magnetoelastic interaction. In combination with stress, we apply a current pulse in the adjacent heavy metal Pt layer to exert SOT shown by red arrow in Fig. 3-1(b), which moves the DW through the nanowire racetrack to the other end of the nanowire. If we reverse the direction of current in the heavy metal layer, it will reverse the direction of DW motion and reset it to the other end.

We have considered edge roughness that is present naturally in a nanoscale racetrack due to lithographic imperfection and pattern transfer process. Authors report [42] ~ 2nm rms edge roughness for 25 nm wide racetrack when they use a combination of electron beam lithography (EBL) and ion beam etching. Authors [36] also report ~ 2nm rms edge roughness for 80 nm wide racetrack using the same method. However, studies have shown higher rms edge roughness for low voltage EBL for racetrack of width 50 nm or higher [42]. For our simulation we have assumed Gaussian distribution for the edge roughness with a rms value of ~ 3nm considering the effect of high electron jitter from mean position for low voltage EBL. In addition to the local structure variation, microstructure in the racetrack such as grain boundary, defects can provide pinning sites and introduce stochasticity in the devices. We did not consider these in our simulation for sake of simplicity. The simulated racetracks have a length of 500 nm, maximum width of 50 nm and thickness of 1 nm. The magnetization dynamics in the presence of Spin Orbit Torque (SOT) is simulated in MUMAX3 [43] using the Landau–Lifshitz–Gilbert–Slonczewski equation as explained in section 1.2.7.

Here, the effective field,  $\vec{H}_{eff}$  accounts for the contributions from demagnetization, PMA, Heisenberg exchange interaction, Dzyaloshinskii–Moriya interaction (DMI), stress induced anisotropy and thermal noise.  $\vec{H}_{eff}$  can be expressed as follows:

$$\vec{H}_{eff} = \vec{H}_{anis} + \vec{H}_{demag} + \vec{H}_{stress} + \vec{H}_{exch} + \vec{H}_{thermal} \quad (1)$$

The racetracks are discretized into  $2 \text{ nm} \times 2 \text{ nm} \times 1 \text{ nm}$  cells which are well within the ferromagnetic exchange length of  $\sqrt{\frac{2A_{ex}}{\mu_0 M_s^2}} = 5.66 \text{ nm}$ . We note that curved edges are difficult to approximate with finite difference method as it depends on staircase approximation. As a result, the demagnetization tensor is not computed properly [45-47]. However, we find similar trend in our result when we decrease the cell size (section 3.4).

PMA induced effective field can be expressed as,  $\vec{H}_{anis}$  :

$$\vec{H}_{anis} = \frac{2K_u}{\mu_0 M_s} (\vec{u} \cdot \vec{m}) \vec{u} \quad (2)$$

Where  $K_u$  is the first order anisotropy constant and  $\vec{u}$  represents the uniaxial anisotropy direction (i.e. perpendicular to plane).

If the electrodes patterned on top of the piezoelectric substrate have dimensions similar to the piezoelectric thickness and separated by one or two times the piezoelectric thickness, maximum stress is generated [48]. In such a scenario, when a positive (negative) voltage is applied in the top electrode pair, the area underneath the electrode becomes stretched (compressed) in the out of plane direction and compressed (stretched) in the in-plane direction. Compression (tension) in the in-plane direction underneath the electrode surface creates tension (compression) in the nanoscale racetrack patterned in between the top electrodes due to strain-displacement compatibility. We assumed our electrodes to be rectangular with width  $b$ =piezoelectric thickness and length  $L$ =racetrack length. This is similar to having  $(L/b)$  number of square electrodes of  $(b \times b)$  dimensions and therefore one can assume this electrode configuration will produce similar amount of stress as mentioned in Ref [46]. Fig. 3-1(b) shows the strain formation in the nanoscale racetrack in such a scenario. Stress produced in the in-plane direction of the racetrack induces anisotropy field due to the magneto-elastic effect in the same direction and modulates the PMA or the anisotropy constant  $K_u$ . The effect of the stress is modeled by the modulating  $K_u$  in the micromagnetic simulation. For simplicity, we did not consider the strain that can be produced in the in-plane direction of the racetrack which is orthogonal to that shown in Fig. 3-1(b).

The effective field due to the interfacial Dzyaloshinskii-Moriya interaction is expressed as follows [43]:

$$\vec{H}_{DM} = \frac{2D}{\mu_0 M_s} \left( \frac{\partial m_z}{\partial x}, \frac{\partial m_z}{\partial y}, -\frac{\partial m_x}{\partial x} - \frac{\partial m_y}{\partial y} \right) \quad (3)$$

Here,  $D$  is the DMI constant and  $m_x$ ,  $m_y$  and  $m_z$  are the x, y and z component of unit magnetization vector  $\vec{m}$  respectively.

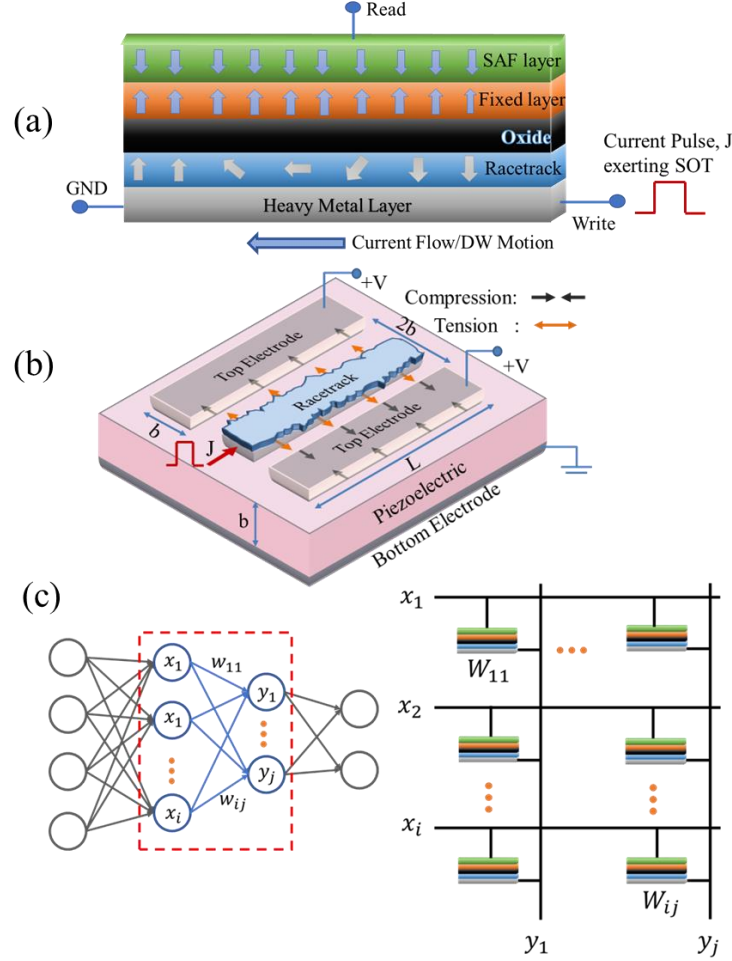


Figure 3-1 (a) Proposed device stack where the nanoscale racetrack act as the magnetic free layer of the MTJ. DW in the racetrack moves when a current is applied to the heavy metal layer underneath the racetrack (b) Stress generation mechanism in rough edge racetrack when a voltage is applied across the piezoelectric. (c) Implementation of layers of DNN with DW based synaptic devices. The devices are arranged in crossbar to provide programmable conductance equivalent to the DNN weights.

Thermal noise induces a random effective field  $\vec{H}_{thermal}$  [49]:

$$\vec{H}_{thermal} = \vec{\eta} \sqrt{\frac{2\alpha kT}{\mu_0 M_S \gamma \Omega \Delta}} \quad (4)$$

Here,  $\vec{\eta}$  is a random variable with Gaussian distribution with mean zero and unit variance and independent (uncorrelated) in each of the 3 cartesian coordinates generated at each time step,  $k$  is Boltzmann constant,

$\Omega$  is the cell volume,  $\Delta$  is the time step size. The parameters for the simulation are presented at table 3-1 [50-52].

Table 3-1: Material parameters used for the CoFe soft layer in the Pt/ CoFe/MgO heterostructure

Parameters	Values
DMI constant (D)	$0.001 \text{ Jm}^{-2}$
Gilbert damping ( $\alpha$ )	0.015
Saturation magnetization ( $M_s$ )	$10^6 \text{ Am}^{-1}$
Exchange constant ( $A_{ex}$ )	$2 \times 10^{-11} \text{ Jm}^{-1}$
Saturation magnetostriction ( $\lambda_s$ )	250 ppm
Perpendicular Magnetic Anisotropy ( $K_u$ )	$7.5 \times 10^5 \text{ Jm}^{-3}$

The synaptic state of the proposed device could be read by the MTJ. For a read voltage applied between the read and GND terminal (as in Fig. 3-1(a)) the resistance is provided by the portion of the racetrack that is parallel (P) and antiparallel (AP) to the fixed layer and a small DW region where the magnetization is transverse to the fixed layer magnetization. The read current also counters a resistance from heavy metal layer however that is small compared to the tunnel magnetoresistance. If we assume the conductance of the racetrack is  $G_{max,P}$  when completely in P state with respect to the fixed layer and  $G_{min,AP}$  when completely in AP state, then for any intermediate position  $q$  of the DW inside the racetrack of length  $L$ , the conductance of the synapse can be expressed as the following:

$$G(q) = G_{max,P} \left( \frac{q}{L} \right) + G_{min,AP} \left( 1 - \frac{q}{L} \right) + G_{DW} \quad (5)$$

## 3.2 Results and Discussion

### 3.2.1. Effect of Edge Roughness on Domain Wall Motion

In rough edge racetrack the racetrack width varies, so local pinning sites are created randomly along the length of the racetrack. Depending on the magnitude of the edge roughness (rms value or standard deviation) the pinning strength of the pinning sites varies. Studies have shown that higher magnitude edge irregularities require higher depinning current to translate DW in the racetrack [53-55]. This would become clear from Fig. 3-2, which plots the stable DW positions for 30 different nanowires of rms edge roughness

of 1.5 nm and another 30 different nanowires with edge roughness of 3 nm. For reproducibility and comparison purpose, the seed value to create the rough edges of different rms roughness nanowires are kept the same. The nanowires PMA was assumed to be  $K_u=7.5 \times 10^5 \text{ J/m}^3$ . The PMA was modulated to  $K_u=8.0 \times 10^5 \text{ J/m}^3$  with a voltage pulse (1.2 ns) and a constant amplitude SOT current pulse of  $24 \times 10^{10} \text{ A/m}^2$  was applied simultaneously for 1.2 ns. All the DWs in the nanowires are nucleated at 450 nm and relaxed before applying the SOT. After the simultaneous withdrawal of voltage and current pulse we relax for 10 ns to find the stabilized DW positions. As expected, the DWs travel higher distances with low edge roughness. However, when we double the edge roughness the DWs mostly get pinned near the initial location. Thus, the strength of pinning sites created by the edge roughness influences the stabilized DW positions in the racetrack, which also determines the operating current for the DW racetrack device.

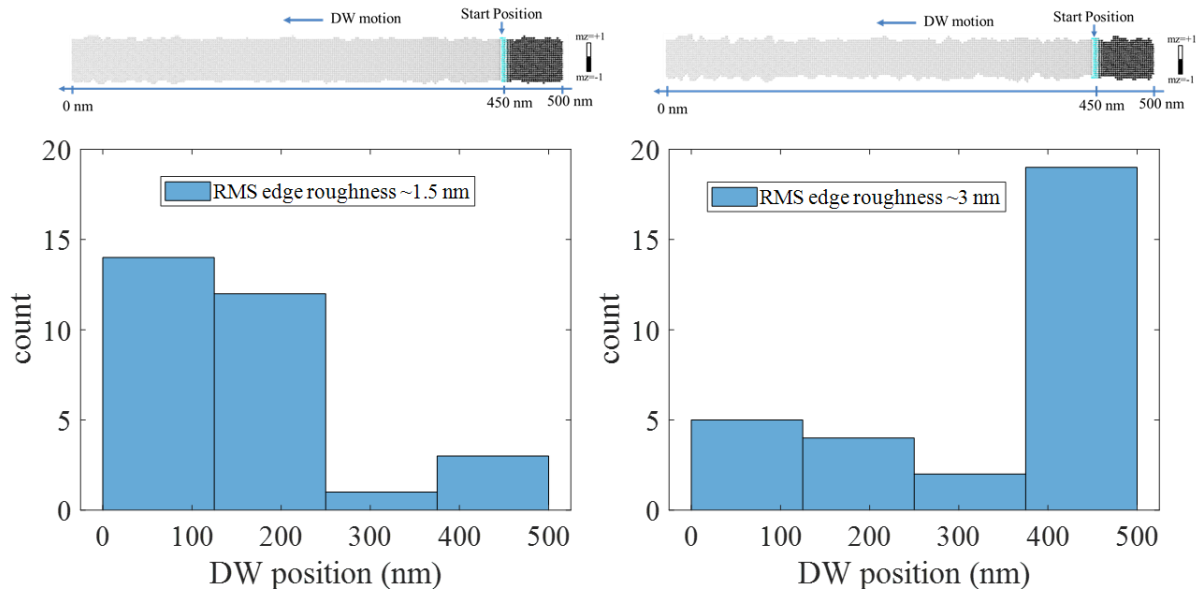


Figure 3-2 Stabilized DW position distribution for 30 different nanowires for two different rms edge roughness values. The PMA of the nanowires is  $K_u=7.5 \times 10^5 \text{ J/m}^3$  which is modulated to  $K_u=8.0 \times 10^5 \text{ J/m}^3$  with a voltage pulse of 1.2 ns. SOT current pulse of  $24 \times 10^{10} \text{ A/m}^2$  is also applied simultaneously for 1.2 ns. All the DWs are nucleated at 450 nm as seen from the top panel. After withdrawing the voltage and current pulse simultaneously the system is relaxed for 10 ns to determine the stabilized DW positions.

In addition to the rms edge roughness, the pinning location distribution or the relative position of the pinning sites from DW start position and center of the racetrack influences the final DW position. The characteristic DW motion equation during the acceleration phase (at the time of SOT excitation) can be found by linearizing the 1-D DW equations [41,56] in  $q-\psi$  axis where  $q(t)$  and  $\psi(t)$  are the DW position and DW magnetization angle in the racetrack respectively.

$$\frac{1 + \alpha^2}{\Delta} \frac{dq}{dt} = \frac{-\gamma H_K}{2} \sin 2\psi + \frac{\pi}{2} \gamma H_{DM} \sin \psi + \alpha \left( \gamma H_{PIN}(q) + \frac{\pi}{2} \gamma H_{SH} \cos \psi \right) \quad (6)$$

$$(1 + \alpha^2) \frac{d\psi}{dt} = -\alpha \left( \frac{-\gamma H_K}{2} \sin 2\psi + \frac{\pi}{2} \gamma H_{DM} \sin \psi \right) + \gamma H_{PIN}(q) + \frac{\pi}{2} \gamma H_{SH} \cos \psi \quad (7)$$

Where,  $H_{DM}$  is the DMI field,  $H_{SH}$  is the damping like spin hall effective field and  $H_K$  is the shape anisotropy field from magnetostatic origin.

$$H_{SH} = \frac{\hbar J \theta}{2 \mu_0 e d M_s} \quad (8)$$

$$H_{DM} = \frac{D}{\mu_0 \Delta M_s} \quad (9)$$

Here,  $J$  is the value of current flowing through the heavy metal layer,  $\theta$  is the spin Hall angle,  $\alpha$  is the damping,  $\gamma$  is the gyromagnetic ratio,  $M_s$  is the saturation magnetization,  $\hbar$  is the reduced Planck constant,  $\mu_0$  is the permeability of free space,  $e$  is the electron charge and  $d$  is the thickness of the nanowire racetrack.

The physical width of DW is  $\pi\Delta$ , and  $\Delta$  can be expressed as:

$$\Delta \sim \sqrt{\frac{A_{ex}}{K_u - \frac{1}{2} \mu_0 M_s^2}} \quad (10)$$

When a current pulse is applied, the DW starts to accelerate and goes from Neel configuration of  $\psi = 0, \pi$  to equilibrium Bloch configuration of  $\psi = \frac{\pi}{2}, -\frac{\pi}{2}$ . Linearizing the above Eq. 1 and 2 with respect to  $\psi = \frac{\pi}{2}$  we can get the following equation:

$$\frac{1 + \alpha^2}{\gamma^2 \Delta} \frac{d^2 q}{dt^2} + \frac{1}{\gamma \Delta} \left( \alpha H_K + \frac{\pi}{2} H_{SH} \right) \frac{dq}{dt} = H_{PIN}(q) H_K + \left( \frac{\pi}{2} \right)^2 H_{DM} H_{SH} \quad (11)$$

This characteristic DW equation is analogous to the Newton-like motion equation with DW velocity,  $v$  :

$$m^* \frac{dv}{dt} + \frac{m^*}{\tau} v = F \quad (12)$$

Where the effective DW mass can be expressed as:

$$m^* = \frac{1 + \alpha^2}{\gamma^2 \Delta} \quad (13)$$

The friction force is:

$$F_{fric} = \frac{m^*}{\tau} v = \frac{1}{\gamma\Delta} \left( \alpha H_K + \frac{\pi}{2} H_{SH} \right) v \quad (14)$$

And the external force is:

$$F = H_{PIN}(q)H_K + \left(\frac{\pi}{2}\right)^2 H_{DM}H_{SH} \quad (15)$$

The pinning field can be expressed as:

$$H_{PIN}(q) = -\frac{1}{2\mu_0 M_s w d} \frac{d[V_{PIN}(q)]}{dq} \quad (16)$$

Where  $V_{PIN}(q)$  is the local pinning potential due to the roughness induced pinning locations and  $w$  is the racetrack width,  $d$  is the thickness.

From the linearized motion equation of the DW we can see that the roughness induced pinning sites induces an attractive force towards the pinning site scaled by the magnetostatic field ( $H_{PIN}(q)H_K$  term in external force equation). This force is added to the SOT current induced force due to  $H_{SH}$ . The lower the pinning potential,  $V_{PIN}(q)$ , the higher the pinning strength and the attractive force exerted to the DW towards the pinning sites. Thus, the pinning sites exert attractive force and help to accelerate the DW. However, the DW takes time to accelerate, and the acceleration time constant can be expressed as:

$$\tau = \frac{1 + \alpha^2}{\gamma \left( \alpha H_K + \frac{\pi}{2} H_{SH} \right)} \quad (17)$$

The acceleration time constant is calculated to be,  $\tau \sim 0.3$  ns when we assume  $K_u = 8.0 \times 10^5 \frac{J}{m^3}$  using,  $H_K = \frac{d \log(2) M_s}{\pi \Delta}$  [57].

As the DW can be associated with an effective mass and it takes finite time to accelerate, the distance of the pinning sites from DW starting position is important. If the DW starting position is close to the pinning site it may not attain enough kinetic energy to overcome that pinning site and thus can be pinned to relatively weaker pinning sites. The relationship between the distance of the pinning site to the DW and the required kinetic energy to overcome that pinning site depends mostly on the competition between demagnetization field,  $H_K$  and pinning site induced field,  $H_{PIN}(q)$  (as indicated by  $H_{PIN}(q)H_K$  term in external force,  $F$  equation). The demagnetization field is maximum at both ends of the rectangular racetrack and starts to decrease and becomes minimum at the center of the racetrack. Similarly, the pinning field induced attraction force is high (low) away from (close to) the pinning site, however the range of this force is much more localized than the demagnetization force. In Fig. 3-3 we plot the depinning current vs the relative distance,



$x_d$  of a pinning location (created by a triangular notch) from the starting position of DW. For higher PMA, such as  $K_u = 8.0 \times 10^5 \frac{J}{m^3}$  the pinning potential is lower than demagnetization (magnetostatic) potential and thus the pinning sites exerts the dominating attractive force. In such case we see from Fig. 3-3 that the depinning current decreases with the distances,  $x_d$  as the DWs can attain sufficient kinetic energy before interacting with the pinning sites. However, with PMA,  $K_u = 7.5 \times 10^5 \frac{J}{m^3}$  or lower, the trend flips. In this case, the attractive force due to demagnetization dominates over the pinning sites and creates an energy well at the center which always pulls the DW towards the center. However, for  $K_u = 7.8 \times 10^5 \frac{J}{m^3}$ , a mixed trend is found. At the center portion of the racetrack the demagnetization field becomes low and the pinning site induced field dominates, as a result the depinning current decrease with distance,  $x_d$ . However, at both ends of the racetrack the demagnetization field dominates. At the near end of the racetrack DW gets help from the demagnetization field to overcome the pinning sites but requires higher energy to overcome both demagnetization and pinning site induced field at the far end.

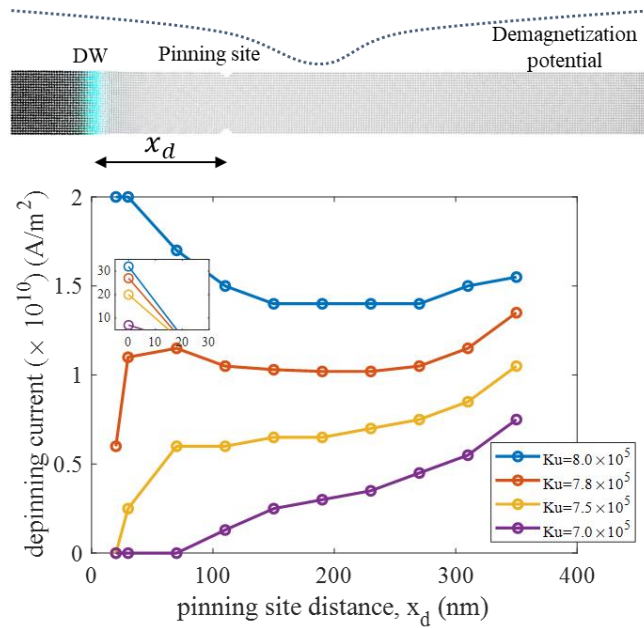


Figure 3-3 DW depinning current with respect to the relative distance,  $x_d$  between the pinning location and the DW starting position for four different PMA ( $K_u$ ). Racetrack of dimension 500 nm x 50 nm with the DW and pinning site (triangular notch) is shown above with a sketch of demagnetization potential.

Thus, in addition to the rms amplitude of the edge roughness the relative position of the pinning sites from the DW starting position influences the stabilized DW position in a nanowire racetrack.

Now we will look at the deceleration phase of DW. After withdrawing the SOT current pulse, DW goes from Bloch configuration of  $\psi = \frac{\pi}{2}, -\frac{\pi}{2}$  to equilibrium Neel configuration of  $\psi = 0, \pi$ . The characteristic equation of DW during deceleration phase can be found by linearizing the 1-D DW equations with respect to  $\psi = 0$ . Linearizing the 1-D equations with respect to  $\psi = 0$  and assuming  $H_{SH} = 0$  we get,

$$\frac{1 + \alpha^2}{\gamma^2 \Delta} \frac{d^2 q}{dt^2} + \frac{\alpha}{\gamma \Delta} \left( H_K - \frac{\pi}{2} H_{DM} \right) \frac{dq}{dt} = - \left( H_K - \frac{\pi}{2} H_{DM} \right) H_{PIN}(q) \quad (18)$$

The deceleration force can be expressed by:

$$F = - \left( H_K - \frac{\pi}{2} H_{DM} \right) H_{PIN}(q) \quad (19)$$

The deceleration time constant is found to be,

$$\tau = \frac{1 + \alpha^2}{\alpha \gamma \left| H_K - \frac{\pi}{2} H_{DM} \right|} \quad (20)$$

The deceleration time constant is calculated to be,  $\tau \sim 3.16$  ns when we assume  $K_u = 8.0 \times 10^5 \frac{J}{m^3}$ . Thus, the DW takes approximately  $3 \times \tau$  to decelerate and settle to an equilibrium position which is approximately 10 ns. We have also found a similar trend from our simulation where the DW settles after 10 ns of SOT withdrawal. From the above linearized equation, the deceleration force is generated by pinning field  $H_{PIN}(q)$ , scaled by the difference of magnetostatic field,  $H_K$  and DMI field  $H_{DM}$ . For a fixed PMA, depending on the position of the DW at the time of SOT withdrawal,  $H_{PIN}(q)$  and  $H_K$  acted upon the DW changes (both functions of position), as a result the deceleration force changes. Thus, the position of the DW at the end of the SOT current pulse also influences the stabilized DW position at the racetrack.

### 3.2.2 Non-thermal statistics due to different edge roughness profiles in different racetracks:

For non-thermal simulations, we simulated the DW motion in 40 different racetracks with different edge roughness profiles. The PMA of the racetracks is considered to be  $7.5 \times 10^5 J/m^3$ . The PMA can be decreased or increased uniformly over the whole racetrack by applying a suitable voltage to the electrodes. The clocking SOT current is applied simultaneously with this voltage pulse. We have assumed that the DW is initialized to a pinning site located at one end of the racetrack. The SOT current translates the DW while the PMA modulation helps to drive the DW to different positions when clocked with SOT for a fixed time. This could be explained as follows. The critical depinning current density  $J_C$  of the DW is related to the anisotropy coefficient  $K_u$  of the racetrack. When  $K_u$  is higher, the potential well of a pinning site becomes

deep, so it requires high depinning current,  $J_C$  to depin a DW sitting in such a potential well or energy minima. On the contrary, lower  $K_u$  is associated with a shallow potential well for the same pinning site hence requires lower threshold current to depin. Fig. 3-4(a) presents a sketch of an example racetrack where the DW is situated at a pinning site located near the right end of the racetrack and Fig. 3-4(b) plots the depinning current versus the anisotropy coefficient for that DW. From Fig. 3-4(b) we can see that critical depinning current  $J_C$  is increased with the increase of anisotropy coefficient  $K_u$ .

The DW velocity at steady state can be expressed by the following [56,58]:

$$v = \frac{\pi}{2} \frac{\gamma \Delta H_{DM}}{\sqrt{(1 + (\frac{J_D}{J - J_C})^2)}} \quad (21)$$

$$J_D = \alpha J H_{DM} / H_{SH} \quad (22)$$

Empirical critical current density  $J_C$  is used to account for the pinning effect which is validated by fitting 1-D DW model to the experimental data [56].

As seen from Fig. 3-4(b), the critical current density  $J_C$  is high for higher  $K_u$ . As a result, for a higher  $K_u$ , for a fixed clocking SOT current  $J > J_C$ , the velocity becomes small as the denominator in Eq. 21 is large compared to the case of lower  $K_u$  for which the denominator is small (low critical current density  $J_C$ ) and velocity is high. In addition, when  $K_u$  increases (decreases) the DW width  $\Delta$  in Eq. 10 decreases (increases) which increases (decreases)  $J_D$  in Eq. 22 and the denominator in Eq. 21, consequently the velocity decreases (increases).

The DW position for different anisotropy constant  $K_u$  is shown in Fig. 3-4(c) for one rough edge racetrack where the SOT current of  $24 \times 10^{10} \text{ A/m}^2$  is applied for fixed 1.2 ns. The change in velocity with the change in  $K_u$  is evident as the DW translates to different distances with the same SOT. After the simultaneous withdrawal of the SOT and strain, the DW further moves at terminal velocity due to the momentum gained because of the SOT. The lower the anisotropy constant the higher the velocity gain and the higher the distance travelled by the DW after the withdrawal of SOT as can be seen for the case of  $K_u = 7.3 \times 10^5 \text{ J/m}^3$ . Notably, the DW for  $K_u = 7.0 \times 10^5 \text{ J/m}^3$  also traveled same distance as  $K_u = 7.3 \times 10^5 \text{ J/m}^3$  as the velocity difference after SOT withdrawal is small and there is no suitable pinning site in between to pin and stop the DW at a different position due to the small velocity difference.

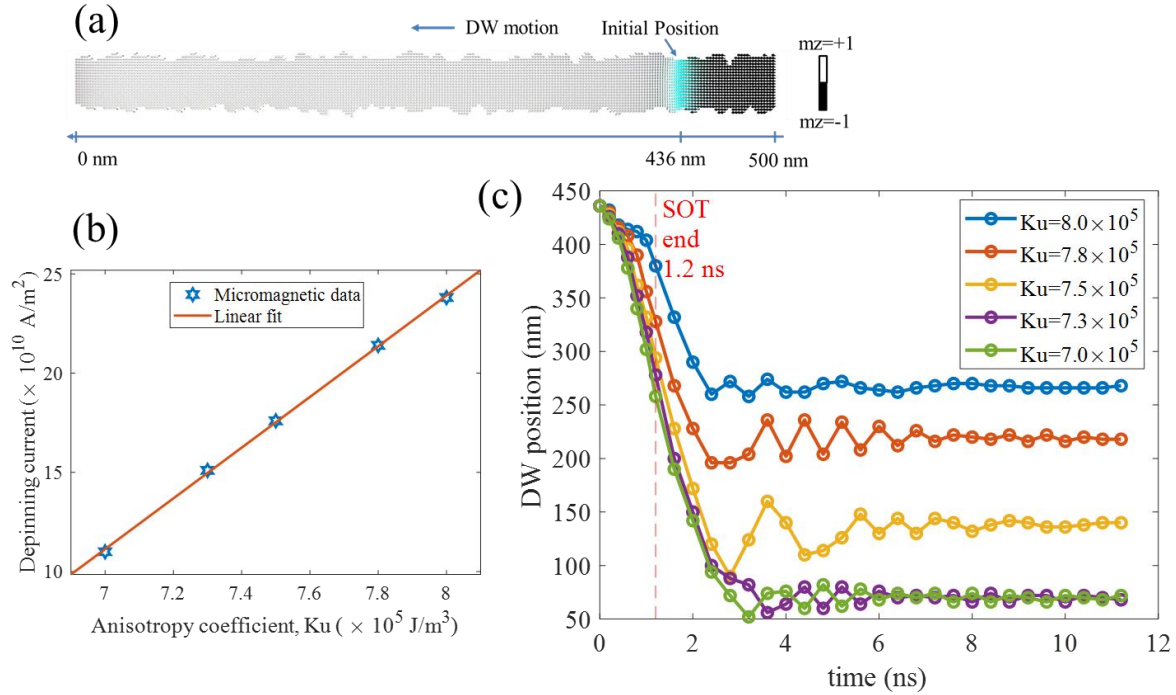


Figure 3-4 a. Initial pinning position of the DW in a PMA rough edge racetrack b. dependence of the DW depinning current on the anisotropy coefficient when the DW in racetrack 3-4a is in the initial pinning position c. DW positions with time in racetrack 3-4a for a fixed duration and amplitude current pulse exerting SOT and different stresses (different  $K_u$ ). The SOT and stress are withdrawn at 1.2 ns. For different stresses respective DWs travel different distances and get pinned to different locations.

We have simulated a total of 40 racetracks of  $\sim 3$  nm rms edge roughness where we varied anisotropy constant values  $K_u$  to 8.0, 7.8, 7.5, 7.3 and 7.0 ( $\times 10^5$ ) J/m $^3$  in each of these racetracks and applied SOT current of fixed amplitude  $24 \times 10^{10}$  A/m $^2$  for 1.2 ns. Each of the DW is initialized to a pinning site located near the right end of the racetrack. After the simultaneous withdrawal of the SOT and stress we wait for 10 ns to allow sufficient time for the DW to decelerate and get pinned to a specific position. We note that, the DWs usually settle within approximately  $\pm 4$  nm of the equilibrium pinning locations after 10 ns of SOT withdrawals which is approximately  $3\times$  of the deceleration time constant calculated from 1-D DW equations. The distribution of the final DW position for the 40 racetracks is shown in Fig. 3-5.

In Fig. 3-5 for each  $K_u$  value we also overlay a gaussian distribution with identical mean and standard deviation of the data used to create the bins. Although the final position distribution does not follow Gaussian distribution, we see that the mean final positions are different for different stress ( $K_u$ ) values (Fig. 3-5(a)-(e)). The mean DW positions shift to the left of the racetracks as we decrease the PMA. The primary source of the distribution of final DW positions for a specific  $K_u$  could be attributed to the interaction of the DWs with the roughness induced pinning sites during the acceleration and deceleration phases of DW motion. During the acceleration phase the kinetic energy (or SOT current) required to overcome a pinning

site depends on the relative distance of the DW from the pinning sites. Different racetracks offer pinning sites at different locations, thus influencing the equilibrium DW positions distribution. Similarly, during the deceleration phase, DW loses momentum due to damping and begins to interact strongly with the edges due to the deceleration force exerted towards the roughness induced pinning sites (as seen from Eq. 19). DW-edge interaction varies among racetracks due to their different roughness profile (distribution of pinning sites is different). Moreover, for different racetracks the DWs begin deceleration from different positions so the deceleration forces acted on the DWs become different. All these factors contribute to the DWs being pinned at random positions for different racetracks. In addition to that, DWs in different racetracks are initialized from pinning sites that have different longitudinal positions and geometry for different racetracks. Pinning site geometry affects the depinning current  $J_C$  vs.  $K_u$  relationship and thus different geometry can add stochasticity to the final DW position. Adding a fixed geometry notch at one end of the racetracks for DW initial location could address this stochasticity (though it cannot be addressed fully due to different stray fields for different racetracks). However, more importantly, significant stochasticity still persists (in spite of the notch to have the same initial DW starting point) due to the above-mentioned DW-edge interaction both at the beginning and end of SOT excitation.

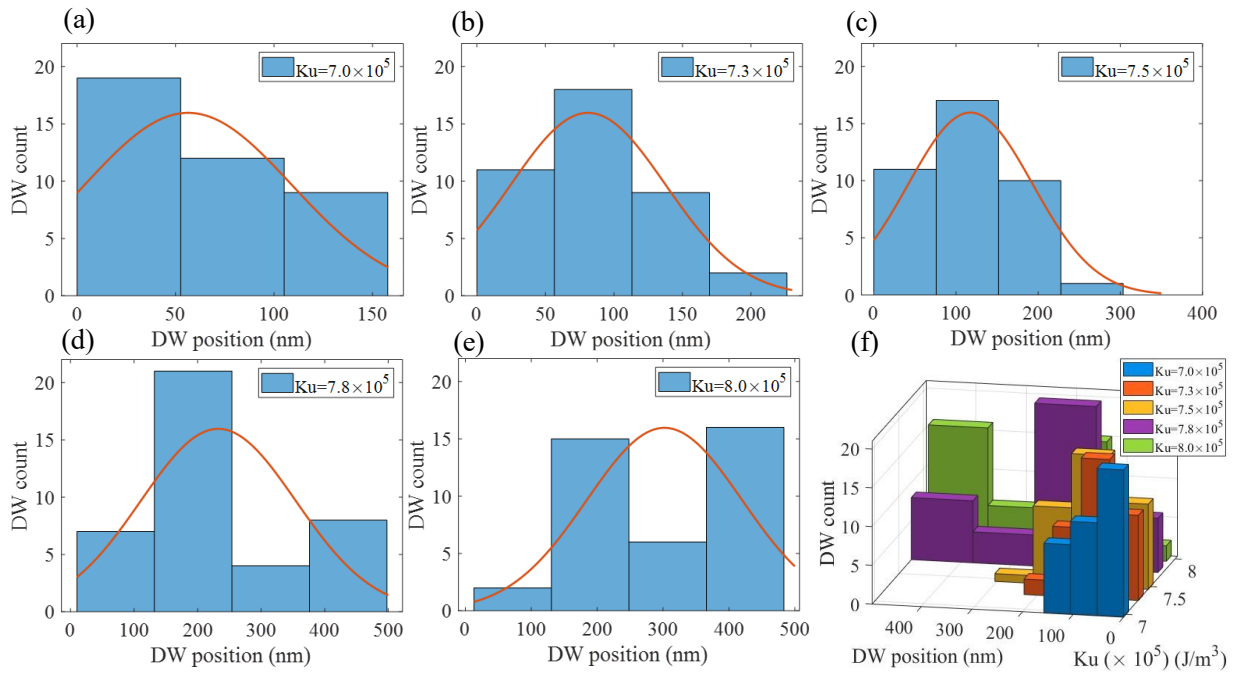


Figure 3-5 a-e. Equilibrium DW positions for 40 different racetracks at T=0 K for a fixed SOT and different stresses correspond to  $K_u$  values of 8.0, 7.8, 7.5, 7.3 and 7.0 ( $\times 10^5$ ) J/m<sup>3</sup>. For each figure in 3-5(a-e) a Gaussian distribution plot is overlaid having a mean and standard deviation identical to the data used to create the bins f. combined plot of 3-5(a-e) shows different mean positions for different  $K_u$  values.

### 3.2.3. Thermal statistics:

At room temperature, the thermal perturbation can dislodge the DW. Hence, edge roughness of  $\sim 3$  nm cannot offer similar pinning effect in thermal cases as in the non-thermal cases. As a result, the depinning current decreases in the presence of room temperature thermal noise for the same racetrack. For thermal simulation, we use a fixed clocking SOT current density of  $15 \times 10^{10} A/m^2$  which is smaller than the current density we use in non-thermal case. The SOT and stress application time are kept the same as before (1.2 ns). After the withdrawal of SOT and stress, we relax for 10 ns (as we did earlier for the non-thermal case). Unlike non-thermal cases, the DWs do not settle to a specific pinning site but oscillate around this pinning site as the thermal energy causes the DW position to fluctuate around the equilibrium position. We found that DWs usually encounter a pinning site within 6 ns of SOT withdrawal. So, a relaxation time of 10 ns is enough for the DWs to reach an equilibrium position. We changed the anisotropy constant,  $K_u$  values to 8.0, 7.8, 7.5, 7.3 and 7.0 ( $\times 10^5$ )  $J/m^3$  and ran the simulation for each  $K_u$  value 100 times considering limited computational resources and time. The equilibrium DW position distribution for one such racetrack of  $\sim 3$  nm rms edge roughness is shown in Fig. 3-6. Here, we also overlay Gaussian distribution with identical mean and standard deviation of the data used to create the bins. The bins in Fig. 3-6(a)-(e) are sized according to the standard deviation of the data. Although the distribution does not follow Gaussian distribution, the mean positions for different  $K_u$  follow the same trend as in non-thermal case where for lower  $K_u$  values the mean DW position shifts to the left. Due to the random variation of the DW internal magnetization angle in the presence of thermal noise, upon encountering a potential barrier (or a well), the DW could overcome the barrier (or gets attracted to the well) in some cases but not in other cases. This leads to a distribution.

The settling time of 10 ns for the DW or a total write time 11.2 ns may indicate a slower device compared to SOT-MRAM based memory device where low switching time is expected. However, for hardware implementation of DNN, 11.2 ns write time is not considered too slow, as different layers in DNN are implemented with separate crossbars (as shown in Fig. 3-1(c)) thus can take advantage of parallel operation. Performing the weighted sum operation during the forward and backward pass of DNN consumes time (read operation), so does the activation function computation. Thus, when a crossbar implements forward pass or backward pass of one layer, the other crossbar devices can be programmed (write operation) to achieve target conductance values.

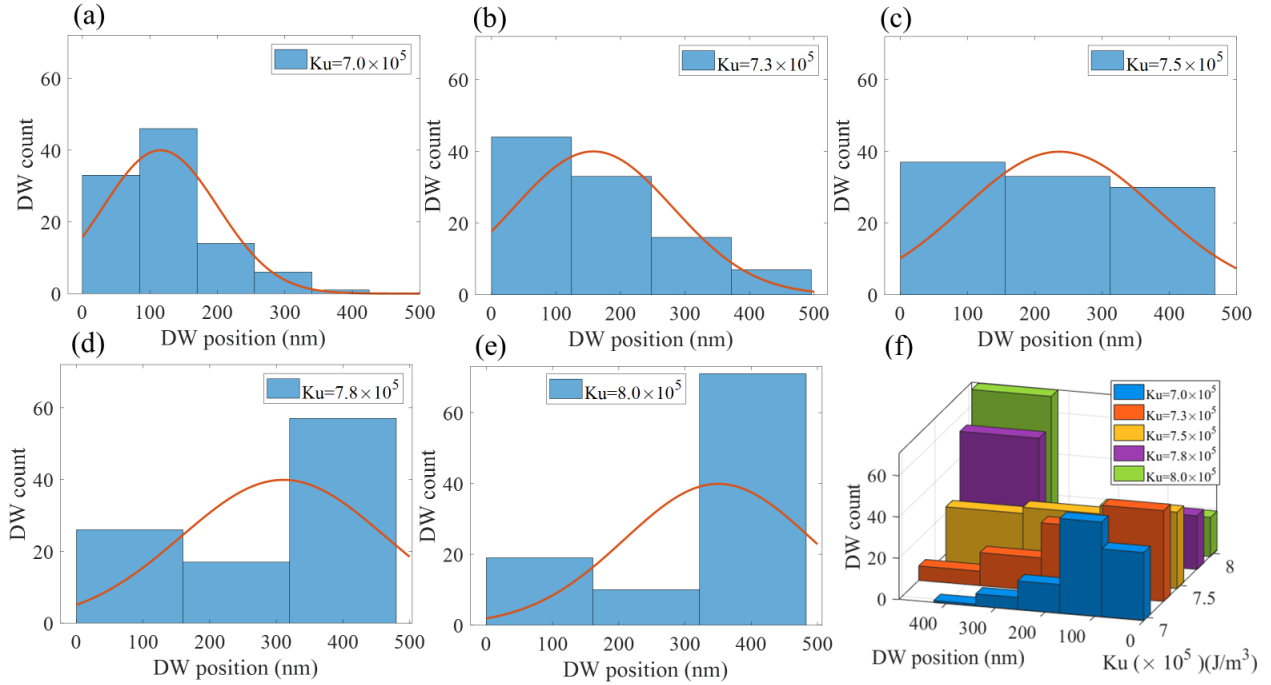


Figure 3-6 a-e. Equilibrium DW positions for one racetrack at  $T=300\text{K}$  for a fixed SOT and different stresses correspond to  $K_u$  values of  $8.0, 7.8, 7.5, 7.3$  and  $7.0 (\times 10^5) \text{ J/m}^3$ . For each figure in 3-6(a-e) a Gaussian distribution plot is overlaid having a mean and standard deviation identical to the data used to create the bins f. combined plot of 3-6(a-e) shows different mean positions for different  $K_u$  values.

### 3.2.4. Determination of Synaptic State

If the number of target states are  $n$ , and the maximum and the minimum conductance of the racetrack are  $G_{max,P}$  and  $G_{min,AP}$ , then  $\sim(G_{max,P} - G_{min,AP})$  can be divided into  $n - 1$  parts to represent one state. In such a scenario, the target conductances for each of the  $n$  states can be  $\sim G_{min,AP}, G_{min,AP} + \frac{G_{max,P} - G_{min,AP}}{n-1}, G_{min,AP} + 2 * \frac{G_{max,P} - G_{min,AP}}{n-1}, \dots, G_{max,P}$ . For any programming voltage pulse, representing by a specific PMA or  $K_u$ , the probability by which any stabilized DW provides conductance  $G$  that is within the range of target conductance  $G_T$  such that  $|G - G_T| < \frac{1}{2} \frac{G_{max,P} - G_{min,AP}}{n-1}$ , is the probability of that state for that programming condition. Fig. 3-7(a) and 3-7(b) plot the cumulative probability of DW device conductance at  $T=300 \text{ K}$  for five and three different programming conditions that implement 5- and 3-state synapse. For the conductance calculation, Eq. 5 is used and the resistance area product and TMR are assumed to be  $4.04 \times 10^{-12} \Omega m^2$  and 120 % [9]. The value of  $G_{DW}$  is small and neglected for calculation. In Fig. 3-7, the black dotted lines represent the target conductance of a state, and the adjacent red dotted lines represent the state boundaries. For 5-state synapse the target conductances are chosen to be 3.22, 3.86, 4.5, 5.14 and 5.78 mS which can be achieved by modulating the PMA to  $8.0, 7.8, 7.5, 7.3$  and  $7.0 (\times 10^5) \text{ J/m}^3$  respectively.

For 3-state synapse the target conductances are chosen to be 3.22, 4.5 and 5.78 mS. Ideally one would want 100% probability for a state for one programming condition or a specific  $K_u$ . However, in the case of stochastic DW, we get a finite probability for all the states for one programming condition. This leads to overlap of states which could degrade the DNN accuracy. These overlaps can be easily addressed by restricting the conductance of a state within the range of a target conductance (given by the adjacent red lines) by programming and then sensing or performing read-verify-write operation in a loop [58]. “Closed loop on device” [60] method can be used to perform read-verify-write operation for on-chip learning and “open loop off device” [61] method can be used for off-chip learning where the target conductance values are calculated beforehand by training a precursor neural network. Comparing Fig. 3-7(a) and (b) we can see that the state boundary is wide for 3-state synapse, thus one state can be programmed with smaller number of attempts.

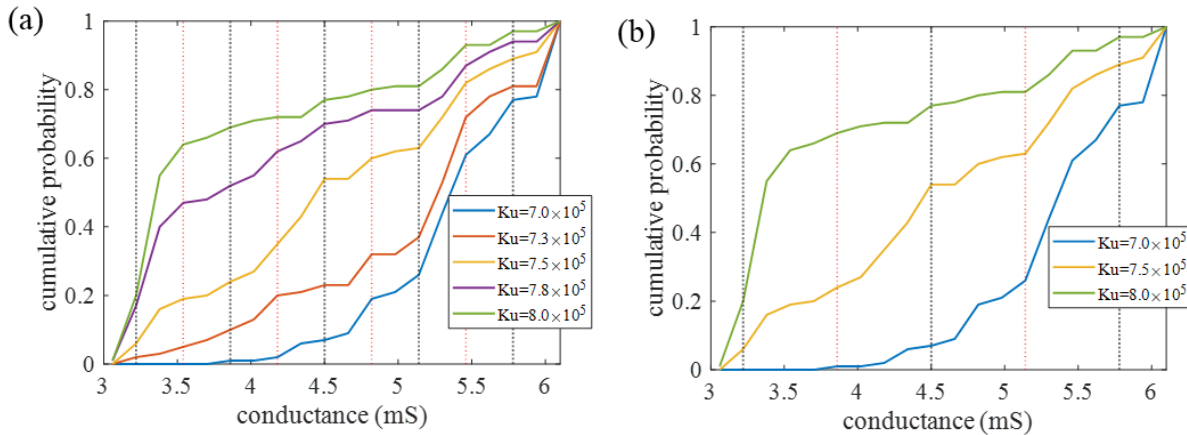


Figure 3-7 (a) Cumulative probability of device conductance for 5 different programming conditions (different  $K_u$ ) implementing a 5-state stochastic synapse. The black dotted lines represent the 5 target conductances for the 5-state synapse. The red dotted lines represent the boundaries of each state to ensure that no overlap happens between adjacent states. (b) Cumulative probability of device conductance for 3-state synapse. The red dotted lines represent the state boundaries. of each state. For 3-state synapse the width of state boundary is high so one state can be programmed with a smaller number of attempts.

While the nanoscale racetrack could be used as a synaptic device after addressing the state overlap issue, however, the presence of device-to-device variation (as in Fig. 3-5) and intra-device variation (as in Fig. 3-6) are also evident. Intuitively such variation could be harmful to the functioning of the DW based synaptic device as an inference engine for classification task, as the synaptic weights obtained after software-based training cannot be programmed accurately during inference stage. However, recent studies [40] have shown that addressing the device variability during the training stage can achieve high inference accuracies that is very close to baseline accuracy (no device variability is assumed) and the accuracy is highest when the level



of noise (because of the device variability) injected during the training is on the same order as the noise of the device used for the inference task.

### 3.3 Energy dissipation:

Energy dissipation in our proposed device depends on charging the piezoelectric layer as well as  $I^2R$  loss of the clocking current through the heavy metal layer. To introduce stress, we have to charge the piezoelectric layer. Energy required to charge this capacitive layer is  $\frac{1}{2} CV^2$ , where  $V$  is the voltage applied and  $C$  is the capacitance of the piezoelectric layer between the metal contacts.

In our proposed device, the racetrack PMA we have considered is  $K_u = 7.5 \times 10^5 \text{ J/m}^3$  and the maximum change of PMA with voltage induced stress is  $\Delta PMA = 0.5 \times 10^5 \text{ J/m}^3$  to achieve  $K_u = 7.0$  or  $8.0 (\times 10^5) \text{ J/m}^3$ . The saturation magnetostriction of CoFe is,  $\lambda_s = 250$  ppm. Using the above values, the maximum amount of required stress,  $\sigma$  is calculated to be,  $\frac{\Delta PMA}{3/2\lambda_s} = 133 \text{ MPa}$ . For CoFe with Young's Modulus of 200 GPa, the required strain is,  $\frac{133 \text{ MPa}}{200 \text{ MPa}} \sim 10^{-3}$ . Previous study [48] showed that  $10^{-3}$  strain is possible in Lead Zirconate Titanate (PZT) piezoelectric with an applied electric field of  $E = 3 \text{ MVm}^{-1}$  when the electrode dimensions are in the same order of the PZT thickness. If we consider our PZT layer to be  $b = 50 \text{ nm}$  thick (same as top electrode or racetrack width as shown in Fig. 3-1(b)) then a voltage of,  $E \cdot b = 0.15 \text{ V}$  applied at the top electrode pair can generate the required strain. If the top electrode length  $L = 500 \text{ nm}$  (same as racetrack length  $500 \text{ nm}$ ) and width  $b = 50 \text{ nm}$  is considered, and relative permittivity of PZT is  $\epsilon_r = 3000$  then the effective capacitance is calculated to be  $\frac{\epsilon_0 \epsilon_r (L \cdot b)}{b} \sim 13.3 \text{ fF}$ . This suggests a  $\frac{1}{2} CV^2$  loss of  $\sim 0.3 \text{ fJ}$  considering two top electrodes on both sides of the racetrack.

For our SOT clocking, we assume resistivity of Pt layer is  $100 \text{ } \Omega \text{ nm}$ . We also assume Pt layer to be  $5 \text{ nm}$  thick, which is greater than the spin diffusion length of  $\sim 2 \text{ nm}$  [44] and the spin hall angle to be  $0.1$  [44]. If a clocking current density of  $24 \times 10^{10} \text{ A/m}^2$  is applied through the Pt layer of length  $500 \text{ nm}$ , width  $50 \text{ nm}$  and thickness  $5 \text{ nm}$  for a clocking period of  $1.2 \text{ ns}$ , then the  $I^2R$  loss incurred is calculated to be  $\sim 0.86 \text{ fJ}$ . Therefore, our proposed DW based device can program the synapse with maximum energy dissipation of approximately  $1.16 \text{ fJ}$ .

Energy consumption to program the proposed synapse to the maximum (or minimum) conductance value is  $1.16 \text{ fJ}$  which is much less than previously reported [10,11]. Recent study has shown DW based synapse with racetrack dimension of  $1000 \text{ nm} \times 50 \text{ nm}$ , where each synaptic state is programmed by applying SOT current pulse for  $3 \text{ ns}$  [9]. In their device they require  $\sim 8.64 \text{ fJ}$  to program the synaptic conductance from

one extreme to the other. While the state-of-the-art phase change memory (PCM) device and the metal oxide resistive random-access memory (RRAM) device can have a smaller footprint, however, the programming energy can be as high as several pJs [9, 62] because these devices involve physical movement of ions. Moreover, the endurance cycle of the of the PCM and the RRAM devices are low compared to spintronic DW devices [63].

### 3.4. Additional details:

#### Demagnetization energy and grid size dependence of simulation results:

Edges in a rough racetrack are difficult to approximate with finite difference method [43] as it relies on stair-case approximation. So, the demagnetization tensor is not calculated properly at the edges. Ref. [45] and ref. [46] prescribed corrections in finite-difference code to compensate for the effects of stair-case approximation. Ref. [47] provided corrections using factors that are computed with finer mesh before simulation with actual (larger) mesh to accurately account for the short-range magnetostatic interaction.

To investigate the extent of magnetostatic interaction, we examined the effect of cell sizes on magnetostatic (demagnetization) and exchange energy densities by performing simulations for different cell sizes of  $2 \times 2 \times 1 \text{ nm}^3$ ,  $1 \times 1 \times 1 \text{ nm}^3$ ,  $0.5 \times 0.5 \times 1 \text{ nm}^3$  and so on. We chose a rectangular region of dimension  $60 \times 50 \times 1 \text{ nm}^3$  as seen in Fig. 3-8 where the length of the rectangle is comparable with physical DW length. The rectangular region is centered around the DW. For each of the cell sizes, we initialized a DW in the racetrack at a particular pinning location and after relaxation we noted the energy densities for that rectangular region. The energy densities are presented in table 3-2. Form table 3-2, we can see that the difference in demagnetization energies becomes small at cell size  $0.5 \times 0.5 \times 1 \text{ nm}^3$  and beyond. We performed the simulation for our reported results using a cell size of  $2 \times 2 \times 1 \text{ nm}^3$ . Reducing it to  $0.5 \times 0.5 \times 1 \text{ nm}^3$  would require replacing 1 cell with 16 finer cells which require significant computational resources. However, we find similar trends for the final DW position when we reduce the cell size. We performed a comparison study for the cell size of  $2 \times 2 \times 1 \text{ nm}^3$  and  $1 \times 1 \times 1 \text{ nm}^3$ .



Figure 3-8 A rectangular region of dimension  $60 \times 50 \times 1 \text{ nm}^3$  marked in red is centered around the DW in a racetrack. The energy densities are calculated for the red rectangular region by changing the cell sizes of the simulation. The computed energy densities are shown in table 3-2.

Table 3-2: Energy densities for simulations with different cell sizes

Energy	$2 \times 2 \times 1$	$1 \times 1 \times 1$	$.5 \times .5 \times 1$	$.25 \times .25 \times 1$
Density ( $J/m^3$ )	$nm^3$	$nm^3$	$nm^3$	$nm^3$
Anisotropy	-58683	-58791	-58037	-58005
Exchange	1023	1065	1012	1017
Demagnetization	46697	46867	46339	46341
Total	-10962	-10859	-10685	-10646

For both of the cell sizes, we initialized the DW at the same position in each of the 40 racetracks and applied the same current pulses (same amplitude and duration). We computed the equilibrium DW position distribution for the two extremes of the PMA,  $K_u = 8.0 \times 10^5$  and  $7.0 \times 10^5 J/m^3$ . For both of the cell sizes studied, DWs translate to longer distances for lower PMA. The distributions are presented in Fig. 3-9 and Fig. 3-10. The mean values that are computed from the distributions are close for both cell sizes for both of the PMAs (anisotropy coefficients,  $K_u$ ). For  $K_u = 8.0 \times 10^5 J/m^3$  the mean values for DW positions are 244.8 nm and 259.5 nm for  $1 \times 1 \times 1 nm^3$  and  $2 \times 2 \times 1 nm^3$  cell sizes respectively. For  $K_u = 7.0 \times 10^5 J/m^3$  the mean values for DW positions are calculated to be 65.7 nm and 87.85 nm for  $1 \times 1 \times 1 nm^3$  and  $2 \times 2 \times 1 nm^3$  cell sizes respectively. Thus, the mean values for respective anisotropy coefficients,  $K_u$  for different cell sizes follow similar trend for equilibrium DW positions.

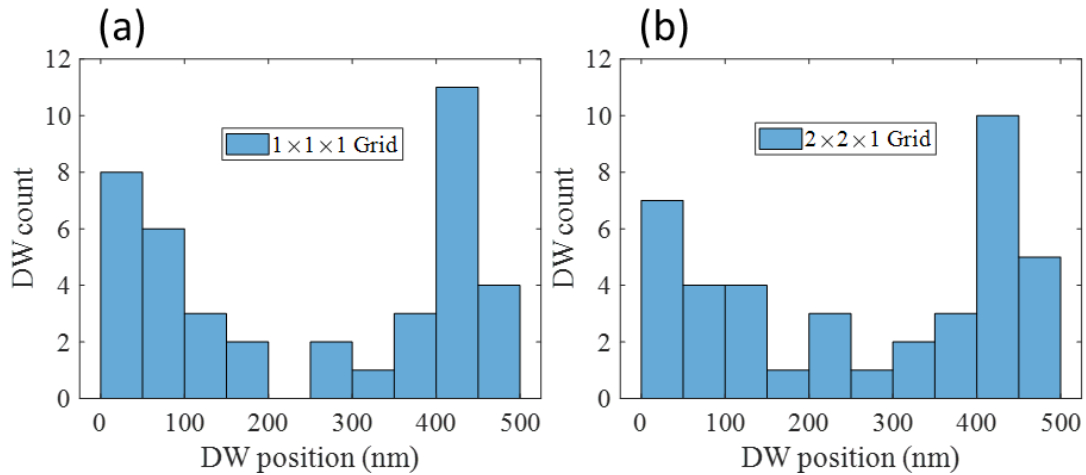


Figure 3-9 Distribution of equilibrium domain wall position in 40 different racetracks for anisotropy coefficient of  $8.0 \times 10^5 J/m^3$  for cell size (a)  $1 \times 1 \times 1 nm^3$  and (b)  $2 \times 2 \times 1 nm^3$ .

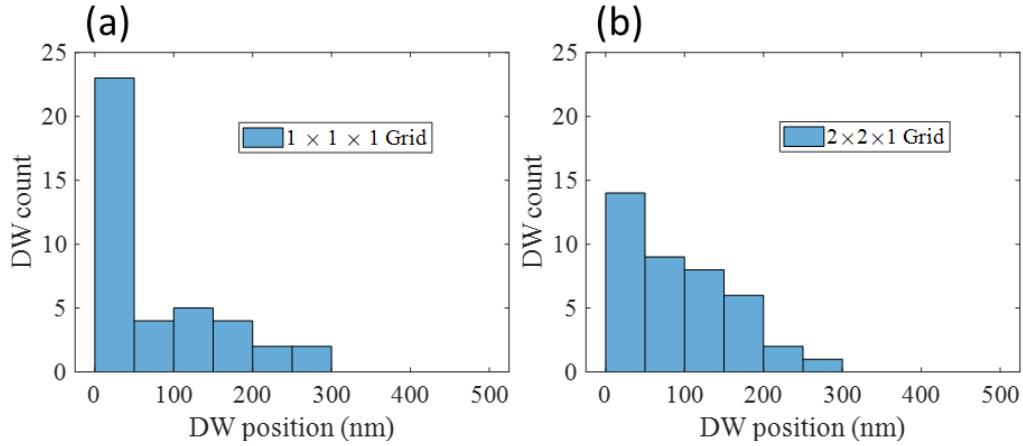


Figure 3-10 Distribution of equilibrium domain wall position in 40 different racetracks for anisotropy coefficient of  $7.0 \times 10^5$   $J/m^3$  for cell size (a)  $1 \times 1 \times 1 \text{ nm}^3$  and (b)  $2 \times 2 \times 1 \text{ nm}^3$ .

### 3.5 Conclusion:

In summary, we have proposed an energy efficient strain-controlled synapse where different synaptic weights have been achieved by applying different values of voltage induced stress in conjunction with a fixed clocking SOT current in chiral DW systems with significant DMI. While a uniform change in stress-induced anisotropy cannot move the DW that is pinned in a trap site, it can influence the potential landscape such that the DW in a low PMA racetrack moves faster than in a high PMA one, when being driven by a fixed SOT current. We have shown that five different mean equilibrium DW positions with five different voltage induced stress values is achievable in a 500 nm long and 50 nm wide racetrack with edge roughness of  $\sim 3$  nm. These suggest the feasibility of a 5-state synapse. A 3-state synapse can be also achieved using three different voltage induced PMA modulation. Recent progress in low precision quantized neural network to achieve near equivalent accuracy to full-precision network makes such a DW based synapse device specifically attractive as a powerful classification tool for edge devices where energy requirement is at a premium.

### References:

- [1] A. Pedram, S. Richardson, M. Horowitz, S. Galal, and S. Kvatinsky, “Dark Memory and Accelerator-Rich System Optimization in the Dark Silicon Era”, IEEE Design & Test, vol. 34, no. 2, pp. 39-50, Apr. 2017. DOI: 10.1109/MDAT.2016.2573586

- [2] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction", in *2011 International Electron Devices Meeting*, pp. 4.4.1-4.4.4, Dec. 2011. DOI: 10.1109/IEDM.2011.6131488
- [3] G. W. Burr, R. M. Shelby, S. Sidler, C. di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi, and H. Hwang, "Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element", *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3498 - 3507, Nov. 2015. DOI: 10.1109/TED.2015.2439635
- [4] I. Boybat, M. L. Gallo, S. R. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian, and E. Eleftheriou, "Neuromorphic computing with multi-memristive synapse", *Nature Communications*, vol. 9, pp. 1-12, Jun. 2018, Art. no. 2514. DOI: 10.1038/s41467-018-04933-y
- [5] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, and H.-S. P. Wong, "An Electronic Synapse Device Based on Metal Oxide Resistive Switching Memory for Neuromorphic Computation", *IEEE Trans. Electron Devices*, vol. 58, no.8, pp. 2729 - 2737, Aug. 2011. DOI: 10.1109/TED.2011.2147791
- [6] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, Q. Zhang, N. Deng, L. Shi, H.-S. P. Wong, and H. Qian, "Face classification using electronic synapses", *Nature Communications*, vol. 8, pp. 1-8, May 2017, Art. no. 15199. DOI: 10.1038/ncomms15199
- [7] A. F. Vincent, J. Larroque, N. Locatelli, N. B. Romdhane, O. Bichler, C. Gamrat, W. S. Zhao, J.-O. Klein, S. G.-Retailleau, and D. Querlioz, "Spin-Transfer Torque Magnetic Memory as a Stochastic Memristive Synapse for Neuromorphic Systems", *IEEE Transactions on Biomedical Circuits and Systems*, vol.9, no. 2, pp. 166 - 174, Apr. 2015. DOI: 10.1109/TBCAS.2015.2414423
- [8] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, and G. W. Burr, "Equivalent-accuracy accelerated neural-network training using analogue memory", *Nature*, vol. 558, pp. 60–67, Jun. 2018. DOI: <https://doi.org/10.1038/s41586-018-0180-5>
- [9] D. Kaushik, U. Singh, U. Sahu, I. Sreedevi, and D. Bhowmik, "Comparing domain wall synapse with other non volatile memory devices for on chip learning in analog hardware neural network", *AIP Advances*, vol. 10, no. 2, pp. 1-7, Feb. 2020, Art. no. 025111. DOI: <https://doi.org/10.1063/1.5128344>
- [10] M. A. Azam, D. Bhattacharya, D. Querlioz, C.A. Ross, and J. Atulasimha, "Voltage control of domain walls in magnetic nanowires for energy-efficient neuromorphic devices", *Nanotechnology*, vol. 31, no. 14, pp. 1-9, Jan. 2020, Art. no. 145201. DOI: <https://doi.org/10.1088/1361-6528/ab6234>

- [11] S. Lequeux, J. Sampaio, V. Cros, K. Yakushiji, A. Fukushima, R. Matsumoto, H. Kubota, S. Yuasa, and J. Grollier, “A magnetic synapse: multilevel spin-torque memristor with perpendicular anisotropy”, *Scientific Reports*, vol. 6, pp. 1-7, Aug. 2016, Art. no. 31510. DOI: 10.1038/srep31510
- [12] D. M. F. Hartmann, R. A. Duine, M. J. Meijer, H. J.M. Swagten, and R. Lavrijsen, “Creep of chiral domain walls”, *Phys. Rev. B*, vol. 100, no. 1, pp. 1-5, Sep. 2019, Art. no. 094417. DOI: 10.1103/PhysRevB.100.094417
- [13] E. Martinez, S. Emori, and G. S. D. Beach, “Current-driven domain wall motion along high perpendicular anisotropy multilayers: The role of the Rashba field, the spin Hall effect, and the Dzyaloshinskii-Moriya interaction”, *Appl. Phys. Lett.*, vol. 103, no. 7, pp. 1-5, Jul. 2013, Art. no. 072406. DOI: <https://doi.org/10.1063/1.4818723>
- [14] A. V. Khvalkovskiy, V. Cros, D. Apalkov, V. Nikitin, M. Krounbi, K. A. Zvezdin, A. Anane, J. Grollier, and A. Fert, “Matching domain-wall configuration and spin-orbit torques for efficient domain-wall motion”, *Phys. Rev. B*, vol. 87, no. 2, pp. 1-5, Jan. 2013, Art. no. 020402(R). DOI: <https://doi.org/10.1103/PhysRevB.87.020402>
- [15] D. Bhowmik, M. E. Nowakowski, L. You, O. Lee, D. Keating, M. Wong, J. Bokor, and S. Salahuddin, “Deterministic Domain Wall Motion Orthogonal To Current Flow Due To Spin Orbit Torque”, vol. 5, pp. 1-10, Jul. 2015, Art. no. 11823. DOI: 10.1038/srep11823
- [16] A. Thiaville, Y. Nakatani, J. Miltat, and N. Vernier, “Domain wall motion by spin-polarized current: a micromagnetic study”, *Journal of Applied Physics*, vol. 95, no. 11, pp. 7049-7051, May 2004. DOI: <https://doi.org/10.1063/1.1667804>
- [17] P. Chureemart, R. F. L. Evans, and R. W. Chantrell, “Dynamics of domain wall driven by spin-transfer torque”, *Phys. Rev. B*, vol. 83, no. 18, pp. 1-8, May 2011, Art. no. 184416. DOI: 10.1088/0953-8984/24/2/024221
- [18] B. Zhang, Y. Xu, W. Zhao, D. Zhu, H. Yang, X. Lin, M. Hehn, G. Malinowski, N. Vernier, D. Ravelosona, and S. Mangin, “Domain-wall motion induced by spin transfer torque delivered by helicity-dependent femtosecond laser”, *Phys. Rev. B*, vol. 99, no. 14, pp. 1-6, Apr. 2019, Art. no. 144402. DOI: <https://doi.org/10.1103/PhysRevB.99.144402>
- [19] N. Lei, T. Devolder, G. Agnus, P. Aubert, L. Daniel, J.-V. Kim, W. Zhao, T. Trypiniotis, R. P. Cowburn, C. Chappert, D. Ravelosona, and P. Lecoeur, “Strain-controlled magnetic domain wall propagation in hybrid piezoelectric/ferromagnetic structures”, *Nature Communications*, vol. 4, pp. 1-7, Jan 2013, Art. no. 1378 DOI: 10.1038/ncomms2386.
- [20] H. T. Chena and A. K. Soh, “Precision electric control of magnetic domain wall motions in a multiferroic bilayer based on strain-mediated magnetoelectric coupling”, *Materials Research Bulletin*, vol. 59, pp. 42-48, Nov. 2014. DOI: <https://doi.org/10.1016/j.materresbull.2014.06.023>

- [21] Q. Wang, J. Z. Hu, C.Y. Liang, A. Sepulveda, and G. Carman, “Voltage-induced strain clocking of nanomagnets with perpendicular magnetic anisotropies”, *Sci. Rep.*, vol. 9, pp. 1-7, Mar. 2019, Art. no. 3639. DOI: <https://doi.org/10.1038/s41598-019-39966-w>
- [22] S. Giordano, Y. Dusch, N. Tiercelin, P. Pernod, and V. Preobrazhensky, “Combined nanomechanical and nanomagnetic analysis of magnetoelectric memories”, *Physical Review B*, vol. 85, no. 15, pp. 1-14, Apr. 2012, Art. no. 155321. DOI: <https://doi.org/10.1103/PhysRevB.85.155321>
- [23] K. Roy, S. Bandyopadhyay, and J. Atulasimha, “Hybrid spintronics and straintronics: A magnetic technology for ultra low energy computing and signal processing”, *Appl. Phys. Lett.*, vol. 99, no. 6, pp. 1-3, Jul. 2011, Art. no. 063108. DOI: <https://doi.org/10.1063/1.3624900>
- [24] J. Atulasimha and S. Bandyopadhyay, “Bennett clocking of nanomagnetic logic using multiferroic single-domain nanomagnets”, *Appl. Phys. Lett.*, vol. 97, no. 17, pp. 1-3, Oct. 2010, Art. no. 173105. DOI: <https://doi.org/10.1063/1.3506690>
- [25] K. Roy, S. Bandyopadhyay, and J. Atulasimha, “Binary switching in a ‘symmetric’ potential landscape”, *Sci. Rep.*, vol. 3, pp. 1-8, Oct. 2013, Art. no. 3038. DOI: [10.1038/srep03038](https://doi.org/10.1038/srep03038)
- [26] N. D’Souza, M. S. Fashami, S. Bandyopadhyay, and J. Atulasimha, “Experimental Clocking of Nanomagnets with Strain for Ultralow Power Boolean Logic”, *Nano Lett.*, vol. 16, no. 2, pp. 1-8, Jan. 2016. DOI: <https://doi.org/10.1021/acs.nanolett.5b04205>
- [27] A. K Biswas, H. Ahmad, J. Atulasimha, and S. Bandyopadhyay, “Experimental Demonstration of Complete 180° Reversal of Magnetization in Isolated Co Nanomagnets on a PMN–PT Substrate with Voltage Generated Strain”, *Nano Lett.* vol. 17, no. 6, pp. 3478–3484, May. 2017. DOI: <https://doi.org/10.1021/acs.nanolett.7b00439>
- [28] A. W. Rushforth, R. R. Robinson, and J Zemen, “Deterministic magnetic domain wall motion induced by pulsed anisotropy energy”, *J. Phys. D: Appl. Phys.*, vol. 53, no. 16, pp. 1-7, Feb. 2020, Art. no. 164001. DOI: <https://doi.org/10.1088/1361-6463/ab6cc7>
- [29] T. Mathurin, S. Giordano, Y. Dusch, N. Tiercelin, P. Pernod, and V. Preobrazhensky, “Stress-mediated magnetoelectric control of ferromagnetic domain wall position in multiferroic heterostructures”, *Appl. Phys. Lett.*, vol. 108, no. 8, pp. 1-5, Feb. 2016, Art. no. 082401. DOI: <https://doi.org/10.1063/1.4942388>
- [30] V. Uhlř, S. Pizzini, N. Rougemaille, J. Novotný, V. Cros, E. Jiménez, G. Faini, L. Heyne, F. Sirotti, C. Tieg, A. Bendounan, F. Maccherozzi, R. Belkhou, J. Grollier, A. Anane, and J. Vogel, “Current-induced motion and pinning of domain walls in spin-valve nanowires studied by XMCD-PEEM”, *Phys. Rev. B*, vol. 81, no. 22, pp. 1-10, Jun. 2010, Art. no. 224418. DOI: <https://doi.org/10.1103/PhysRevB.81.224418>

- [31] X. Jiang, L. Thomas, R. Moriya, M. Hayashi, B. Bergman, C. Rettner, and S. S.P. Parkin, “Enhanced stochasticity of domain wall motion in magnetic racetracks due to dynamic pinning”, *Nature Communications*, vol. 1, pp. 1-5, Jun. 2010, Art. no: 25. DOI: 10.1038/ncomms1024
- [32] J. P. Attan’e, D. Ravelosona, A. Marty, Y. Samson, and C. Chappert, “Thermally Activated Depinning of a Narrow Domain Wall from a Single Defect”, *Phys. Rev. Lett.*, vol. 96, no. 14, pp. 1-4, Apr. 2006, Art. no. 147204. DOI: <https://doi.org/10.1103/PhysRevLett.96.147204>.
- [33] R. Lewis, D. Petit, L. Thevenard, A. V. Jausovec, L. O’Brien, D. E. Read, and R. P. Cowburn, “Magnetic domain wall pinning by a curved conduit”, *Appl. Phys. Lett.*, vol. 95, no. 15, pp. 1-3, Oct. 2009, Art. no. 152505. DOI: <https://doi.org/10.1063/1.3246154>
- [34] D. Petit, A.-V. Jausovec, D. Read, and R. P. Cowburn, “Domain wall pinning and potential landscapes created by constrictions and protrusions in ferromagnetic nanowires”, *J. Appl. Phys.*, vol. 103, no. 11, pp. 1-6, Jun. 2008, Art. no. 114307. DOI: <https://doi.org/10.1063/1.2936981>
- [35] M. Albert, M. Franchin, T. Fischbacher, G. Meier, and H. Fangohr, “Domain wall motion in perpendicular anisotropy nanowires with edge roughness”, *Journal of Physics: Condensed Matter*, vol. 24, no. 2, pp. 1-14, Dec. 2011, Art. no. 024219. DOI: 10.1088/0953-8984/24/2/024219
- [36] S. Dutta, S. A. Siddiqui, J. A. Currivan-Incorvia, C. A. Ross, and M. A. Baldo, “The Spatial Resolution Limit for an Individual Domain Wall in Magnetic Nanowires”, *Nano Lett.*, vol. 17, no. 9, pp. 5869–5874, Aug. 2017. DOI: <https://doi.org/10.1021/acs.nanolett.7b03199>
- [37] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations”, *The Journal of Machine Learning Research* vol. 18, no. 187, pp. 1-30, Apr. 2017.
- [38] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, “Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference”, *arXiv:1712.05877*, Dec. 2017.
- [39] F. Li, B. Zhang, and B. Liu, “Ternary weight networks”, *arXiv:1605.04711*, May 2016.
- [40] V. Joshi, M. Le Gallo, S. Haefeli, I. Boybat, S. R. Nandakumar, C. Piveteau, M. Dazzi, B. Rajendran, A. Sebastian, and E. Eleftheriou, “Accurate deep neural network inference using computational phase-change memory”, *Nature Communications*, vol. 11, pp. 1-13, May 2020, Art. no: 2473. DOI: <https://doi.org/10.1038/s41467-020-16108-9>
- [41] S. Emori, U. Bauer, S.-M. Ahn, E. Martinez, and G. S. D. Beach, “Current-driven dynamics of chiral ferromagnetic domain walls”, *Nat. Mater.*, vol. 12, pp. 611–616, Jun. 2013. DOI:10.1038/NMAT3675
- [42] J. A. Currivan, S. Siddiqui, S. Ahn, L. Tryputen, G. S. D. Beach, Marc A. Baldo, Caroline A. Ross, “Polymethyl methacrylate/hydrogen silsesquioxane bilayer resist electron beam lithography process



- for etching 25 nm wide magnetic wires”, J. Vac. Sci. Technol. B, vol. 32, no. 2, pp. 1-5, Feb. 2014, Art. no. 021601. doi: <https://doi.org/10.1116/1.4867753>
- [43] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. V. Waeyenberge, “The design and verification of MuMax3”, AIP Advances, vol. 4, no. 10, pp. 1-22, Oct. 2014, Art. no. 107133. DOI: <https://doi.org/10.1063/1.4899186>
- [44] L. Liu, R. A. Buhrman, and D. C. Ralph, “Review and Analysis of Measurements of the Spin Hall Effect in Platinum”, arXiv:1111.3702, Mar. 2012.
- [45] C. J. G.-Cervera, Z. Gimbutas, and Weinan E, “Accurate numerical methods for micromagnetics simulations with general geometries”, Journal of Computational Physics, vol. 184, no. 1, pp. 37-52, Jan. 2003. DOI: [https://doi.org/10.1016/S0021-9991\(02\)00014-1](https://doi.org/10.1016/S0021-9991(02)00014-1)
- [46] G. J. Parker, C. Cerjan, and D. W. Hewett, “Embedded curve boundary method for micromagnetic simulations”, Journal of Magnetism and Magnetic Materials, vol. 214, no. 1-2, pp. 130-138, May. 2000. DOI: [https://doi.org/10.1016/S0304-8853\(00\)00043-3](https://doi.org/10.1016/S0304-8853(00)00043-3)
- [47] M. J. Donahue, and R. D. McMichael, “Micromagnetics on Curved Geometries Using Rectangular Cells: Error Correction and Analysis”, IEEE Transactions on Magnetics, vol. 43, no. 6, pp. 2878 - 2880, Jun. 2007. DOI: 10.1109/TMAG.2007.892865
- [48] J. Cui, J. L. Hockel, P. K. Nordeen, D. M. Pisani, C.-Y. Liang, G. P. Carman, and C. S. Lynch, “A method to control magnetism in individual strain-mediated magnetoelectric islands”, Appl. Phys. Lett., vol. 103, no. 23, pp. 1-5, Dec. 2013, Art. no. 232905. DOI: <https://doi.org/10.1063/1.4838216>
- [49] G. Bertotti, I. D. Mayergoyz, and C. Serpico, “Stochastic magnetization dynamics,” in *Nonlinear Magnetization Dynamics in Nanosystems*, Amsterdam, The Netherlands: Elsevier, 2009, ch 10, pp-271-345.
- [50] C. Bilzera, T. Devolder, J.-V. Kim, G. Counil, and C. Chappert, “Study of the dynamic magnetic properties of soft CoFeB films”, J. Appl. Phys., vol. 100, no. 5, pp. 1-4, Sep. 2006, Art. no. 053903. DOI: <https://doi.org/10.1063/1.2337165>
- [51] M. Belméguenai, M. S. Gabor, Y. Roussigné, A. Stashkevich, S. M. Chérif, F. Zighem, and C. Tiusan, “Brillouin light scattering investigation of the thickness dependence of Dzyaloshinskii-Moriya interaction in Co<sub>0.5</sub>Fe<sub>0.5</sub> ultrathin films”, Physical Review B, vol. 93, no. 17, pp. 1-8, May 2016, Art. no. 174407. DOI: <https://doi.org/10.1103/PhysRevB.93.174407>
- [52] D. Hunter, W. Osborn, K. Wang, N. Kazantseva, J. H.-Simpers, R. Suchoski, R. Takahashi, M. L. Young, A. Mehta, L. A. Bendersky, S. E. Lofland, M. Wuttig, and I. Takeuchi, “Giant magnetostriction in annealed Co<sub>1-x</sub>Fex thin-films”, Nature Communications, vol. 2, pp. 1-7, Nov. 2011, Art. no: 518. DOI: 10.1038/ncomms1529

- [53] S. Dutta, S. A. Siddiqui, J. A. Currivan-Incorvia, C. A. Ross, and M. A. Baldo, Micromagnetic modeling of domain wall motion in sub-100-nm-wide wires with individual and periodic edge defects, *AIP Advances* 5, 127206 (2015), doi:<https://doi.org/10.1063/1.4937557>
- [54] T. Suzukia, S. Fukami, N. Ohshima, K. Nagahara, and N. Ishiwata, Analysis of current-driven domain wall motion from pinning sites in nanostrips with perpendicular magnetic anisotropy, *Journal of Applied Physics* 103, 113913 (2008), doi:<https://doi.org/10.1063/1.2938843>
- [55] Takashi Komine, Hiroshi Murakami, Takahiro Nagayama, Ryuji Sugita, Influence of Notch Shape and Size on Current-Driven Domain Wall Motions in a Magnetic Nanowire, *IEEE Transactions on Magnetics*, Vol. 44, no.11, pp. 2516 - 2518, Nov. 2008, DOI: 10.1109/TMAG.2008.2002614
- [56] J. Torrejon, E. Martinez, and M. Hayashi, “Tunable inertia of chiral magnetic domain walls”, *Nature Communications*, vol. 7, pp. 1-7, Nov. 2016, Art. no: 13533. DOI: 10.1038/ncomms13533
- [57] E. Martinez, S. Emori, N. Perez, L. Torres, and G. S. D. Beach, “Current-driven dynamics of Dzyaloshinskii domain walls in the presence of in-plane fields: Full micromagnetic and one-dimensional analysis”, *Journal of Applied Physics*, vol. 115, no. 21, pp. 1-14, Art. no. 213909, May 2014.
- [58] A. Thiaville, S. Rohart, É. Jué, V. Cros, and A. Fert, “Dynamics of Dzyaloshinskii domain walls in ultrathin magnetic films”, *EPL (Europhysics Letters)*, vol. 100, no. 5, pp. 1-6, Dec. 2012, Art. no. 57002. DOI: 10.1209/0295-5075/100/57002
- [59] A. Mohanty, X. Du, P.-Yu Chen, J. Seo, S. Yu, Y. Cao, “Random sparse adaptation for accurate inference with inaccurate multi-level RRAM arrays”, in *2017 IEEE International Electron Devices Meeting (IEDM)*, Dec. 2017, DOI: 10.1109/IEDM.2017.8268339
- [60] M. Hu, H. Li, Y. Chen, Q. Wu, G. S. Rose. “BSB training scheme implementation on memristor-based circuit”, in *2013 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, Apr. 2013. DOI: 10.1109/CISDA.2013.6595431
- [61] B. Liu, H. Li, Y. Chen, X. Li, T. Huang, Q. Wu, and M. Barnell, “Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems”, in *2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Jan. 2015. DOI: 10.1109/ICCAD.2014.7001330
- [62] T. Hirtzlin, M. Bocquet, B. Penkovsky, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, and D. Querlioz, “Digital Biologically Plausible Implementation of Binarized Neural Networks With Differential Hafnium Oxide Resistive Memory Arrays”, *Frontiers in Neuroscience*, Vol. 13, pp.1-14, Jan. 2020, Art. no. 1383, doi: <https://doi.org/10.3389/fnins.2019.01383>
- [63] H.-S. Philip Wong, and S. Salahuddin, Memory leads the way to better computing, *Nature Nanotechnology*, vol 10, pp. 191-194, Mar. 2015. doi: <https://doi.org/10.1038/nnano.2015.29>

## Chapter 4: Energy Efficient Learning with Low Resolution Stochastic Domain Wall Synapse for Deep Neural Networks

In Chapter 3, we demonstrated voltage-controlled domain wall synapses using racetrack memory devices. The synaptic states that can be achieved with small-footprint racetracks are limited and exhibit stochastic behavior. In this chapter we demonstrate extremely low resolution quantized (nominally 5-state) synapses with large stochastic variations in synaptic weight can be energy efficient and achieve reasonably high testing accuracies compared to deep neural networks (DNNs) of similar sizes using floating precision synaptic weights. We propose both in-situ and ex-situ training algorithms, based on modification of the algorithm proposed by Hubara et al. [1] which works well with quantization of synaptic weights, and train several 5-layer DNNs on MNIST dataset using 2-, 3- and 5-state DW devices as a synapse. For in-situ training, a separate high precision memory unit preserves and accumulates the weight gradients which prevents the accuracy loss due to weight quantization. For ex-situ training, a precursor DNN is first trained based on weight quantization and characterized DW device model. Moreover, a noise tolerance margin is included in both of the training methods to account for the intrinsic device noise. The highest inference accuracies we obtain after in-situ and ex-situ training are  $\sim 96.67\%$  and  $\sim 96.63\%$  which is very close to the baseline accuracy of  $\sim 97.1\%$  obtained from a similar topology DNN having floating precision weights with no stochasticity. Large inter-state interval due to quantized weights and noise tolerance margin enables in-situ training with significantly lower number of programming attempts. Our proposed approach demonstrates a possibility of at least *two orders of magnitude* energy savings compared to the floating-point approach implemented in CMOS. This approach is specifically attractive for low power intelligent edge devices where the ex-situ learning can be utilized for energy efficient non-adaptive tasks and the in-situ learning can provide the opportunity to adapt and learn in a dynamically evolving environment.

Deep neural networks (DNNs) have proven to be successful in image recognition and other big data driven classification tasks. However, implementing a DNN with traditional von-Neumann computing is time consuming [2] as it requires shuttling a large number of synaptic weight data stored in the memory to the processing unit to perform matrix-vector multiplication during the forward propagation and backward propagation stages. Moreover, shuttling data between the computational unit and memory unit is energy intensive [3], which hinders the implementation of such DNNs in edge devices where energy is at a premium [4-5].

In-memory computing has been widely explored to reduce the physical separation between computation and memory units. In-memory computing is a non-von-Neumann computing paradigm where the computational memory units are arranged in a way that certain computational tasks take place in the memory itself [6-7]. Matrix vector multiplication, the most computationally intensive part of a DNN [8], has been demonstrated with in-memory computing [9-10]. When the computational memory units are connected in a crossbar and programmed to provide conductances equivalent to the DNN weights [11-12], the matrix-vector multiplication operation can be implemented in single time step [6,8] and with minimal data movement. Computational memory such as phase change random access memory (PCM) [13-14], resistive random-access memory (RRAM) [15-16], arranged in a crossbar array have been shown to classify handwritten digits [11,17] and recognize human faces [18]. However, these analog memory devices have stochastic and non-linear responses and provide limited resolution for synaptic weights. To achieve higher classification accuracy, these issues should be addressed with appropriate training algorithms.

Recently spintronic memory devices have been widely explored for in-memory DNN implementation because of their non-volatility, high endurance, high speed of access, high scalability and compatibility with CMOS technology [2,19-24]. Among these spintronic devices, domain wall (DW) based computational memory [19-20] is promising and these devices can be programmed with a low energy budget [25]. However, similar to other analog devices, DW devices have limitations such as their stochastic behavior [26-29] and low resolution due to the relatively small on/off ratio of magnetic tunnel junctions (MTJs) which are 7:1, at best, at room temperature [30].

With recent advances in computing, researchers have shown fast and energy efficient implementation of DNNs with low resolution synaptic weights [1,31-34] where the weights' values can only be binary (1-bit or 2-state) [31]. However, for updating weights, gradients are calculated in full precision to achieve high accuracy [1]. This idea of keeping full precision gradient information for training a network with limited precision synaptic weights can be useful for a DNN that is built from energy efficient DWs or other analog low-resolution devices.

Apart from the low resolution, stochasticity and non-linear response of the analog devices should be addressed during training to achieve higher classification accuracy [11]. To address the stochasticity of the analog devices, both online (in-situ) and offline (ex-situ) training of the DNN are proposed. For online training, multiple devices per synapse have been proposed with [35] or without 'periodic carry' [36] to address device variability and noise. In another work, a 3T1C module (consists of 3 CMOS transistors and 1 capacitor) is used in conjunction with the stochastic PCM device to accumulate small linear updates and then periodically transfer them to the non-volatile PCM [37]. However, with online training, using the techniques mentioned above during the weight update stage, each of the synaptic weights in the crossbar

array are updated. This has great implications for the endurance of the devices as well as the energy consumed in training the device. Recently, a mixed precision framework [38-39] has been proposed where large computational loads, such as the weighted sum operation (matrix-vector multiplication) along with the conductance updates, are performed in a low precision computational memory unit and the weight updates are accumulated in a high precision unit. Using this framework, a large variety of DNNs have been shown to achieve high classification accuracy with significantly smaller number of weight updates [38].

In contrast to the online training, for offline training the DNN is trained in software and the actual devices are programmed based on the learned weights from software. In this case, hardware nonidealities are characterized first and then included in the training process. To address stochasticity of the devices, Gaussian noise injection for the DNN weights has been proposed [40] and has shown excellent accuracy. Random gaussian noise is also added to the ternary weights (3-state weight) of a DNN [41]. Variation aware offline training algorithm is reported in [42-43] where the variation in device conductances and device defects are first characterized and then incorporated during the training of the DNN in software. In another case involving a deep convolutional neural network, the optimal weights for convolution layers and fully connected layers are learned via offline training before the fully connected layers' weights being updated by online training [44].

Although a significant amount of research has focused on addressing the device variability and non-linearity, a largely unexplored area is quantized (low resolution) learning with these non-volatile devices. Even a high-resolution device (or a low granularity device) can behave as a low-resolution device, when device variability is taken into account. The limited numbers of synaptic states provided by low-resolution devices can strongly impact accuracy. At the same time, in neuromorphic computing applications, these devices offer advantages. An inherent benefit of the low-resolution devices can be their large inter-state interval. It provides an opportunity to train a neural network with significantly lower number of the weight updates with a standard learning rate. However, accuracy metrics need to be acceptable while using these devices for in-situ learning.

In this study, we demonstrate that such low resolution and stochastic non-volatile devices can perform highly accurate in-situ classification tasks while taking advantage of significantly lower number of device updates (energy efficiency). In our proposed algorithm, weight gradients are accumulated in high precision (digital domain) before quantizing this information to program the low-resolution devices in analog domain. Using rigorous device model and simulations, we show that such in-situ training can achieve comparable accuracy to that of a 32-bit precision synapse. This demonstrates that there is a possible low-resolution weight space, which can provide an optimal solution to the classification problem with highly energy-efficient non-volatile devices. In addition, we have shown ex-situ training strategies to achieve high

classification accuracy for DNN implemented with these highly stochastic (non-Gaussian) and extremely low resolution (nominally 2-state, 3-state, and 5-state for synaptic weights) analog DW based computational memory devices.

The rest of the chapter is organized as follows. In the methods section, we detail the architecture of the DW device that can work as a synapse in the DNN and discuss the in-situ and ex-situ learning algorithms of such DW device based DNN. For both of the learning algorithms we adapt quantized neural network learning algorithm [1,31] with several modifications including the weight deviation tolerance from target weight to account for the programming noise intrinsic to such stochastic DW devices. For ex-situ training, we also incorporate the statistical distribution of the DW device conductance during the training, which helps to achieve higher test accuracy. This is followed by the results and discussion section, and then a conclusion.

#### **4.1. Methods:**

##### **DW based Nano-Synapse and Micromagnetic Modeling for Device Stochasticity:**

We model our synapse on a magnetic DW based nanodevice, which is non-volatile in nature. Once the memory state (here the synaptic weight) is written, the information is retained for a long time. For the nano-synapse device, we simulated a thin ferromagnetic racetrack having a dimension of  $600 \text{ nm} \times 60 \text{ nm} \times 1 \text{ nm}$  with a DW initialized and stabilized in a notch at one end. In addition, we assume several engineered notches at regular intervals along the racetrack. The racetrack dimensions and notch intervals are shown in Fig. 4-1a. Moreover, we consider edge irregularities (rms roughness of  $\sim 2 \text{ nm}$ ) in the racetrack to mimic the effect of lithographic imperfections by randomly removing or adding some finite difference cells from the edges [45-46]. We assume the racetrack is on top of a heavy metal layer that is patterned on top of the piezoelectric layer (see Fig. 4-1b). An insulator (MgO layer) and a reference ferromagnetic layer (one could also add a synthetic antiferromagnet (SAF) layer to cancel dipole coupling from this fixed layer) are stacked on top of the racetrack, these two layers combined with the racetrack ferromagnetic layer (free layer) forms a magnetic tunnel junction (MTJ) (see Fig. 4-1b), which facilitates the readout of the device. With this configuration, a combination of fixed amplitude and fixed time current pulse “or clocking signal” injected in the heavy metal layer and a varying amplitude “control” voltage pulse applied across the piezoelectric translates the domain wall into different distances along the racetrack. Different positions of the DW lead to different conductances of the MTJ thus forming a voltage programmable non-volatile synapse.

### Magnetization Dynamics:

The magnetization dynamics of the domain wall (DW) in the racetrack ferromagnetic layer which governs the conductance (synaptic state) evolution of the nano-synapse is simulated in MUMAX3 [47] using the Landau–Lifshitz–Gilbert–Slonczewski equation. The simulation parameters are listed in Table 4-1. The simulation details can be found in [29] and in section 1.2.7.

Table 4-1: Simulation parameter

Parameters	Values
DMI constant (D)	$0.0006 \text{ Jm}^{-2}$
Gilbert damping ( $\psi$ )	0.03
Saturation magnetization ( $M_s$ )	$10^6 \text{ Am}^{-1}$
Exchange constant ( $A_{ex}$ )	$2 \times 10^{-11} \text{ Jm}^{-1}$
Saturation magnetostriction ( $\lambda_s$ )	250 ppm
Perpendicular Magnetic Anisotropy ( $K_u$ )	$7.5 \times 10^5 \text{ Jm}^{-3}$

### Mapping Domain Wall Position to Conductivity:

The distribution of equilibrium DW positions for five different programming voltages, represented by different perpendicular magnetic anisotropy (PMA) coefficient,  $K_u$ , in addition to fixed amplitude and fixed time spin orbit torque (SOT) generating current pulse ( $35 \times 10^{10} \text{ A/m}^2$  applied for 1 ns) in the presence of room temperature thermal noise are shown in Fig. 4-1c. The mean equilibrium DW positions vary for different  $K_u$ , which implies that different programming voltages can be chosen for different synaptic states. For example, one can select five, three, or two different programming voltages to implement a 5-state, 3-state or 2-state synapse. The number of states that can be obtained from the device is limited due to the fact that with the presence of Dzyaloshinskii–Moriya interaction (DMI), SOT current can cause DW tilting long after the current stimulus is withdrawn [48]. Thus, in the presence of room temperature thermal noise and structural irregularities (edge roughness) DW motion becomes significantly stochastic. As a result, an average variance of  $\sim 90 \text{ nm}$  can be seen for different DW mean positions in our modelled racetrack. Considering all of these factors contributing to the stochasticity, we choose a maximum 5-state for our modelled device as higher number of states can cause larger overlaps between the states, which is detrimental for DNN performance. While it is possible to increase the number of states by increasing the racetrack dimension (increase area footprint) and/or increase the notch depth (increases energy as operating current increases), the main contributions of our study is to show that we can use extremely low precision

(5-state, 3-state, etc.) non-volatile synapses for in-situ (and ex-situ) DNN training and achieve classification accuracy that are comparable to that of full precision (32-bit) DNNs.

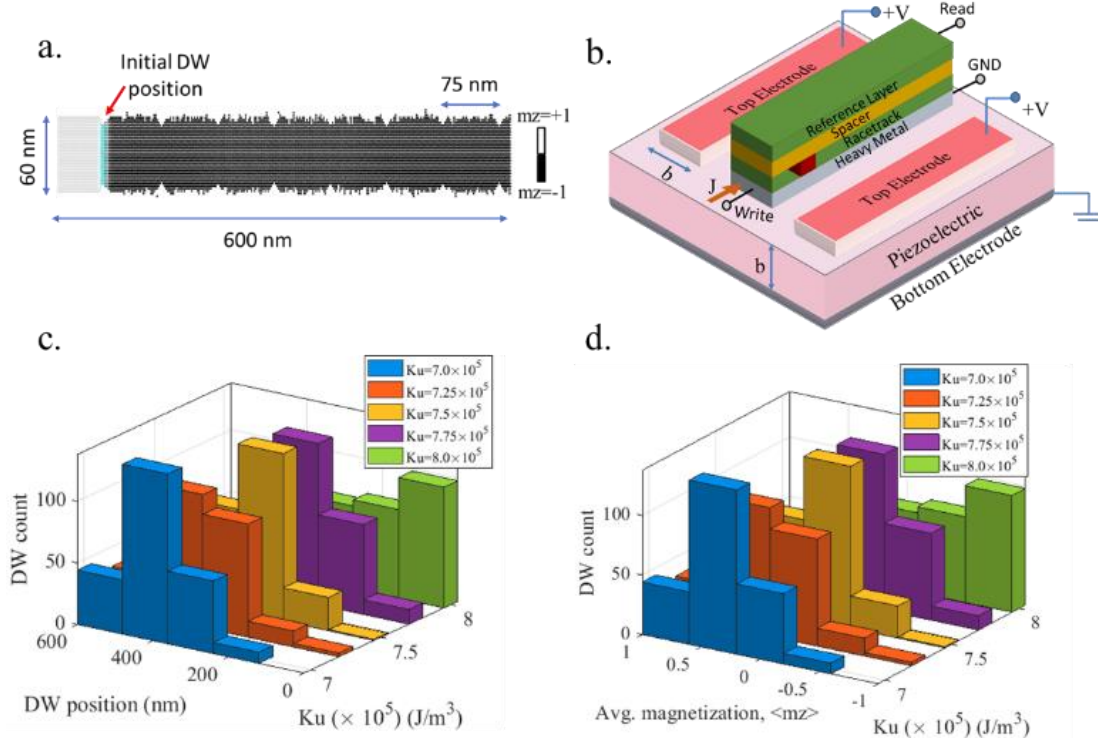


Figure 4-1 a. Micromagnetic configuration of a ~2 nm rms rough edge racetrack with perpendicular magnetic anisotropy (PMA). Engineered notches are placed regularly at 75 nm intervals. A DW is initialized at a notch 60 nm from the left of the racetrack. b. DW based nano-synapse device: racetrack ferromagnet/insulator/reference ferromagnet (MTJ) on top of a heavy metal layer on a piezoelectric substrate. A fixed amplitude current pulse,  $J$  in the heavy metal layer along with different amplitude voltage pulse,  $V$  across the piezoelectric changes the perpendicular anisotropy (PMA or  $K_u$  constant) of the racetrack and translates the DW (shown in red rectangle) into different longitudinal positions along the racetrack. c. Distribution of equilibrium DW positions in the racetrack (shown in Fig. 4-1a) at room temperature  $T=300K$  for a fixed SOT generating current pulse of  $J = 35 \times 10^{10} A/m^2$  applied for 1 ns and five different PMA coefficients,  $K_u$  (corresponds to five different programming voltages). Different mean positions for different  $K_u$  implies that 5-state, 3-state or 2-state stochastic synapses can be implemented by choosing 5,3 or 2 different programming voltages. d. distribution of average perpendicular magnetization,  $\langle m_z \rangle$  (which is equivalent to DNN weights according to Eq. 4) derived directly from DW positions.

Equilibrium DW positions shown in Fig. 4-1c can be linearly mapped to a conductance value by means of the following equations:

$$G_{synapse} = \frac{G_{max} + G_{min}}{2} + \frac{G_{max} - G_{min}}{2} \langle m_z \rangle \quad (1)$$



where,  $\langle m_z \rangle$  is the average magnetization moment of ferromagnetic racetrack along z-direction and the reference ferromagnetic layer magnetization is assumed to point upward, +z-direction. The distribution of  $\langle m_z \rangle$  is shown in Fig. 4-1d, which can be derived directly from DW position.  $G_{max}$  and  $G_{min}$  are the maximum and minimum conductance of the synaptic device which occur when the racetrack and reference layer magnetizations are parallel and anti-parallel respectively.

## 4.2. Learning of Fully Connected DNN with Domain Wall Nano-synapse

### Crossbar with DW Devices:

We assume a crossbar architecture for the DW devices (Fig. 4-2b) that implements a fully connected DNN (Fig. 4-2a). The task of the DNN studied here is classification of handwritten digits from MNIST database [49]. The network is trained with the MNIST training images each having  $28 \times 28$  pixels or a total of 784 pixels with intensity values ranging from 0-255. The pixel intensities of the image converted to binary values acts as input to the DNN and the corresponding linearly mapped voltages are supplied to the crossbar using peripheral circuits. We have considered 3 hidden layers for the DNN and the numbers of neurons for input layer, hidden layers and output layer are chosen to be 784-392-196-98-10. The reason for the choice is discussed in the results section. The conductance of the devices can be scaled linearly to represent the weights  $W_{ij}$  of the DNN [19].

$$G_{ij}^{synapse} = \frac{G_{max} + G_{min}}{2} + \frac{(G_{max} - G_{min})W_{ij}}{2 W_{max}} \quad (2)$$

Here,  $W_{max}$  is the maximum absolute value for the weights of the DNN. DNN weights,  $W_{ij}$  can be both positive and negative; however, the DW devices can only provide positive conductance values. To address the issue, one can add a parallel conductance,  $G_p = \frac{G_{max} + G_{min}}{2}$  to each of the cross-points in the crossbar and feed this parallel conductance with a voltage that is of opposite polarity to the voltage applied to the DW device [19]. This parallel reference conductance,  $G_p$  can be achieved using a similar dimension DW device as the nano-synapse with the DW being driven and pinned at the center of the racetrack (in this case,  $\langle m_z \rangle \sim 0$  in Eq. 1). An engineered notch placed at the center of the racetrack can provide further pinning strength to the DW in addition to the demagnetization potential minima, which acts near the center of the racetrack [29]. Stochasticity that could arise in programming the parallel conductances is not considered in our study.

Two separate rows supplied with opposite polarity voltages connect the synaptic devices and parallel conductances to a single column of the crossbar (bit line (BL)) as shown in Fig. 4-2b. The additional read

line (RL), write word line (WWL) and source line (SL) shown in Fig. 4-2b are required to enable read and write operation. The WWL for the parallel conductances are not shown for the sake of simplicity. To read the column sum, RL is activated and WWL is deactivated, whereas a specific device can be read by activating RL and supplying read voltage to the corresponding input line. To program a device, WWL is activated and RL is deactivated and SL and BL are made high or low depending on the direction of the current in addition to a write voltage being supplied across the piezoelectric thickness. The effective conductance,  $G_{ij}$ , at each cross-point would be,

$$G_{ij} = \frac{(G_{max} - G_{min})W_{ij}}{2 W_{max}} \quad (3)$$

Combining Eq. 1 and 2 and considering  $W_{max} = 1$ , one can get,

$$W_{ij} = \langle m_z \rangle \quad (4)$$

From the above equation it is clear that if we train the DNN shown in Fig. 4-2a with weights (both positive and negative) that are derived from micro-magnetics (see Fig. 4-1d) for different programming conditions, we are effectively implementing a hardware DNN with DW devices shown in Fig. 4-2b given that the peripheral circuitry is designed to provide the appropriate scaling.

During backward propagation, linearly scaled voltages corresponding to the output layer,  $L$ , error signal,  $\delta_i^L = y_i^L - d_i^L$ , are supplied to the crossbar, where  $y_i^L$  and  $d_i^L$  are the predicted and desired outcomes of the output layer's neuron  $i$ .

### **Backpropagation and Learning Algorithm:**

For the training of the DNN, we update the weights by calculating the gradient of a cost function with respect to the weights. We considered mean square error,  $C = \frac{1}{2} \sum (y_i^L - d_i^L)^2$  as our cost function where the gradient of the cost function with respect to the output of the output layer neuron  $i$  is expressed as ,  $\delta_i^L = (y_i^L - d_i^L)$  (we also call it error ). Once the output layer's errors are determined, the preceding layer's errors can be calculated using the backpropagation equation,  $\delta_i^l = W_{ij} \delta_j^{l+1}$ , which is different from the backpropagation equation,  $\delta_i^l = W_{ij} \delta_j^{l+1} f'_{l+1}$  reported in Ref [50] where  $f'_{l+1}$  is the gradient of the activation function of layer  $l + 1$  neuron (we use sigmoid function as activation in our simulation). In other words, we do not backpropagate the gradients of activation function as it does not achieve high testing accuracy with quantized weights. Finally, the derivative of the cost function with respect to the weights is calculated, which determines the weight update signal,  $\Delta W_{ij}$  for the weights connected between layer  $l$  neuron  $i$  and layer  $l + 1$  neuron  $j$ ,

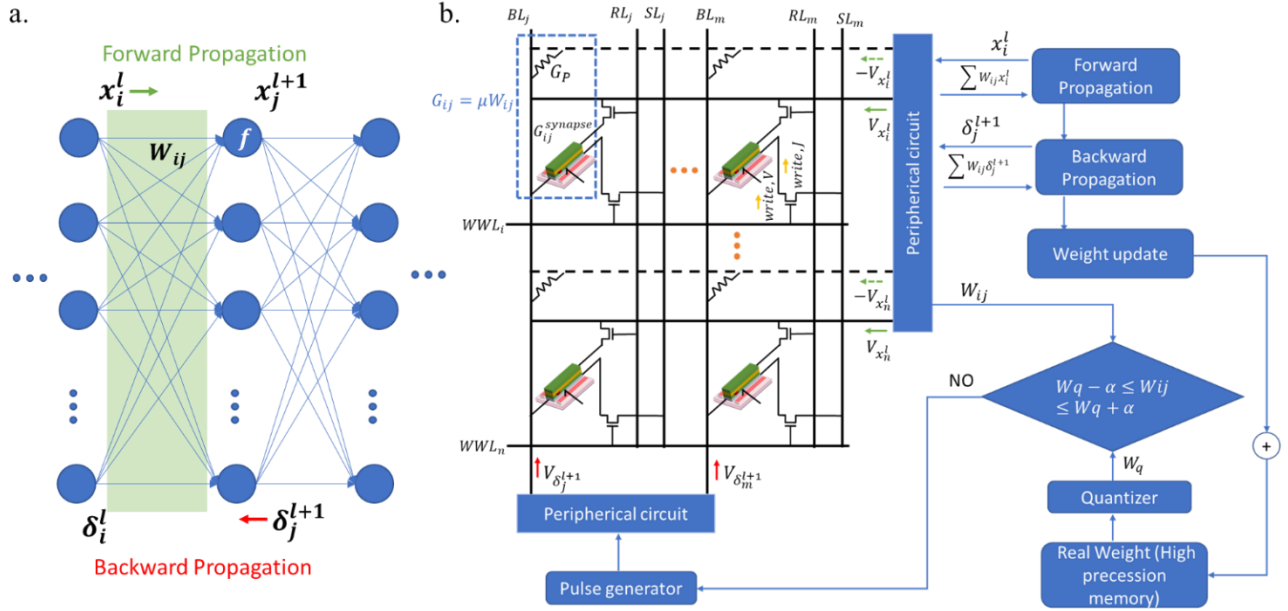


Figure 4-2 a. Architecture of a fully connected deep neural network (DNN). Any neuron  $i$  in layer  $l$  is connected to neuron  $j$  in layer  $l + 1$  with synaptic weight  $W_{ij}$ . At forward propagation, inputs to neuron  $j$  are summed and passed through an activation function  $f$  to generate its output,  $x_j^{l+1}$ . At backward propagation, errors of layer  $l + 1$  neurons are back propagated to calculate the error  $\delta_i^l$  of neuron  $i$  in layer  $l$ . b. Implementation of the DNN in crossbar with DW devices. The peripheral circuit and the crossbar shown here implements DNN functionalities of only one layer (“ $l$ ”) and the next layer (“ $l + 1$ ”) and the number of rows in the crossbar are determined by the number of neurons in layer,  $l$  and the number of columns by the number of neurons in layer,  $l + 1$  (shaded region in Fig. 4-2a). At each cross point of the crossbar there is a DW device with conductance  $G_{ij}^{synapse}$  and a parallel conductance  $G_P$ . The effective conductance at each cross point is equivalent to the DNN weights  $W_{ij}$  such that  $G_{ij} = \mu W_{ij}$ . Inputs and errors of neurons are scaled to voltages before feeding them into the crossbar. The flow of the training algorithm is shown at the right-hand side of the crossbar. For each of the DW devices there is a corresponding high precision weight (real weight) that is stored in a separate high precision memory unit. These high precision weights are updated after a forward and backward pass before passing it through a quantizer (i.e., 2, 3 or 5-level quantization, depending on the number of states of the device). The DW device conductances,  $G_{ij}$  (or the corresponding device weights,  $W_{ij} = \frac{G_{ij}}{\mu}$ ) are updated when they fall outside the prescribed range of the target quantized weights,  $W_q$ . Figure idea adopted from [38].

$$\Delta W_{ij} = \eta x_i^l \delta_j^{l+1} f'_{l+1} \quad (5)$$

Here,  $\eta$  denotes the learning rate. For our learning algorithm we propose to store the updated weights in a separate high precision memory unit. That way, the gradients with respect to the weights can be calculated accurately [1]. We note that these high precision weights are different from the actual synaptic weights (or equivalent conductances) provided by the DW device that are quantized and of low precision. However, we use these high precision weights to update the DW device weights (conductances). As we apply

stochastic gradient decent for optimization, these high precision weights are updated at each forward pass with an input image.

As DW devices can only provide limited resolution in their synaptic weights we adopt weight quantization in our training algorithm. For that, the high precision weights are quantized at each forward pass during the training. For weight quantization, we use the following sets of functions in the manner of Ref [51]:

$$\text{clip}(m, a, b) = \min(\max(m, a), b)$$

$$\Delta = \frac{b - a}{n - 1} \tag{6}$$

$$q = \left[ \text{round} \left( \frac{\text{clip}(m, a, b) - a}{\Delta} \right) \right] \times \Delta + a$$

where,  $q$  is the quantized value of the real valued number  $m$ ,  $[a; b]$  is the quantization range and  $n$  is the level of quantization. The level of quantization depends on the number of distinct states (without significant overlap) the device that is used to implement the DNN crossbar arrays is capable of providing. After quantization, a programming pulse is generated to update the DW device weights to the quantized value, a target that is similar to the quantized neural network learning algorithm [1]. We note that, the cost gradients with respect to the prior quantization quantities is zero, so to backpropagate gradients through weight quantization we apply “straight through estimator” approach similar to that used in [1]. Typically, two types of training are possible for a DNN implemented with DW nano-synapse device: in-situ and ex-situ. In in-situ training the DNN is trained and tested in hardware. In contrast, in ex-situ training, a precursor DNN is first trained in software and then the DW devices are programmed to provide the equivalent learned weights prior to testing.

#### 1) In-situ Training:

Here, we describe in detail the step-by-step in-situ training algorithm presented in Algorithm 4-1 and shown in Fig. 4-2b. This Algorithm 1 is based on the modification of quantized neural algorithms presented in Ref [1]. For each DW device in crossbar arrays there is a corresponding high precision weight that is stored in a separate digital memory unit to accumulate the weight gradients in full precision. Initially, these high (full) precision weights are chosen at random from a gaussian distribution. After each forward and backward pass in the analog crossbar array, these weights are updated according to Eq. 5. Then, these weights are clipped and quantized so that they lie between -1 to 1. After that, a programming pulse is sent to the DW device to update its synaptic weight value to the quantized value. For example, in 5-level quantization (5-state for the synaptic device) the quantized weights can be of any value from the set  $W_q \in$

(-1, -0.5, 0, 0.5, 1). Five different programming voltages can be applied to the device, which results in  $K_u = 8, 7.75, 7.5, 7.25$  and  $7.0 (\times 10^5) J/m^3$  to achieve five different quantized weights of -1, -0.5, 0, 0.5 and -1 respectively as seen from Fig. 4-1d (DW device weights,  $W_{ij} = \langle m_z \rangle$  according to Eq. 4). Because of the significant spread that exists in the DW device weights (or the  $\langle m_z \rangle$  distribution) due to the stochastic nature of the device we introduce a noise tolerance hyperparameter called alpha,  $\alpha$  (real valued) during training, as after applying a programming pulse (fixed current + control voltage) the device weights can be of a value other than the desired quantized weight. For instance, if we want to program a DW device to a quantized weight of  $W_q$  then we would allow any values for the device weights that satisfy the condition,  $W_q - \alpha \leq W_{ij} \leq W_q + \alpha$ . Therefore, at each iteration following quantization, we read the states of the DW device (costs read energy but that is typically much lower than write energy) and only if it falls outside the noise tolerance margin, a programming pulse is sent to the device to write the corresponding quantized weight. However, due to the large inter-state interval in quantized learning, a quantized weight does not change at each forward pass (the backpropagated errors update the weights slowly due to low learning rate). Instead, it typically changes only after several passes. Therefore, the noise tolerance condition need not to be satisfied strictly at each iteration. Furthermore, if a DW device weight is programmed outside the tolerance margin, it is not rectified in the current iteration, as it has a chance to satisfy the window in the next several iterations. This relaxation over noise tolerance condition speeds up the training process without losing accuracy. Again, the DW device, which already satisfies the tolerance margin, need not be programmed for next several iterations due to same reason of the quantized weights not being updated frequently. Introduction of noise tolerance hyperparameter,  $\alpha$ , is critical during the training of this stochastic device based DNN. Without  $\alpha$ , the DW device needs to be programmed a significantly large number of times to achieve a particular quantized weight. On the other hand, a high value of  $\alpha$  allows more imprecise weight update or higher variation of the DW device weights from the target values, which will degrade the accuracy. Thus, a proper balance needs to be found for selecting the value of  $\alpha$  so that it not only ensures high classification accuracy but also low programming energy. Once the DNN is trained, the learned DW device's weights (or conductances) remains the same during testing, as these devices are non-volatile.

We have chosen two representative noise tolerance limits for our study that are  $\alpha = 0.15$  and  $\alpha = 0.25$ . Studies have shown that during training a Gaussian noise of standard deviation,  $\sigma$  that is up to 7.5% of the maximum magnitude of DNN weights does not degrade test accuracies significantly when no inference noise is assumed [40]. This motivates us to consider a noise tolerance limit of  $\alpha = 0.15$  that is 15% of the maximum DNN weights (most of the weights in Gaussian distribution lies within  $2\sigma \sim 15\%$ ). However, the DW device we studied here does have inference noise due to the device stochasticity. Furthermore, we

choose a maximum noise tolerance of  $\alpha = 0.25$  that is 25% of the maximum DNN weights so that the state overlaps between two adjacent states can be restricted for 5-state networks (half of the interstate interval for 5-state is 0.25).

We note that for 3-level quantization (3-state device), DW device can be programmed with control voltages to generate PMA of  $K_u = 8, 7.5, \text{ and } 7.0 (\times 10^5) J/m^3$  that can achieve quantized weights of -1, 0, and 1 respectively. For 2-level quantization (2-state device), the devices can be programmed to  $K_u = 8$  and  $7.0 (\times 10^5) J/m^3$  to achieve weights of -1 and 1 respectively. During in-situ training, the device weights are selected randomly from the  $\langle m_z \rangle$  distribution of corresponding  $K_u$  to program the DW device to a target quantized value. Although we have computed 250 instances for each of the programming conditions (in Fig. 4-1c-d) due to the limitation in computational resources, we note that there are dominant pinning sites in the racetrack because of the notches. As a result, the DWs tend to be stuck in or close to those pinning sites in most cases rather than the pinning sites offered by the rough edges of the racetrack (see section 4.5 Fig. 4-9). Hence, generating more instances will likely follow the probability distribution, which already exists in the current distribution.

## 2) Ex-situ Training

In this section, we discuss the steps of ex-situ training algorithm. The goal of ex-situ training is to achieve high testing accuracy in hardware although a precursor DNN is first trained in software. For this training, we also adopt weight quantization and allocate a separate memory in software where we store the high precision weights (similar to in-situ training) in addition to the DNN weights. The training algorithm shown in Fig. 4-2b remains the same for ex-situ training. After each iteration (forward and backward pass), high precision weights are updated and then quantized. Ideally, these quantized weights should be used as DNN weights for the next iteration in case of deterministic quantized neural network learning [1]. However, as we are dealing with a stochastic device for our inference engine, we include stochastic behavior of synaptic weights during learning. This stochasticity is obtained from a statistical distribution of the device (shown in Fig. 4-1d) rather than from uniform random distribution [1] or Gaussian distribution [40, 52]. For example, in 5-level quantization if the quantized weight is 0 then the DNN weight can be of any values selected randomly from the  $\langle m_z \rangle$  distribution of  $K_u = 7.5 (\times 10^5) J/m^3$  which is responsible for generating quantized weight of 0 (see Fig. 4-1d). The noise tolerance margin  $\alpha$  is also used during ex-situ training. This will relax the stringent requirement of programming a stochastic DW to a predetermined learned weight value and potentially save a large number of programming attempts. More importantly, if the DNN becomes aware of the statistical distribution of the device during training it can perform well during inference as the same device based DNN is used for inference.

Once the ex-situ training is accomplished, the DNN weights (or the high precision weights) are quantized and transferred to the DW devices by suitable programming. Here, the learned weights and the programmed weights may not be the same due to the programming noise. During the programming, we allow the same noise tolerance margin,  $\alpha$  that is used during training. Thus, any programmed device weight,  $W_{ij}$  need to satisfy,  $W_q - \alpha \leq W_{ij} \leq W_q + \alpha$  for a target-quantized weight of  $W_q$ . The devices can be programmed by repeated programming or performing read-verify-write operation in a loop, which is called ‘‘Open loop off device’’ method [53]. As we have already trained our network with stochastic distribution of weights by introducing finite  $\alpha$ , the network is expected to perform well during testing when we allow the same noise tolerance level for programming the device.

---

**Algorithm 4-1** In-situ training of a quantized neural network with crossbar array of DW devices.  $L$  is the number of layers,  $C$  is the cost function,  $\lambda$  is the learning rate decay and  $\alpha$  is the noise tolerance margin for writing the DW devices. Quantize () specifies how to quantize a weight with  $n$ -level quantization and Clip () specifies how to clip the weights based on Eq. 6. Update () specifies how to update weights once their gradients are calculated using stochastic gradient decent. These updated weights are accumulated in full precision (32-bit) in high precision memory unit. Program () specifies sending a voltage and current pulse to the DW device to write its conductance to a target quantized weight.

---

**Require:** a set of inputs and desired label  $(a^0, d^L)$ , previous DW device weights  $W_{ij,device}$  and corresponding full precision weights  $W_{ij,fp}$ , previous learning rate  $\eta$ .

**Ensure:** updated full precision weights  $W_{ij,fp}^{t+1}$ , corresponding DW device weights  $W_{ij,device}^{t+1}$  and updated learning rate  $\eta^{t+1}$ .

{ 1. Forward propagation in analog DW device crossbar: }

**for**  $k = 1$  to  $L$  **do**

$$a^k \leftarrow a^{k-1} W_{ij,device}^k$$

**end for**

{ 2. Backward propagation in analog DW device crossbar: }

{ Gradients are computed in digital unit built from CMOS devices }

Compute gradient  $g_{a^L} = \frac{\partial C}{\partial a^L}$  from  $a^L$  and  $d^L$

**for**  $k = L$  to 1 **do**

$$g_{a^{k-1}} \leftarrow g_{a^k} W_{ij,device}^k$$

$$g_{W_{ij}^k} \leftarrow g_{a^k}^T a^{k-1}$$

**end for**

{3. Accumulating the gradients in full precision in digital unit and update DW devices:}

**for**  $k = 1$  to  $L$  **do**

$$W_{ij,fp}^{k,t+1} \leftarrow \text{Update} (W_{ij,fp}^{k,t}, \eta, g_{W_{ij}^k})$$

$Level \leftarrow n$  //  $n$  represents maximum number of states of DW device

$$W_{ij,q}^{k,t+1} \leftarrow \text{Quantize} (\text{Clip} (W_{ij,fp}^{k,t+1}, -1, 1), Level)$$

**if**  $|W_{ij,device}^{k,t} - W_{ij,q}^{k,t+1}| > \alpha$  **do**

$$W_{ij,device}^{k,t+1} \leftarrow \text{Program} (W_{ij,device}^{k,t}, W_{ij,q}^{k,t+1})$$

**end if**

$$\eta^{t+1} \leftarrow \lambda \eta^t$$

**end for**

---

### Testing the DNN:

During the testing stage, we computed the predicted class for all the image samples from the MNIST test dataset using the trained DNN and compared it to the desired class. The percentage accuracy is calculated by dividing the total number of accurate predictions to the total number of test samples. During the testing stage, we consider two scenarios depending on in-situ or ex-situ training. When both the training and testing is performed on simulated hardware, the testing accuracy we record is termed online testing accuracy. In contrast, when the training is performed offline (ex-situ) in software, and we program the hardware (simulated device in this case) prior to testing according to the learned weights then the testing accuracy we record is termed offline testing accuracy.



### 4.3. Results and Discussions

#### DNN Configuration Selection:

The focus of our paper is to demonstrate the ability to classify images using a DW device based DNN and benchmark its performance against a DNN with floating precision (32-bit) weights. The topography of the benchmark DNN can be arbitrary as the inference accuracy varies across the spectrum of the parameters such as hidden layer number, layer size ratio (ratio of neurons between a layer and the previous layer) and learning rate constant (see Fig. 4-8 in section 4.5). Thus, one can select multiple configurations for the DNN and achieve good accuracy. We select a benchmark DNN architecture consisted of a network with three hidden layers, an initial learning rate of 0.007 and a layer size ratio of  $\frac{1}{2}$ . Also, we assume a learning rate decay of 10 % after each epoch and use stochastic gradient decent method as the optimizer. The selection criteria are detailed in section 4.5. After training the selected benchmark DNN for 10 epochs, the test accuracy we achieve is 97.1 %. We note that there are opportunities to improve the accuracy further in terms of topography, batch normalization, dropout layer and selection of different optimizers. However, the main goal of this study is to show how well a stochastic and low precision DW based DNN can perform in comparison to a similar architecture floating precision DNN. The selected topography mentioned above is used throughout the study to implement the DW device based DNN.

#### Online (In-Situ) Training:

After determining the DNN topography we investigate the test accuracies of the DNNs that are built from 2-state, 3-state and 5-state DW devices and trained with the proposed online training algorithm. For simplicity, we did not consider additional hardware non-idealities that could arise from peripheral circuits or unresponsive devices as these factors would automatically be included as constraints during the online training [40] and would not result in a significant degradation in performance compared to our current work.

The effectiveness of the proposed in-situ training algorithm is evident from Fig. 4-3a and Fig. 4-3b which plots the in-situ training accuracies for different state devices for low ( $\alpha = 0.15$ , 15% of maximum possible absolute weight) and high ( $\alpha = 0.25$ , 25% of maximum possible absolute weight) noise tolerance margin respectively. The results are also compared with baseline accuracy (accuracy of the same topography DNN with floating precision weights and no stochasticity). The training accuracies for DW device based DNNs increase with the number of device states and almost reach the baseline accuracy of  $\sim 99.6\%$  for low noise tolerance of  $\alpha = 0.15$  as can be seen from Fig. 4-3a. However, the training accuracies with high noise tolerance,  $\alpha = 0.25$  become slightly lower (see Fig. 4-3b) as these networks allow higher deviation from the

target quantized weights. Nonetheless, competitive training accuracies are achieved for both 3- and 5- state devices with high noise tolerance margins.

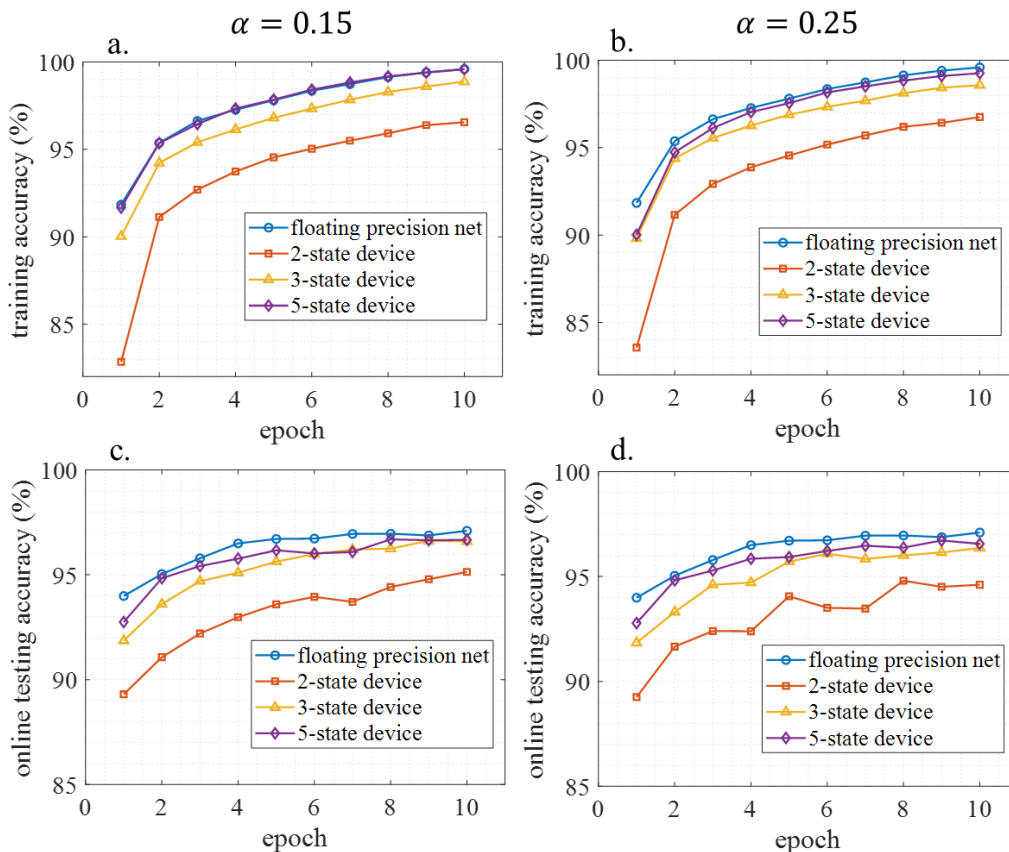


Figure 4-3 Online training accuracy and online testing accuracy for DNNs with different state DW devices for two different noise tolerance margins. These accuracies are compared with a DNN trained and tested with full precision weights and no stochasticity (baseline accuracy) a. and b. show the online training accuracies with the numbers of epochs for  $\alpha = 0.15$  and  $0.25$  respectively. c. and d. show online testing accuracies with numbers of epochs for  $\alpha = 0.15$  and  $0.25$  respectively.

After each epoch of the in-situ training we test the DNN with test images from MNIST dataset and compute the online test accuracy. Fig 4-3c and 4-3d plots online testing accuracies for low and high level of noise tolerance margin respectively. The baseline (DNN with floating precision weights and no stochasticity) test accuracies are plotted for comparison. For low noise tolerance margin of  $\alpha=0.15$ , the test accuracy is highest for 5-state device and reaches  $\sim 96.67\%$  after 10 epochs of training. This accuracy is very close to the baseline test accuracy of  $\sim 97.1\%$ . It is important to note that the 3-state device based DNN achieves a test accuracy of  $\sim 96.6\%$  after 10 epochs of training, which is similar to a 5-state device. When the noise tolerance margin is increased to  $\alpha=0.25$ , the test accuracies for 5-state and 3-state devices are  $\sim 96.56\%$  and  $\sim 96.36\%$  after 10 epochs of training. Thus, a maximum decrease of accuracy of  $\sim 0.74\%$  from 32-bit

precision weight is recorded for a 3-state stochastic weight. We note that, the test accuracies for 2-state device are  $\sim 95.14\%$  and  $\sim 94.64\%$  for low and high noise tolerance margin respectively. Thus, the same topography networks for 2-state does not achieve comparable test accuracies. Changing the topography, such as increasing the number of neurons in hidden layers, can increase the accuracy of binary DNN [54].

Next, we analyze the total number of programming pulses that are applied to the DW devices during the course of the online training at various epochs. Because the network updates the high-precision weights, a single weight may have its high precision value updated many times before crossing the threshold to update the DW device weight. As the number of device updates is dependent on the number of times a high precision weight crosses the threshold, the larger the threshold the fewer the updates. Between the 2, 3 and 5 state networks the 5-state has the smallest threshold, which increases the number of DW device updates as seen in Fig. 4-4. These DNNs are also compared with a DNN trained with floating precision weights (and no stochasticity). In floating precision DNN, all the weights are updated each time a training image is passed to the network. Thus, although the network is better trained with an increasing number of epochs, the weight update count remains almost constant as seen in Fig. 4-4. In contrast, for DNNs with limited state DW devices with the proposed training method, the programming instances decrease significantly with the number of epochs. As expected, with low noise tolerance margin the DNNs with DW devices become more selective and require higher number of weight updates during the course of the training (though this is much smaller than the case of floating precision weights).

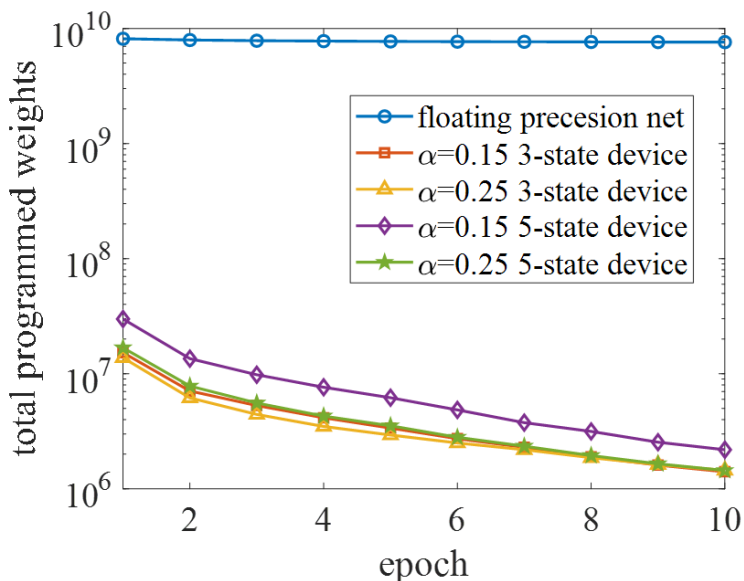


Figure 4-4 Comparison of the total number of programmed weights with the number of training epochs for different networks. A significantly lower number of weights are updated during the proposed online training compared to the floating precision weight network of the same architecture.

In Fig. 4-5, we show the convergence of DW device weights during the training. DNN weights whose noise tolerance are higher will converge to a value quicker, on an average, than a weight with a lower noise tolerance. In Fig. 4-5a and 4-5b, the DW device weights fall within  $\pm\alpha$  of the quantized weight value. In both cases, the DW device weight is closer to the high precision value than the quantized weight, which tends to provide a higher accuracy for our DW based DNN.

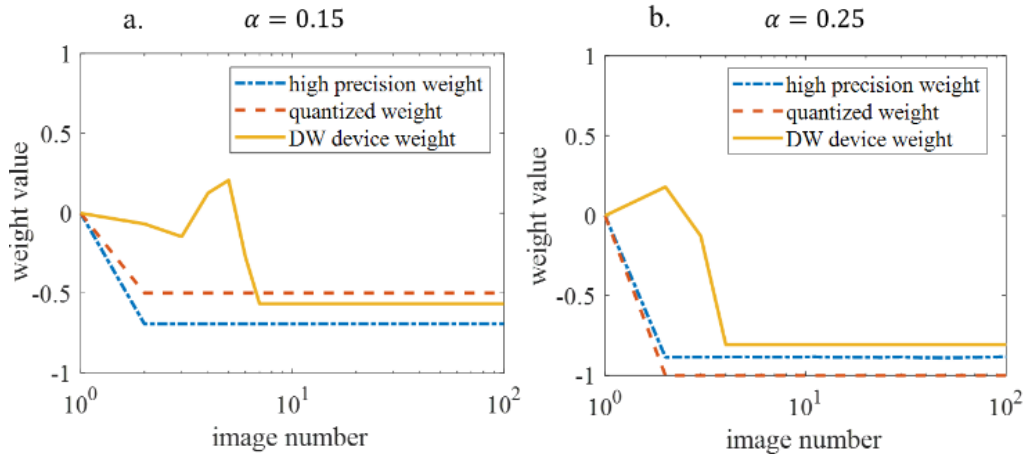


Figure 4-5 Weight evolution of high precision weight, quantized weight and the DW device weight during the first few training images for two different noise tolerance margin a.  $\alpha = 0.15$  b.  $\alpha = 0.25$ . The synaptic weight shown here is connected between the neurons located in hidden layers 2 and 3.

### Offline (Ex-Situ) Training:

In this section, we first analyze the effectiveness of our proposed ex-situ training by comparing it with other techniques. For that, we train several precursor DNNs in software using different offline training algorithms (Fig. 4-6) and then test the DNNs, which are built from DW synaptic devices (3- state and 5-state hardware). Each of the DNNs are trained offline with a total of 10 epochs (train with entire training dataset 10 times) and prior to the testing the DW devices are programmed according to the weights that are learned offline. These results are shown in Fig. 4-6a and 4-6b when we consider a low ( $\alpha = 0.15$ ) and high ( $\alpha = 0.25$ ) value of noise tolerance margin to program the devices. In both Fig. 4-6a and 4-6b, for each of the hardware test accuracies, a corresponding software accuracy is presented side by side with green and yellow bar. When the exact learned weights (no programming noise is considered while transferring the learned weights to the device) are used to test the DNNs we call it software accuracy.

When offline training is performed with both the floating precision and quantized weights cases, the test accuracies are low for low noise tolerance margin, as can be seen from Fig. 4-6a. After floating precision weight training, the learned weights need to be converted to 3- or 5-state to program the DW devices. Thus, for both of the 3- and 5-state hardware the test accuracies degrade compared to software accuracy of ~

97.1%. Converting floating precision learned weights to 5-state compatible weights (5-level quantization) generates smaller deviations compared to the 3-state weight (3-level quantization). Thus the 5-state device provides higher test accuracy which is  $\sim 87\%$  compared to the 3-state which is only  $\sim 10\%$ .

Training with quantized weights (as proposed in Ref [1]) improves the test accuracies to  $\sim 90\%$  for 5-state device (see Fig. 4-6a) as the network becomes aware about the limited states of the weights during the training period. However, the test accuracy remains low (software accuracy is  $\sim 96.74\%$ ). The accuracy loss is mainly due to the deviation of the programmed weights from the learned weights. We note that, with floating precision training, weight deviations occur in two ways: converting the floating precision weights to quantized weights and during the programming of the device where the target quantized weights are not achieved deterministically. However, with quantized training only the latter deviation occurs during the testing stage.

In contrast, with our proposed training which we call quantized + stochastic training, the test accuracy increases and reaches up to  $\sim 96.63\%$  for 5-state device, which is very close to the software accuracy of  $\sim 96.67\%$ . The accuracy improvement can be attributed to even smaller deviation of the programmed weights from the learned weights. Unlike quantized training, in our proposed training the weight quantization is also accompanied by training the DNN weights according to the statistical distribution of the device. As a result, during back propagation, the high precision weights are updated depending on the weighted sum performed over the imprecise DNN weights (which are mapped from the stochastic distribution of the device as in Fig. 4-1d). In other words, the high precision weights are being tuned based on the stochastic signature of the device. Thus, the statistical distribution of the device is embedded in the learning. When the same devices are used for testing, the distribution matches better and this plays an important role in improving test accuracy. This finding is also supported by other works [40, 52]. Ref [40] shows that the DNN trained with Gaussian distributed weights of a certain standard deviation performs better when a weight distribution of same standard deviation is used for inference.

With high programming noise, for both floating precision and quantized training, the programmed weights deviate more from the learned weights because of the higher noise tolerance. Thus, the test accuracies for 3- or 5-state hardware degrade significantly compared to the software-based accuracies as seen from Fig. 4-6b. In contrast, with our proposed training method, the DNNs are made aware of the statistical distribution of the device thus resulting in significantly higher test accuracies compare than other offline training methods. (Note that as the device statistics are not Gaussian and instead heavily dominated by the pinning positions, training with Gaussian distributed weights does not improve accuracy and was not employed).

We also studied the evolution of offline test accuracies with the number of epochs for different state devices, which are presented in section 4.5 in Fig. 4-10. The influence of noise tolerance margin  $\alpha$  on training accuracy, online testing accuracy and offline testing accuracy for DNN with limited state device (5-state) is shown in section 4.5 in Fig. 4-11, which shows that offline testing accuracy is affected most by the choice of different  $\alpha$ .

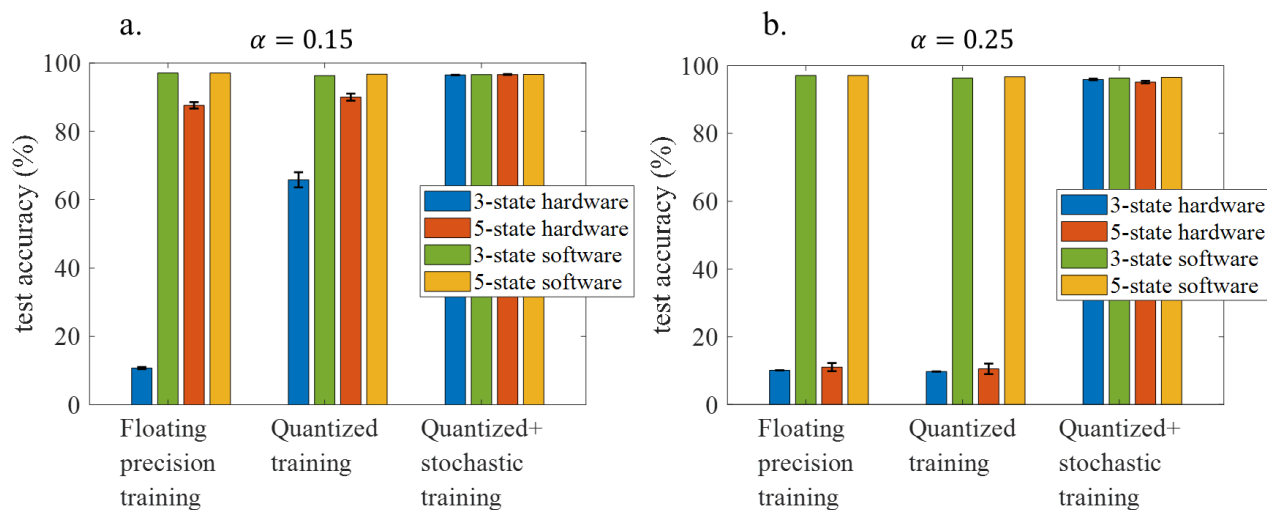


Figure 4-6 Testing accuracy comparison of 3-state and 5-state DW device based DNNs for different ex-situ training algorithms with a programming noise tolerance margin of a.  $\alpha = 0.15$  b.  $\alpha = 0.25$ . The networks are trained offline with floating precision weights, quantized weights, and stochastic quantized weights derived from micromagnetic simulation. Each of the networks is trained with a total number of 10 epochs. Once training is done, the 3-state and 5-state DW devices are programmed based on the quantized value of trained weights prior to testing. For different training algorithms and for each of the test accuracies of DNNs built from 3- and 5-state hardware, a corresponding software test accuracy (no programming noise is considered, and exact trained weights are used for testing the DNN) is plotted side by side with green and yellow bar. The error bar seen in the figure is calculated from 10 different test trials. For both noise tolerance margins, the test accuracy is highest when the DNNs are trained with proposed training algorithm (quantized + stochastic).

#### 4.4. Energy Dissipation

The energy required to program a DW synapse is determined from charging the piezoelectric layer with a voltage pulse,  $\frac{1}{2}CV^2$  and  $I^2R$  loss due to the SOT current in the heavy metal layer. The maximum change in PMA is  $\Delta PMA = 0.5 \times 10^5 J/m^3$ . For magnetic racetrack of CoFe the saturation magnetostriction is,  $\lambda_s=250$  ppm. Thus, the maximum required stress,  $\sigma$  is,  $\frac{\Delta PMA}{\frac{3}{2}\lambda_s}=133$  MPa and the strain is,  $\frac{133 MPa}{200 GPa} \sim 10^{-3}$ , considering the Young's Modulus of CoFe to be 200 GPa. When the electrode dimensions are in the same order as the piezoelectric thickness, previous study [55] demonstrated that  $10^{-3}$  strain is possible in Lead

Zirconate Titanate (PZT) with an applied electric field of  $E=3$  MV/m. If we consider PZT layer to be  $b=60$  nm thick (same as top electrode or racetrack width as illustrated in Fig. 4-1(b)) then a voltage of,  $E*b = 0.18$  V applied between the top electrode pair and the bottom electrode can generate the required strain. For a top electrode of length  $L=600$  nm (same as racetrack length 600 nm) and width  $b=60$  nm and a relative permittivity of PZT  $\epsilon_r=3000$ , the effective capacitance is calculated to be  $\frac{\epsilon_0\epsilon_r(L*b)}{b} \sim 16$  fF. This predicts a  $\frac{1}{2}CV^2$  loss of  $\sim 0.5$  fJ, considering two top electrodes on each side of the racetrack.

The heavy metal layer is considered to be Pt and for  $600 \times 60 \times 5$  nm<sup>3</sup> dimension Pt layer the resistance is calculated to be  $200 \Omega$  assuming the resistivity of Pt to be  $100 \Omega\text{nm}$ . The heat loss in the heavy metal layer is calculated to be  $2.2$  fJ for a fixed SOT generating current pulse of magnitude  $35 \times 10^{10}$  A/m<sup>2</sup> applied for 1 ns. Thus, the maximum energy dissipation to program a synapse is calculated to be  $2.7$  fJ.

### 1) In-situ Training

With in-situ training, the highest inference accuracy is achieved for a 5-state device when a low noise margin is considered during the training. However, with a higher noise tolerance margin similar test accuracy is obtained with fewer device updates as can be seen from Fig. 4-4. For 5-state device, if we consider a noise tolerance margin of  $\alpha=0.25$ , the total number of weight updates are calculated to be  $\sim 48$  million after running the training for 10 epochs. *Thus, the energy dissipation to program the DNN synapses is calculated to be  $\sim 13$  pJ for one inference event followed by the weight updates (10000 test images in MNIST).*

### 2) Ex-situ Training

With ex-situ training, highest inference accuracy is achieved for 5-state device when the noise margin to program the DW devices is considered to be low. Fig. 4-7 shows the cumulative probability of the DW device weights for different programming condition for a 5-state device. The solid black line represents the target quantized weights of 1, 0.5, 0, -0.5 and -1 (in this case -0.833) which can be achieved by a combination of fixed SOT current pulse and a varying amplitude voltage pulse which modulates the anisotropy of the racetrack to  $K_u = 7, 7.25, 7.5, 7.75$  and  $8.0 (\times 10^5) \text{ J/m}^3$  respectively. The adjacent red dotted line in the figure shows the noise margin ( $\alpha = 0.15$ ) that is allowed while programming the DW device to a specific quantized state. From Fig. 4-7, it can be seen that the probability of programming the DW device weight to a quantized value of 1 is the lowest which is  $\sim 6\%$  meaning a number of  $\sim 20$  attempt is required to program the device. If we consider the worst-case scenario, then after ex-situ training prior to the inference, we need 20 programming pulses to program each of the DW devices implementing the DNN

weights. Thus, for our network topology of 784-392-196-98-10 neurons, the energy dissipation to program the DW synapse is 2.8 pJ per inference event.

The energy dissipation to program the DW devices in in-situ training is found to be  $5\times$  the dissipation incurred in ex-situ training, which is a moderately low provided that the training is performed over the entire 60000 training images for 10 times. This low-dissipation in-situ training is possible due to distinct features of proposed training algorithm that benefits from weight quantization and noise tolerance margin. Large inter-state interval in quantized learning helps to reduce the number of weight updates. Moreover, once the device is programmed within the noise tolerance margin, further write operation is avoided with a simple low cost read operation. We note that onsite learning is attractive in power constraint edge devices, where the learning itself needs to adapt and respond to a continuously evolving environment. Embedded medical systems [56], real time intrusion detection [57], and dialect specific speech recognition systems can benefit from such onsite learning. Ex-situ learning can perform inference tasks in edge devices with energy efficient manner (given the training is performed over cloud server), however the benefit can only apply to non-adaptive tasks.

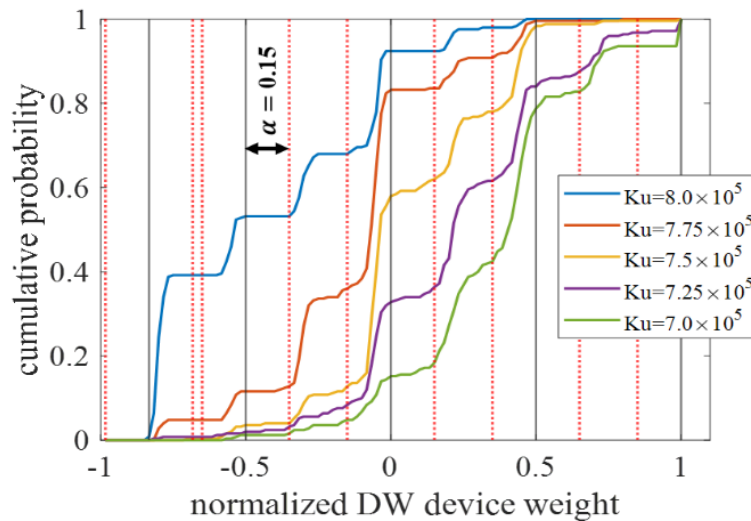


Figure 4-7. Cumulative probability of normalized DW device weights for 5-state device under different programming conditions denoted by different  $K_u$ . Black solid line represents the target quantized weights and the adjacent dotted red lines represent the programming noise tolerance margin of  $\alpha = 0.15$ .

Finally, the accuracy and energy consumption of our proposed DW based approach is compared with state-of-the-art techniques in the literature. The accuracies that we achieve for 5-state DW are comparable to the RRAM [54] and PCM [38] and better than the DW approach presented in [48] that can provide 32-states



For energy comparison purpose, we first calculate the energy consumption of our proposed in-situ approach, including the energy expenditure for performing forward and backward propagation in the analog domain (crossbar devices) and weight gradient accumulation in the digital domain (high precision memory update). The details about energy calculation can be found in the next section. In addition, we estimate the energy consumption of a similar architecture deep neural network (DNN) with 32-bit precision weights (see section 4.5). Our proposed approach demonstrates a possibility of  $\sim 165\times$  energy saving compared to the 32-bit precision DNN implemented with on-chip CMOS static random-access memory (SRAM).

The estimated energy consumption of  $\sim 26$  nJ per inference is comparable with state-of-the-art non-volatile technologies such as RRAM [54] and PCM [38]. Moreover, our proposed 5-state DW based DNN consumes less power compared to 32-state DW based DNN [48] for each synaptic weight update event as a  $50 \mu A$  and 1 ns duration current pulse is used to program the 32-state synapses as opposed to our synapse that requires  $21 \mu A$  and 1 ns duration current pulse (Note that SOT clock dominates the energy consumed in our case). Further, the DW-based approach presented in [25] consumes an energy  $\sim 8.64$  fJ to program the synapse from one extreme conductance to the other, compared to our  $\sim 2.7$  fJ. However, Ref [25] does not take thermal noise and edge irregularities into consideration that could significantly reduce the number of distinguishable states due to device stochasticity. Finally, our algorithm ensures the number of times the weights are programmed are also significantly lower, making the training cost very small.

## 4.5 Additional details

### Selection of Deep Neural Network (DNN) Architecture:

The network topography we aim to optimize, in terms of training and testing accuracy, uses floating precision (32-bit) weights for the synapses. This optimized network will be used to benchmark the performance of the domain wall (DW) based DNN of the same network topography.

To select the optimal network topography, multiple characteristics are considered: training accuracy, testing accuracy, network size, and overall fitness. An algorithm is designed to create, train, and test networks; each having a unique set of parameters such as number of hidden layers, the ratio of the number of neurons in a layer to the number of neurons in the previous layer or “layer size ratio” and a learning rate; the ranges for each being 0 to 9, 0 to 1, and 0 to 0.009 respectively. Parameters that remain consistent, during optimization, are the learning rate decay and the starting epoch of learning rate decay. All the networks are trained using only one epoch resulting in underfitting networks where the training accuracy was lower than the testing accuracy. It is also important to note that many networks, when trained for multiple epochs, can

achieve high training accuracies but at the expense of overfitting to the training data therefore limiting the number of epochs, to one, places priority on having high initial accuracies relative to other network topographies rather than achieving independent high accuracies.

The learning rate decay is chosen to be 10 % for each epoch with the decay starting from the very first epoch. For example, if the initial learning rate is 0.007, the final learning rate at the end of the first epoch would be 90% of the initial learning rate i.e., 0.00693. The networks are trained on the full set of 60,000 training images and tested on the full set of 10,000 testing images from the MNIST handwritten digits database. From hidden layer number variation results (not shown here) we find that the accuracy increases with the increase of the number of hidden layers, however, when the hidden layer number is greater than 3 the accuracy does not increase appreciably. Therefore, we select a total of 3 hidden layers for our DNN configuration. Fig. 4-8 shows the height map for training accuracy, testing accuracy and the average of training and testing accuracies for different topography with 3-hidden layers. We can see from Fig. 4-8 that the accuracies are higher when the network's learning rate is within 0.004 and 0.007 and the ratio of the number of neurons in a layer to the number of neurons in the previous layer is within 0.4375 and 0.5625. When the algorithm was left to run multiple times with different initialized weights, the topography that possessed the maximum accuracy changed each time but remained in the region of maximum training and testing accuracy as previously described. Thus, one can select several configurations of DNN and achieve good accuracies if the topography parameters fit within the previously mentioned region.

We choose a layer size ratio of 0.5, whereas for the selected network the number of neurons for the first hidden layer is  $0.5 \times$  the number of neurons of the input layer and so on. We choose a learning rate constant of 0.007 as it would provide sufficient learning capacity with the increase of training epoch (learning rate decays by 10 % at each epoch). Thus, the final architecture consists of a network with three hidden layers, an initial learning rate of 0.007 and a layer size ratio of 0.5.

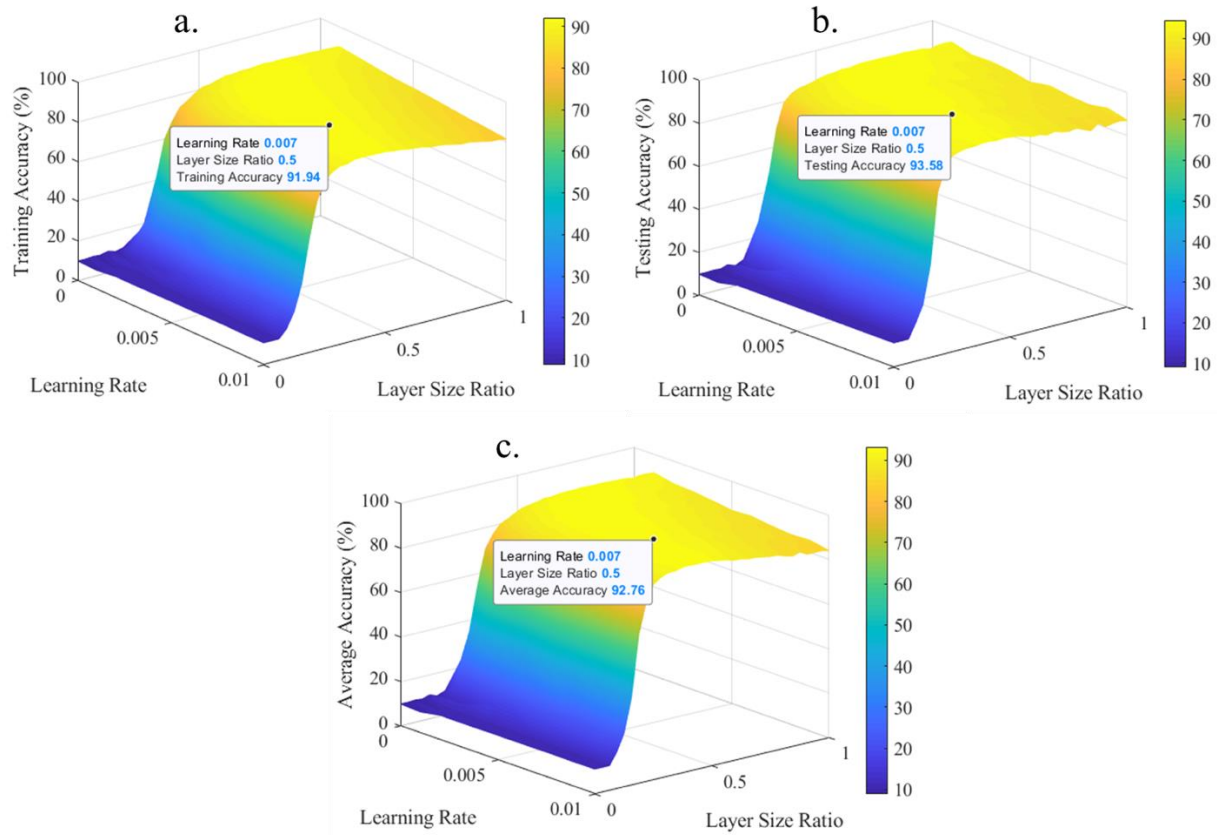


Figure 4-8 Height maps show the a. training b. testing and c. average of training and testing accuracies of a 3-hidden layer DNN for varying topographies of the network. The topographic features that are varied are the learning rate and the ratio of the number of neurons in a layer to the number of neurons in the previous layer or “layer size ratio”. The highlighted data point is the final topography chosen for our DW device based DNN as it gives high accuracies for a small number of synapses.

### Equilibrium DW position distribution with respect to dominant pinning sites (notches):

The following Fig. 4-9b-f shows the equilibrium DW position distribution for the racetrack shown in Fig. 4-9a for five different programming conditions represented by  $K_u = 8, 7.75, 7.5, 7.25$  and  $7.0 (\times 10^5) J/m^3$ . Although a few of the DWs are pinned stochastically due to the pinning sites offered by the edge roughness ( $\sim 2\text{nm rms}$ ) and room temperature ( $T=300\text{ K}$ ) thermal noise, however, they are predominantly pinned at or near the location of the engineered notches. Thus, the distribution is heavily influenced by the dominant pinning site locations or the notches.

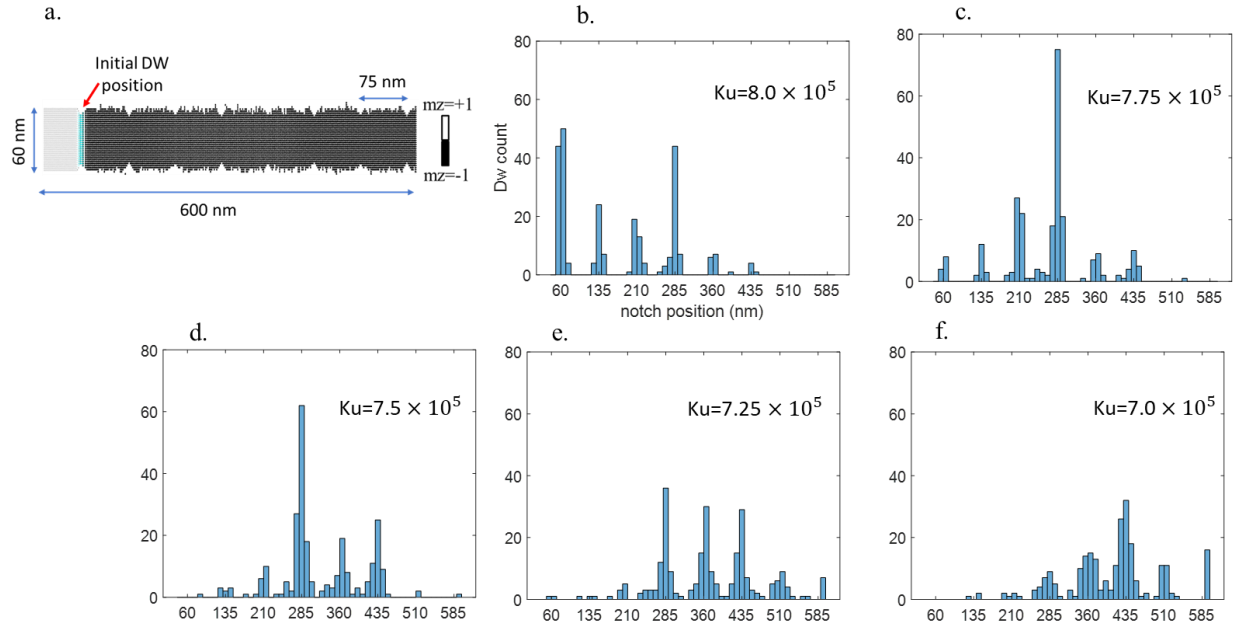


Figure 4-9 a. Racetrack of dimension  $600 \text{ nm} \times 60 \text{ nm} \times 1 \text{ nm}$  with rms edge roughness of  $\sim 2 \text{ nm}$  hosting a DW at an initial position of  $60 \text{ nm}$  from the left end. Engineered notches starting from  $60 \text{ nm}$  to the left of the racetrack are placed at a regular interval of  $75 \text{ nm}$ . b.-f. Distribution of equilibrium DW positions along the racetrack shown in Fig. 4-9a for different programming conditions represented by different PMA coefficient,  $K_u$ . The DWs are primarily pinned at or around the notches, thus the distribution become dominated by different notch locations for different programming conditions.

### Evolution of offline testing accuracies with training epochs:

The offline testing accuracies with the number of epochs are presented in Fig. 4-10a and Fig. 4-10b for both low and high noise tolerance margins. The baseline testing accuracies when using floating precision weights and no stochasticity are also plotted for comparison. After each epoch of training, the trained weights are collected and the devices, in the crossbar arrays, are programmed. As we have allowed for a noise tolerance margin,  $\alpha$ , during testing to program the devices, the programmed weights of the network during each testing trial could be different but within the range of tolerance. Therefore, we consider a total of 10 different trials for testing the DNNs. The error bar shown in the figure is computed from 10 different test trials after each epoch of the training. After 10 epochs of training for 5-state and 3-state devices as seen from Fig. 4-10a, for a noise tolerance margin of  $\alpha=0.15$ , the offline testing accuracies approaches the baseline testing accuracies. The highest offline test accuracy of  $\sim 96.63\%$  is achieved for 5-state device which is very close to the baseline test accuracy of  $\sim 97.1\%$ . However, for  $\alpha=0.25$ , the test accuracies degrade for all the devices as can be seen from Fig. 4-10b. The highest offline test accuracy for 5-state, 3-state and 2-state devices are obtained after 10 epochs of training which are  $\sim 95.14\%$ ,  $\sim 95.93\%$  and  $\sim 94.24\%$  respectively.

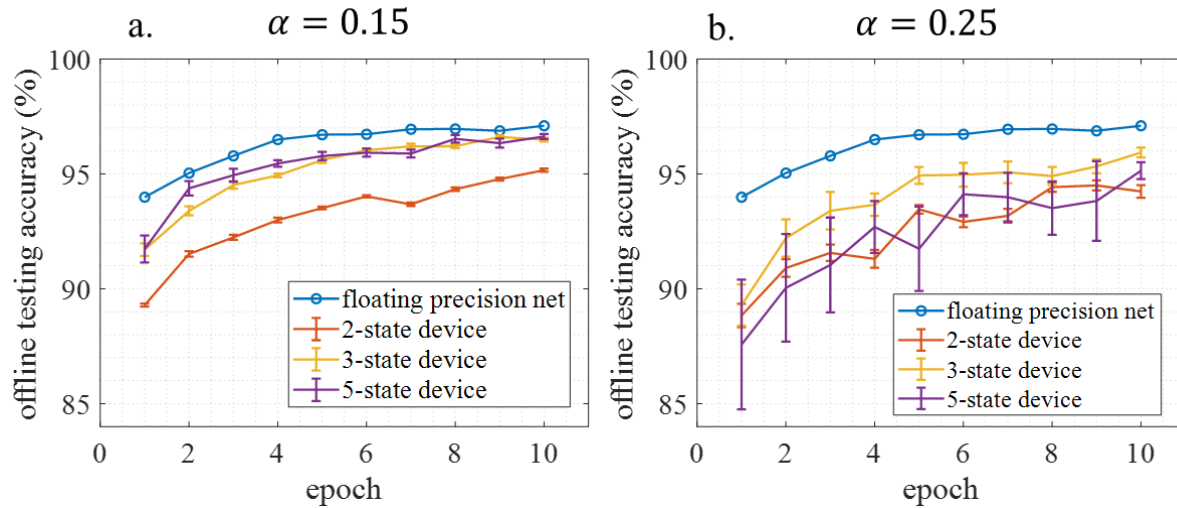


Figure 4-10 Offline testing accuracies for DNNs of different state DW devices with two different noise tolerance margins, a.  $\alpha = 0.15$  and b.  $\alpha = 0.25$  used during the training and programming of the devices. The accuracies are compared with a DNN trained and tested with 32-bit (floating) precision weights and no stochasticity (baseline accuracy). Error bar is calculated for a total of 10 different test trials.

### Influence of Noise Tolerance Margin on Accuracy:

To examine the influence of a noise tolerance margin on network accuracies; training accuracy, online testing accuracy (training is done in-situ) and offline testing accuracy (training is done ex-situ) of 5-state DW based DNNs are compared in Fig. 4-11 for high and low level of noise tolerance margin,  $\alpha$ . From Fig. 4-11a and 4-11b, we can see that the difference in noise tolerance margins does not appear to have a significant effect on the training or online testing accuracies. This is due to the fact that backpropagation is performed over the imprecise DW device weight which is then used to update the high precision weights for the network. In effect the potential values selected at random for the DW device weights are then known by the network and, through training, changes the DW device weights accordingly based on the quantized value of the high precision weights. Once trained in-situ, the learned weights remain the same during testing due to DW device non-volatility. In contrast, after ex-situ training the learned weights are transferred to the DNN, for testing, by programming the DW device with a noise tolerance margin that is used during the training. When the devices are programmed prior to testing, programming noises are added to the DW device weights. The higher the noise tolerance margin, the higher the amount of noise that could be added to the programmed device weights. Thus, the offline testing accuracy degrades with higher noise tolerance margin. For example, if the trained weight is 0.24 then the quantized weight would be 0 and after programming, with a noise tolerance margin of  $\alpha = 0.25$ , the DW device weight could be -0.24. Thus, with the noise tolerance margin of  $\alpha = 0.25$ , the maximum deviation of the programmed weight from the learned

weights could be  $2\alpha=0.5$  whereas for a low noise tolerance margin  $\alpha=0.15$  the maximum deviation could be  $2\alpha=0.3$ .

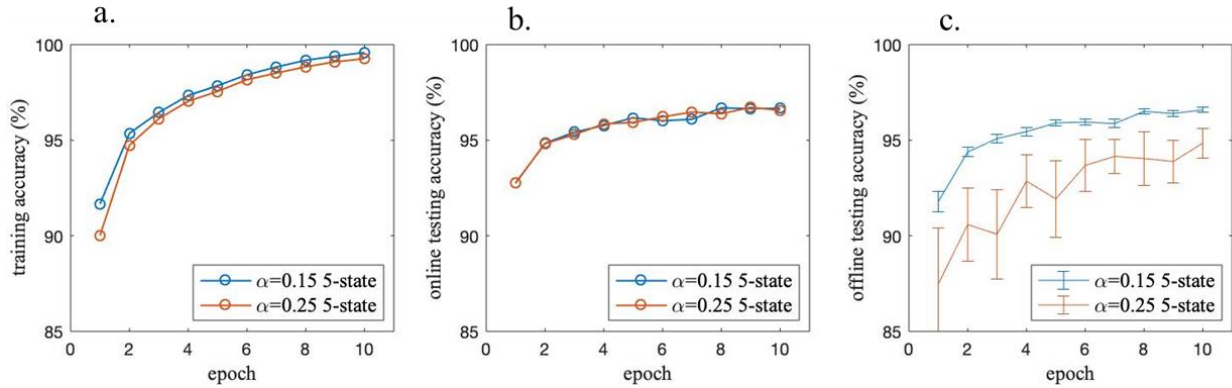


Figure 4-11 a. Training accuracies b. Online testing accuracies and c. Offline testing accuracies for a 5-state DW device based DNN for two different noise tolerance margins of  $\alpha$ . The training accuracy and online testing accuracy does not change appreciably for different noise tolerance margins. Offline testing accuracy decreases with high noise tolerance margin due to the higher deviation of device weights during the programming of the devices.

### Energy Dissipation Estimation of the Proposed Technique:

The scope of the operations performed in the analog and digital domains for our proposed in-situ training of a DNN is presented in Fig. 4-12. The energy consumption in the analog domain depends on performing matrix vector multiplication during forward and backward propagation and updating the DW device weights. In the digital domain, energy is spent for computing neurons activation, error gradients, and accumulating gradients for weight update. During analog computation, energy is consumed in several stages, such as serially in and out data to and from the crossbar rows and columns, performing digital to analog conversion of the input data to read voltage pulses using Pulse-width modulation (PWM), regulating the column voltage to a specific value so that a corresponding read voltage drops across the DW devices, reading the analog weighted sum in the crossbar arrays, and performing analog to digital conversion (ADC) of the weighted sum before sending them to the crossbar for implementing the other layers of the DNN. The main supply voltage, read voltage, and the clock frequency are assumed to be 0.8 V, 100 mV and 2 GHz. We consider 8-bit resolution for the PWM and ADC and assume the bit shifting energy to be 2 fJ/bit, PWM counter energy to be 50 fJ per counting step and the energy to buffer the PWM comparators output to be 10 fJ per turn on or turn off event [38]. The bias current for the operational transconductance amplifier used to regulate the column voltage is assumed to be 50  $\mu A$  [38] and considered to be turned on for one read pulse duration (2 ns). Further, the mean resistance of the DW-based MTJ device (and the devices for parallel conductances) is considered to be 20 k $\Omega$  [58] when computing the energy for the analog weighted sum. The ADC energy for 8-bit conversion is assumed to be  $\sim 330$  fJ at a conversion delay of 15 ns, based

on [59]. Considering all the above parameters the energy consumption in the analog computation stage is estimated to be  $\sim 8$  nJ for forward propagation and  $\sim 2.5$  nJ for backward propagation for each of the training images. The detailed guidelines for energy calculation are presented in Ref. [38]. For DW device weight updates, an average of  $\sim 80$  devices are updated during each training instance which results in an energy consumption of 220 fJ per training instance with a write energy of  $\sim 2.7$  fJ/update. Thus, the total energy in the analog computation stage is estimated to be  $\sim 10.5$  nJ per training instance. The energy dissipation in transistors for switching is negligible ( $\sim 100$  aJ [60]), thus omitted during the energy estimation.

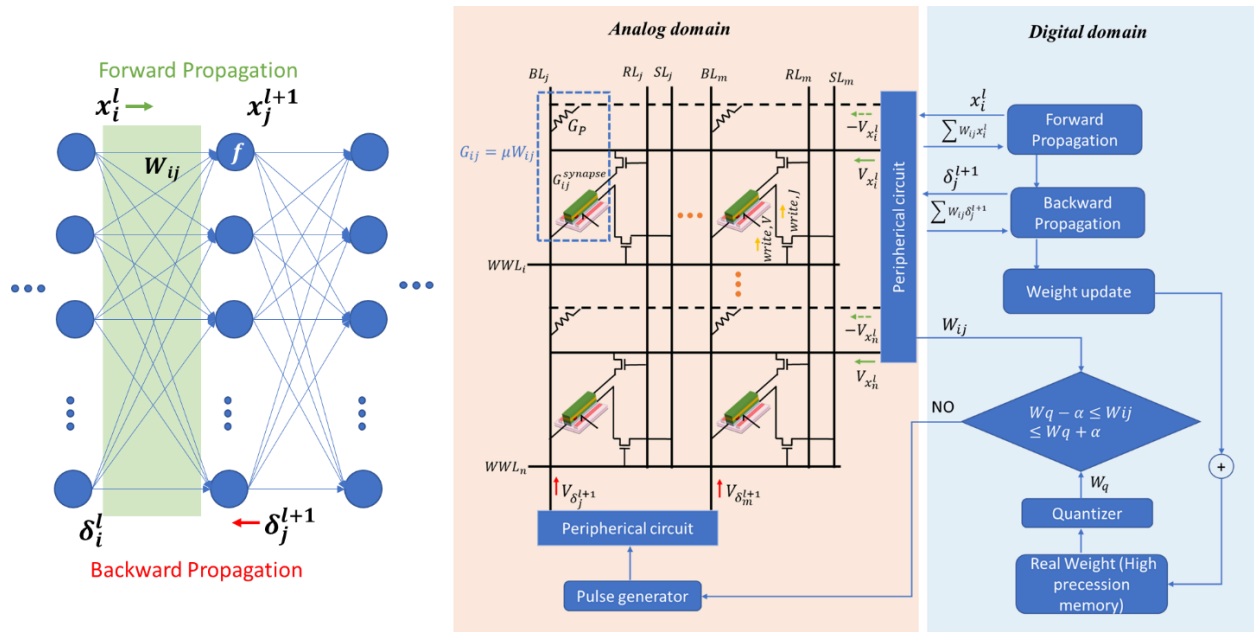


Figure 4-12 Proposed in-situ training of two successive DNN layers (green shadowed region). The scope of operations that are performed in analog and digital domain during the training is shown in two different colored boxes.

For DW device weight updates, gradients are first accumulated in a digital unit using 32-bit precision memory (see Fig. 4-12). However, for weight gradient computation, if the neuron activations (forward propagation) and error gradients (backward propagation) are quantized to 3-bit, then a significantly lower number of 32-bit memory access is possible due to a small number of non-zero entries (due to quantization) without negatively impacting accuracy [38, 61]. Further, if the memory is accessed a number of times close to 1% of the total synapses (that is  $\sim 4044$  synapses, which is 1% of 4,043,48 synapses), the weight update energy in the digital unit can be estimated to be  $\sim 122$  nJ based on Ref. [38]. Moreover, in our design we require a quantizer (i.e., ADC) for quantizing the accumulated gradient before comparing it with existing DW device weights (by reading the device conductance) to allow for the DW device update within a noise margin. The quantization operation needs to be performed whenever there is an update in the 32-bit

precision memory (i.e., 4044 times) and read (and comparator) operation is performed for  $\sim 4044$  times plus the extra number of times when the devices are updating within a targeted noise margin. Considering a total of  $\sim 20$  attempts to program the device within noise margin (worst case scenario), a total of  $\sim 80 \times 20$  read (and comparator) operations is required on top of the number of quantization operation. By considering an analog read operation is preceded by PWM input signal (4-bit PWM suffices in this case as PWM output voltage resolution does not affect reading the device conductance), read voltage regulation and followed by an ADC operation, the total read energy is estimated to be  $\sim 13.8$  nJ per training instances for reading the device conductances and the corresponding parallel conductances. Assuming the comparator energy to be equal to that of ADC energy  $\sim 330$  fJ, the total energies for quantizer, read (and comparator) operations are estimated to be  $\sim 17$  nJ per training instance.

Furthermore, the amount of energy dissipated in a digital unit for computing neuron activations (in forward propagations), error gradients (in backward propagations) and correctly addressing the analog DW devices for sending write pulses can be extended from the application specific integrated circuit (ASIC) design implemented by Ref. [38] using on-chip static random access memory (SRAM). In their design, the energy consumption in forward and backward passes on a digital unit are shown to be  $\sim 9$  nJ and  $\sim 3$  nJ per training instance. Since the number of synapses of our architecture are twice that of Ref. [38], we can have an estimate of energy consumption for our architecture of  $\sim 18$  nJ and  $\sim 6$  nJ. Thus, the total energy consumption in the digital unit is estimated to be  $\sim 163$  nJ per training inference. Combining the analog and digital units' energy together, the energy consumption is estimated to be  $\sim 174$  nJ per training instance and  $\sim 26$  nJ per inference instance (energies for the forward propagations only).

Next, we estimate the energy computation of a similar architecture 32-bit DNN. For the 32-bit system architecture, we refer to the design proposed by Ref. [38] which demonstrates an ASIC design optimized with on-chip SRAM (i.e., where off-chip data communications are avoided) using modern 14 nm low power plus (LPP) technology. The DNN architecture used in [38] is 784-250-10 (neurons in different layers). Thus, for our network architecture, of 784-392-196-98-10, the synapse counts are doubled which results in twice more memory registers and weight updates, as in 32-bit precision system all the weights are updated at each training instance. Extending their design for our DNN architecture, the energy consumption during in-situ training of 32-bit precision weights can be estimated to be  $\sim 29$   $\mu$ J per training instance.

Therefore, the energy consumption of our proposed design is significantly lower than the 32-bit precision DNN implemented with on-chip SRAM and demonstrates a possibility of  $\sim 165$  times more energy savings with in-situ training. Furthermore, the estimated energy consumption of  $\sim 26$  nJ per inference instance is comparable with state-of-the-art resistive random-access memory (RRAM) [54] and phase change memory (PCM) [38].



## 4.6 Conclusion

We have shown that DNNs with extremely low resolution and stochastic DW device-based synapses can achieve high classification accuracy when trained with appropriate learning algorithms. In this study, both in-situ and ex-situ training algorithms are presented for DNNs that are implemented with 2-state, 3-state and 5-state DW devices. For in-situ training, a high precision memory unit is employed to preserve and accumulate the weight gradients, which are quantized to obtain target conductance for updating the low precision DW devices. A noise tolerance margin further allows for random deviations of the programmed conductances from the target conductance values. For ex-situ training, a precursor DNN is first trained in software by performing weight quantization and considering a noise tolerance margin from the quantized weight and later tested with an equivalent DNN of DW devices programmed with the same noise margin. While the energy dissipation statistics for programming the DNN synapses shows that ex-situ method is energy efficient, however, the in-situ training comes with an opportunity to learn and adapt to the changing environment with only  $5\times$  more dissipation (despite the fact that the in-situ training is performed over a vast number of training images for many epochs). This technology is specifically attractive for low power intelligent edge devices of future IoT where energy requirement is at a premium.

Although our quantization-aware, stochastic domain wall (DW) device-based DNN demonstrated high accuracy in classification tasks, the effectiveness of the proposed methods for other neural network architectures, such as convolutional, recurrent, long short-term memory (LSTM), transformer, and autoencoder-based unsupervised learning, requires further investigation. Some of these architectures are discussed in Chapter 9.

### References:

- [1] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations", *The Journal of Machine Learning Research* vol. 18, no. 187, pp. 1-30, Apr. 2017.
- [2] H.-S. Philip Wong, and S. Salahuddin, "Memory leads the way to better computing, *Nature Nanotechnology*", vol 10, pp. 191-194, Mar. 2015.doi: <https://doi.org/10.1038/nnano.2015.29>
- [3] A. Pedram, S. Richardson, M. Horowitz, S. Galal, and S. Kvatinsky, "Dark Memory and Accelerator-Rich System Optimization in the Dark Silicon Era", *IEEE Design & Test*, vol. 34, no. 2, pp. 39-50, Apr. 2017. DOI: 10.1109/MDAT.2016.2573586
- [4] F. Jiang, K. Wang, L. Dong, C. Pan, W. Xu, and K. Yang, "Deep-Learning-Based Joint Resource Scheduling Algorithms for Hybrid MEC Networks", *IEEE Internet of Things Journal*, vol. 7, no.7, pp. 6252 - 6265, July 2020, DOI: 10.1109/JIOT.2019.2954503.

- [5] F. Jiang, L. Dong, K. Wang, K. Yang, and C. Pan, "Distributed Resource Scheduling for Large-Scale MEC Systems: A Multiagent Ensemble Deep Reinforcement Learning With Imitation Acceleration", *IEEE Internet of Things Journal*, vol. 9, no.9, pp. 6597 - 6610, May 2022, DOI: 10.1109/JIOT.2021.3113872.
- [6] A. Sebastian, M. L. Gallo, R. K.- Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing", *Nature Nanotechnology*, vol. 15, pp. 529–544, Mar. 2020, DOI: <https://doi.org/10.1038/s41565-020-0655-z>
- [7] D. Ielmini, and H.-S. Philip Wong, "In-memory computing with resistive switching devices", *Nature Electronics*, vol. 1, pp. 333–343, Jun. 2018, DOI: <https://doi.org/10.1038/s41928-018-0092-2>
- [8] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves, S. Lam, N. Ge, J. Yang, and R. S. Williams, "Dot-Product Engine for Neuromorphic Computing: Programming 1T1M Crossbar to Accelerate Matrix-Vector Multiplication", in *53rd ACM/EDAC/IEEE Design Automation Conference (DAC)*, Austin, TX, USA, Jun. 2016.
- [9] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves, Z. Li, J. P. Strachan, P. Lin, Z. Wang, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, "Analogue signal and image processing with large memristor crossbars", *Nature Electronics*, vol. 1, pp. 52–59, Dec. 2017, DOI: <https://doi.org/10.1038/s41928-017-0002-z>
- [10] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola, L. L. Sanches, I. Boybat, M. L. Gallo, K. Moon, J. Woo, H. Hwang and Y. Leblebici, "Neuromorphic computing using non-volatile memory", *Advances in Physics: X*, vol. 2, no.1, pp. 89-124, Dec. 2016, DOI: <https://doi.org/10.1080/23746149.2016.1259585>
- [11] G. W. Burr, R. M. Shelby, S. Sidler, C. di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi, and H. Hwang, "Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element", *IEEE Trans. Electron Devices* vol. 62, no. 11, pp. 3498 - 3507, Nov. 2015. DOI: 10.1109/TED.2015.2439635
- [12] M. Prezioso, F. M.- Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors", *Nature*, vol. 521, pp. 61–64, May 2015, DOI: <https://doi.org/10.1038/nature14441>
- [13] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction", in *2011 International Electron Devices Meeting*, pp. 4.4.1-4.4.4, Dec. 2011. DOI: 10.1109/IEDM.2011.6131488

- [14] T. H. Lee, D. Loke, K.-J. Huang, W.-J. Wang, and S. R. Elliott, "Tailoring Transient-Amorphous States: Towards Fast and Power-Efficient Phase-Change Memory and Neuromorphic Computing", *Adv. Mater.*, vol. 26, no. 44, pp. 7493-749, Nov. 2014, DOI: <https://doi.org/10.1002/adma.201402696>
- [15] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, and H.-S. P. Wong, "An Electronic Synapse Device Based on Metal Oxide Resistive Switching Memory for Neuromorphic Computation", *IEEE Trans. Electron Devices*, vol. 58, no.8, pp. 2729 - 2737, Aug. 2011. DOI: 10.1109/TED.2011.2147791
- [16] J. Woo, K. Moon, J. Song, S. Lee, M. Kwak, J. Park, and H. Hwang, "Improved Synaptic Behavior Under Identical Pulses Using AlOx /HfO2 Bilayer RRAM Array for Neuromorphic Systems", *IEEE Electron Device Letters*, vol. 37, no. 8, pp. 994 - 997, Aug. 2016, DOI: 10.1109/LED.2016.2582859
- [17] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, W. Song, J. P. Strachan, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks", *Nature Communications*, vol. 9, pp. 1-8, Jun. 2018, Art. no. 2385, DOI: <https://doi.org/10.1038/s41467-018-04484-2>
- [18] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, Q. Zhang, N. Deng, L. Shi, H.-S. P. Wong, and H. Qian, "Face classification using electronic synapses", *Nature Communications*, vol. 8, pp. 1-8, May 2017, Art. no. 15199. DOI: 10.1038/ncomms15199.
- [19] D. Bhowmik, U. Saxena, A. Dankar, A. Verma, D. Kaushik, S. Chatterjee, and U. Singh, "On-chip learning for domain wall synapse based Fully Connected Neural Network", *Journal of Magnetism and Magnetic Materials*, vol. 498, pp.1-11, Nov. 2019, Art. no. 1654342, DOI: <https://doi.org/10.1016/j.jmmm.2019.165434>.
- [20] A. Sengupta, Y. Shim, and K. Roy, "Proposal for an All-Spin Artificial Neural Network: Emulating Neural and Synaptic Functionalities Through Domain Wall Motion in Ferromagnets", *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 6, pp. 1152 -1160, Dec. 2016, DOI: 10.1109/TBCAS.2016.2525823
- [21] D. Zhang, Y. Hou, L. Zeng, and W. Zhao, "Hardware Acceleration Implementation of Sparse Coding Algorithm With Spintronic Devices", *IEEE Transactions on Nanotechnology*, vol. 10, pp. 518 - 531, May 2019, DOI: 10.1109/TNANO.2019.2916149
- [22] A. F. Vincent, J. Larroque, N. Locatelli, N. B. Romdhane, O. Bichler, C. Gamrat, W. S. Zhao, J.-O. Klein, S. G.-Retailleau, and D. Querlioz, "Spin-Transfer Torque Magnetic Memory as a Stochastic Memristive Synapse for Neuromorphic Systems", *IEEE Transactions on Biomedical Circuits and Systems*, vol. 9, no. 2, pp. 166 - 174, Apr. 2015, DOI: 10.1109/TBCAS.2015.2414423
- [23] M. Alamdar, T. Leonard, C. Cui, B. P. Rimal, L. Xue, O. G. Akinola, T. P. Xiao, J. S. Friedman, C. H. Bennett, M. J. Marinella, and J. A. C. Incorvia, "Domain wall-magnetic tunnel junction spin-orbit

- torque devices and circuits for in-memory computing”, *Appl. Phys. Lett.*, vol. 118, pp. 1-6, Mar. 2021, Art. no. 112401, DOI: <https://doi.org/10.1063/5.0038521>
- [24] M.-C. Chen, A. Sengupta, and K. Roy, “Magnetic Skyrmion as a Spintronic Deep Learning Spiking Neuron Processor”, *IEEE Transactions on Magnetics*, vol. 54, no. 8, pp. 1-7, Aug. 2018, Art. no.1500207, DOI: 10.1109/TMAG.2018.2845890
- [25] D. Kaushik, U. Singh, U. Sahu, I. Sreedevi, and D. Bhowmik, “Comparing domain wall synapse with other non volatile memory devices for on chip learning in analog hardware neural network”, *AIP Advances*, vol. 10, no. 2, pp. 1-7, Feb. 2020, Art. no. 025111. DOI: <https://doi.org/10.1063/1.5128344>
- [26] V. Uhlíř, S. Pizzini, N. Rougemaille, J. Novotný, V. Cros, E. Jiménez, G. Faini, L. Heyne, F. Sirotti, C. Tieg, A. Bendounan, F. Maccherozzi, R. Belkhou, J. Grollier, A. Anane, and J. Vogel, “Current-induced motion and pinning of domain walls in spin-valve nanowires studied by XMCD-PEEM”, *Phys. Rev. B*, vol. 81, no. 22, pp. 1-10, Jun. 2010, Art. no. 224418. DOI: <https://doi.org/10.1103/PhysRevB.81.224418>
- [27] X. Jiang, L. Thomas, R. Moriya, M. Hayashi, B. Bergman, C. Rettner, and S. S.P. Parkin, “Enhanced stochasticity of domain wall motion in magnetic racetracks due to dynamic pinning”, *Nature Communications*, vol. 1, pp. 1-5, Jun. 2010, Art. no: 25. DOI: 10.1038/ncomms1024
- [28] J. P. Attané, D. Ravelosona, A. Marty, Y. Samson, and C. Chappert, “Thermally Activated Depinning of a Narrow Domain Wall from a Single Defect”, *Phys. Rev. Lett.*, vol. 96, no. 14, pp. 1-4, Apr. 2006, Art. no. 147204. DOI: <https://doi.org/10.1103/PhysRevLett.96.147204>.
- [29] W. A. Misba, T. Kaisar, D. Bhattacharya, J. Atulasimha, “Voltage-Controlled Energy-Efficient Domain Wall Synapses With Stochastic Distribution of Quantized Weights in the Presence of Thermal Noise and Edge Roughness”, *IEEE Transactions on Electron Devices*, vol. 69, no. 4, pp. 1658 - 1666, Sep. 2021, doi: 10.1109/TED.2021.3111846
- [30] S. Ikeda, J. Hayakawa, Y. Ashizawa, Y. M. Lee, K. Miura, H. Hasegawa, M. Tsunoda, F. Matsukura, and H. Ohno, “Tunnel magnetoresistance of 604% at 300K by suppression of Ta diffusion in CoFeBMgO/CoFeB pseudo-spin-valves annealed at high temperature”, *Appl. Phys. Lett.* vol. 93, pp.1-3, Aug. 2008, Art. no. 082508, DOI: <https://doi.org/10.1063/1.2976435>
- [31] M. Courbariaux, Y. Bengio, and J.-P. David, “Binaryconnect: Training deep neural networks with binary weights during propagations”, in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Montreal, BC, Canada, Dec. 2015, vol.2, pp. 3123–3131.
- [32] H. Zhang, J. Li, K. Kara, D. Alistarh, J. Liu, C. Zhang, “ZipML: Training Linear Models with End-to-End Low Precision, and a Little Bit of Deep Learning”, in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, NSW, Australia, Aug. 2017, Vol. 70, pp. 4035–4043.

- [33] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, “Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients”. arXiv:1606.06160, Feb. 2018.
- [34] D. Miyashita, E. H. Lee, B. Murmann, “Convolutional Neural Networks using Logarithmic Data Representation”, arXiv:1603.01025, Mar. 2016.
- [35] S. Agarwal, R. B. J. Gedrim, A. H. Hsia, D. R. Hughart, E. J. Fuller, A. A. Talin, C. D. James, S. J. Plimpton, M. J. Marinella, “Achieving Ideal Accuracies in Analog Neuromorphic Computing Using Periodic Carry”, in *2017 Symposium on VLSI Technology*, Kyoto, Japan, Jun. 2017.
- [36] I. Boybat, M. L. Gallo, S. R. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian, and E. Eleftheriou, “Neuromorphic computing with multi-memristive synapse”, *Nature Communications*, vol. 9, pp. 1-12, Jun. 2018, Art. no. 2514. DOI: 10.1038/s41467-018-04933-y.
- [37] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, and G. W. Burr, “Equivalent-accuracy accelerated neural-network training using analogue memory”, *Nature*, vol. 558, pp. 60–67, Jun. 2018. DOI: <https://doi.org/10.1038/s41586-018-0180-5>
- [38] S. R. Nandakumar, M. L. Gallo, C. Piveteau, V. Joshi, G. Mariani, I. Boybat, G. Karunaratne, R. K-Aljameh, U. Egger, A. Petropoulos, T. Antonakopoulos, B. Rajendran, A. Sebastian, and E. Eleftheriou, “Mixed-Precision Deep Learning Based on Computational Memory”, *Frontiers in Neuroscience*, vol. 14, pp. 1-17, May. 2020, Art. no. 406, DOI:10.3389/fnins.2020.00406
- [39] M. L. Gallo, A. Sebastian, R. Mathis, M. Manica, H. Giefers, T. Tuma, C. Bekas, A. Curioni, and E. Eleftheriou, “Mixed-precision in-memory computing”, *Nature Electronics*, vol. 1, pp. 246–253, Apr. 2018, DOI: <https://doi.org/10.1038/s41928-018-0054-8>
- [40] V. Joshi, M. Le Gallo, S. Haefeli, I. Boybat, S. R. Nandakumar, C. Piveteau, M. Dazzi, B. Rajendran, A. Sebastian, and E. Eleftheriou, “Accurate deep neural network inference using computational phase-change memory”, *Nature Communications*, vol. 11, pp. 1-13, May 2020, Art. no: 2473. DOI: <https://doi.org/10.1038/s41467-020-16108-9>
- [41] G. Boquet, E. Macias, A. Morell, J. Serrano, E. Miranda, and J. L. Vicario, “Offline Training for Memristor-based Neural Networks”, in *28th European Signal Processing Conference (EUSIPCO)*, Amsterdam, Netherlands, Jan. 2021, DOI: 10.23919/Eusipco47968.2020.9287574
- [42] L. Chen, J. Li, Y. Chen, Q. Deng, J. Shen, X. Liang, and L. Jiang, “Accelerator-friendly Neural-network Training: Learning Variations and Defects in RRAM Crossbar”, in *Design, Automation & Test in Europe Conference & Exhibition*, Lausanne, Switzerland, Mar. 2017, DOI: 10.23919/DATE.2017.7926952

- [43] B. Liu, H. Li, Y. Chen, X. Li, Q. Wu, and T. Huang, “Vortex: Variation-aware training for memristor X-bar”, in 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, Jun. 2015, DOI: 10.1145/2744769.2744930.
- [44] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, “Fully hardware-implemented memristor convolutional neural network”, *Nature*, vol. 577, pp. 641-646, Jan. 2020, DOI: <https://doi.org/10.1038/s41586-020-1942-4>
- [45] E. Martinez, L. L.-Diaz, L. Torres, C. Tristan, and O. Alejos, “Thermal effects in domain wall motion: Micromagnetic simulations and analytical mode”, *Phys. Rev. B*, vol. 75, pp. 1-11, May. 2007, Art. no. 174409, DOI: <https://doi.org/10.1103/PhysRevB.75.174409>
- [46] S. Dutta, S. A. Siddiqui, J. A. C.-Incorvia, C. A. Ross, and M. A. Baldo, Micromagnetic modeling of domain wall motion in sub-100-nm-wide wires with individual and periodic edge defects, *AIP ADVANCES*, vol. 5, pp. 1-9, Aug. 2015, Art. no. 127206, DOI: <https://doi.org/10.1063/1.4937557>
- [47] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. V. Waeyenberge, “The design and verification of MuMax3”, *AIP Advances*, vol. 4, no. 10, pp. 1-22, Oct. 2014, Art. no. 107133. DOI: <https://doi.org/10.1063/1.4899186>
- [48] S. Liu, T. Xiao, C. Cui, J.-A. C. Incorvia, C. H. Bennett, and M. J. Marinella, A domain wall-magnetic tunnel junction artificial synapse with notched geometry for accurate and efficient training of deep neural networks, *Appl. Phys. Lett.* vol.118, Art. no. 202405, pp. 1-7, May 2021, doi: 10.1063/5.0046032
- [49] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition”, *Proceedings of the IEEE*, vol. 86, no.11, pp-2278 - 2324, Nov. 1998, DOI: 10.1109/5.726791
- [50] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors”, *Nature*, vol. 323, pp. 533–536, Oct. 1986, DOI: <https://doi.org/10.1038/323533a0>
- [51] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, “Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference”, arXiv:1712.05877, Dec. 2017.
- [52] T. Hirtzlin, M. Bocquet, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, D. Querlioz, “Outstanding Bit Error Tolerance of Resistive RAM-Based Binarized Neural Networks”, arXiv:1904.03652, Apr. 2019.
- [53] B. Liu, H. Li, Y. Chen, X. Li, T. Huang, Q. Wu, and M. Barnell, “Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems”, in 2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), Jan. 2015. DOI: 10.1109/ICCAD.2014.7001330

- [54] T. Hirtzlin, M. Bocquet, B. Penkovsky, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, and D. Querlioz, “Digital Biologically Plausible Implementation of Binarized Neural Networks With Differential Hafnium Oxide Resistive Memory Arrays”, *Frontiers in Neuroscience*, Vol. 13, pp.1-14, Jan. 2020, Art. no. 1383, doi: <https://doi.org/10.3389/fnins.2019.01383>
- [55] J. Cui, J. L. Hockel, P. K. Nordeen, D. M. Pisani, C.-Y. Liang, G. P. Carman, and C. S. Lynch, “A method to control magnetism in individual strain-mediated magnetoelectric islands”, *Appl. Phys. Lett.*, vol. 103, no. 23, pp. 1-5, Dec. 2013, Art. no. 232905. DOI: <https://doi.org/10.1063/1.4838216>
- [56] T. Dalgaty, N. Castellani, C. Turck, K.-E. Harabi, D. Querlioz and E. Vianello, “In situ learning using intrinsic memristor variability via Markov chain Monte Carlo sampling”, *Nature Electronics*, Vol. 4 ,pp. 151-161, Jan. 2021, doi: <https://doi.org/10.1038/s41928-020-00523-3>
- [57] M. S. Alam, B. R. Fernando, Y. Jaoudi, C. Yakopcic, R. Hasan, T. M. Taha, G. Subramanyam, “Memristor Based Autoencoder for Unsupervised Real-Time Network Intrusion and Anomaly Detection”, in *Proceedings of the International Conference on Neuromorphic Systems*, pp. 1-8, Art. no. 2, Jul. 2019, doi: <https://doi.org/10.1145/3354265.3354267>.
- [58] J. M. Slaughter, R.W. Dave, M. Durlam, G. Kerszykowski, K. Smith, K. Nagel, B. Feil, J. Calder, M. DeHerrera, B. Garni, and S. Tehrani, “High Speed Toggle MRAM with MgO-Based Tunnel Junctions”, in *IEEE International Electron Devices Meeting, 2005. IEDM Technical Digest, Dec. 2005*, DOI: 10.1109/IEDM.2005.1609496
- [59] N. Gupta, H. Shrimali, A. Makosiej, A. Vladimirescu, and A. Amara, “Energy Efficient Comparator-Less Current-Mode TFET-CMOS Co-Integrated Scalable Flash ADC”, *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug. 2021, DOI: 10.1109/MWSCAS47672.2021.9531911
- [60] N. D’Souza, A. Biswas, H. Ahmad, M. S. Fashami, M. M. A. Rashid, V. Sampath, D. Bhattacharya, M. A. Abeed, J. Atulasimha and S. Bandyopadhyay, Energy-efficient switching of nanomagnets for computing: straintronics and other methodologies, *Nanotechnology*, Vol. 29, no. 44, pp. 1-49, Aug. 2018. doi: <https://doi.org/10.1088/1361-6528/aad65d>
- [61] M. J. Marinella, S. Agarwal, A. Hsia, I. Richter, R. J.-Gedrim, J. Niroula, S. J. Plimpton, E. Ipek, and C. D. James, “Multiscale Co-Design Analysis of Energy, Latency, Area, and Accuracy of a ReRAM Analog Neural Training Accelerator”, *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 1, Mar. 2018, pp. 86 - 101, DOI: 10.1109/JETCAS.2018.2796379

## **Chapter 5: Spintronic Physical Reservoir for Autonomous Prediction and Long-Term Household Energy Load Forecasting**

In the previous Chapter 4, we explored the implementation of highly accurate DNNs for classification tasks using domain wall synapse-based racetrack memory devices. In this chapter, we examine another intriguing chiral magnetic texture, such as skyrmions, as a physical reservoir for autonomous time series forecasting tasks—typically performed by recurrent neural networks (RNNs) and long short-term memory (LSTM) networks.

With the growing use of artificial neural networks (ANNs) in temporal data processing tasks, the cost of training for complex ANNs is an escalating concern. Physical reservoir computing (RC), a variation of RNNs, obviates the need for most data intensive matrix vector multiplication in the recurrent layer by evolving the RC's internal states with the inherent nonlinear dynamics and short-term memory. In this chapter, we show that magnetic skyrmion confined in a fixed geometry forming the soft layer of a magnetic tunnel junction (MTJ) can work as a RC and perform autonomous long-term prediction of temporal data. Our proposed skyrmion reservoir allows for manipulation of spin dynamics with ultra energy efficient voltage controlled magnetic anisotropy modulation (VCMA) method. Furthermore, the boundary effect on the skyrmion from the geometric edges provides necessary consistency property of the reservoir. We employ our proposed reservoir for the modeling and prediction of the chaotic time series such as Mackey-Glass and dynamic time-series data, such as household building energy loads. For autonomous run, the predicted output is fed to the input of the reservoir. By comparing our spintronic physical RC approach with energy load forecasting algorithms, such as LSTMs and RNNs, we conclude that the proposed framework presents good performance in achieving high predictions accuracy, while also requiring low memory and energy both of which are at a premium in hardware resource and power constrained edge applications. Further, the proposed approach is shown to require very small training datasets and at the same time being at least  $16\times$  energy efficient compared to the sequence-to-sequence LSTM for accurate household load predictions. Higher endurance, fast processing and well-established technology (i.e., MTJ) to integrate with CMOS technology makes such spintronic reservoir attractive over other emergent memory device-based reservoirs.

Recurrent neural networks (RNNs) [1,2] are shown to be more suitable in temporal data processing tasks than the traditional feedforward neural networks (FNNs) because of the recurrent connections among constituent neurons. However, RNNs often suffer from vanishing and exploding gradients problem due to



the long-term dependencies that could arise in the recurrent layers. To circumvent these issues variations of RNN is proposed, i.e., long short-term memory (LSTM) [3] and reservoir computing (RC) [4,5]. In RC, the reservoir consists of an RNN which maps temporal inputs to higher dimensional features due to the short-term memory property that exists in the reservoir and a read-out layer that analyzes the features stored as reservoir states. The RNN connections are fixed and only the read-out layer is trained [4]. Thus, the training can be performed with simple learning rules such as linear regression which makes RC much simpler to implement with low training cost. Recently software-based RC systems have been shown to achieve state-of-the-art performance in speech recognition tasks [6] and superior performances in forecasting tasks, such as prediction of financial systems [7], water inflow [8], and chaotic system prediction [9].

Since the essence of RC is to employ non-linearity to transform input to high dimensional space, any physical dynamic non-linear system can work as a reservoir. For a typical RNN implemented on hardware, the required training is performed for all layers of the neural network [10]. This can be implemented on neuromorphic chips [11]. However, in RC the inference is performed using physical phenomena, and only linear regression is used to train weights between select physical reservoir states and the output. This makes information processing much faster and involves low training cost. These features make physical systems the preferred candidate for hardware implementations of RC. Towards this end, the choice of physical reservoir remains explorative, and researchers investigated electronic [12, 13], photonic [14, 15], memristive [16, 17], spintronic [18-22] reservoir and so on. The spintronic reservoir is most attractive due to its significantly higher endurance cycle and faster information processing. For instance, spintronic computational memory has an endurance over  $\sim 10^{15}$  cycle and write speed of  $\sim 1-10$  ns [23]. Whereas the phase-change memory based memristor device exhibits an endurance of  $\sim 10^9$  cycle and has a write speed  $\sim 100$  ns [24]. Also, the read (i.e., magnetoresistance) and write techniques (i.e., spin torques) of spintronic devices and associated integration with CMOS technologies are well established due to their historical use as magnetic hard drives, sensors and magnetic random-access memory devices [25]. In spintronic reservoir only one small-scale non-linear node (i.e., nanomagnets) can essentially capture non-linear dynamics. In comparison, electronic reservoirs (i.e., Mackey-Glass nonlinear circuit element with delayed feedback) [12], use a number of active (such as op-amps) and passive elements (such as resistors) for such capability which could cause prohibitive energy and memory footprint. While the photonic reservoir allows for faster processing, compact design with short time delays requires extremely fast input and output processing, a significant drawback for practical implementation [5]. Moreover, short range exchange interaction and long-range dipolar interaction in spintronic systems provides the opportunity to couple magnetic nodes

without any physical interconnections and allow more complex coupling otherwise absent in above-mentioned alternative reservoirs.

Various devices concepts are proposed for spintronic reservoir including nanoscale magnetic structures such as spin torque nano-oscillators [18], planner ensemble of nanomagnets interacting in dipole coupled [26,27] and spin wave mediated [20,28] systems and artificial spin ice (ASI) [29]. Chiral magnetic textures such as domain walls (DWs) [21,30], skyrmions [22,31] and skyrmion lattice [19,32] are also proposed. Individually accessing the dense arrays of ASI and planar nanomagnets [26] with addressable magnetic tunnel junction (MTJ) remains a fabrication challenge with modern technology. Furthermore, recent experiments with ferromagnetic resonance with ASI [29] and interconnected ring arrays with magnetic DWs [30] require external magnetic field for dynamic interaction, which is energy prohibitive. Moreover, low loss propagation of spin wave requires higher quality crystal growth (Yttrium Iron Garnet) and fabrication of nano-antennas to excite and detect spin waves which are prone to Ohmic losses [33]. While the spin-torque nano-oscillator based MTJ [18] shows excellent performance with a greater promise of low area footprint and energy cost, the fast-oscillatory signal cannot be used directly for postprocessing (microwave diode is used to capture amplitude variation).

In contrast, chiral spin texture, skyrmion, confined in a fixed geometry working as a RC allows for external magnetic field free control and direct processing of the generated magnetoresistance signal. Moreover, skyrmion shows higher mobility and has low pinning potential than other chiral structures such as DWs, thus can be excited and translated with extremely low amplitude excitation [19]. In addition, confined skyrmion textures stacked within a magnetic tunnel junction (MTJ) device allows for ultra-energy efficient voltage controlled magnetic anisotropy (VCMA) modulation [34-39]. Furthermore, confinement of skyrmion provides necessary repulsion from boundary which ensures an essential consistency property to the reservoir [40], where the skyrmion needs to be relaxed to the same energy minima upon withdrawing the input excitation.

Although physical RCs have been shown to implement prediction tasks, most of the works attempt to perform one-step ahead prediction. Long term prediction is important for real-world data as future evolution can facilitate more informed decisions and customized policy making. However, multi-step prediction itself is challenging due to the non-linear nature of most real-world data and typically inaccurate predictions of the immediate future accumulate very fast and cause divergence in the future predictions over longer times. Authors in [41] shown multi-step prediction for high spatial dimension data using spatiotemporal transformation and encoder-decoder like reservoir. However, this approach can fall short for univariate data such as individual household power consumption prediction. In household load forecasting tasks, usually the real time power value is readily available from wattmeter, however, the voltage, current values and

other parameters are unknown. In [42], autonomous multi-step prediction has been shown for chaotic Mackey-Glass (MG) series by feeding delayed output to the reservoir. However, the reservoir response is transformed using a non-linear function which could be costly, and also diminishes the benefits obtained from linear reservoir operation. Recently, delayed inputs [43] and polynomial transformation of the delayed inputs [44] are used to improve the prediction performance of the reservoir, however, these works attempt to solve the optimized one-step ahead prediction. In this study, we have shown multi-step autonomous prediction using spintronic magnetic skyrmion based RC system. For autonomous prediction, the predicted output is fed directly to the input. To adequately extract the non-linearity that arises in the reservoir dynamics, we include several previous states of the reservoir during the training. This obviates the need to perform any non-linear (i.e. trigonometric, polynomial) transformation of the reservoir states.

For long-term prediction our skyrmion based RC employs the virtual node concept originally proposed in [45] and shown in Fig. 5-1. Instead of fabricating a large number of reservoir nodes to increase the reservoir dimensionality, a single physical nonlinear node subjected to delayed feedback acts as a chain of virtual nodes. First, we have shown long term prediction for chaotic MG time series since it has been frequently used for benchmarking forecasting tasks. Moreover, the long-term prediction can diverge more quickly due to the sensitivity of the chaotic systems to the error. Next, we have shown individual household power demand forecasting, which is an active research area. According to a recent study, the amount of energy wasted in a commercial building can reach up to 40% if energy consumption is not properly maintained [46]. With energy management system (EMS), a commercial building can save up to 25.6% of its total energy consumption [47]. EMS in a building requires accurate load forecasting to maintain stability, improve performance, and detect abnormal system behavior. However, accurate long-term forecasting, especially in a single household building, is very challenging due to the volatile and univariate nature of the household power consumption data. Traditional statistical approaches such as auto-regressive moving average (ARIMA) [48], time-series statistical model [49] suffer from low prediction accuracy due to the parameters assumed in the model and the complexity of the systems. Machine learning based models such as RNN [50] and LSTM [51] offer more flexibility in this regard as they do not depend on the parameters of the system. Instead, they are driven by the observed past and present data, however, with significant training costs. RC can perform the prediction tasks with much more efficiency due to its low training cost and thus is very suitable to be implemented in edge computing platforms which are equipped with low power devices.

We use three decoupled and patterned skyrmion devices as our reservoir where the temporal correlation of the inputs is captured by the inherent short-term memory of the breathing skyrmions. Upon excitation, the skyrmions undergoes oscillations and the skyrmion states are read at a regular interval and processed with

linear regression for the prediction. Autonomous prediction up to 30-time steps for the MG time series and up to 23 hours (equivalent to 23-time steps as only hourly demand is typically required) for the household power prediction have been demonstrated by using the predicted output as the input for the next time step prediction.

The rest of the chapter is organized as follows. In the section 5.1, we detail the architecture of the skyrmion reservoir and corresponding magnetization dynamics, in section 5.2 we describe the process to set up and train and test the reservoir and in section 5.3 we discuss the results and summarize our findings in the conclusion (section 5.4).

## 5.1. Model:

### Proposed Physical Reservoir:

Fig. 5-1a shows a conventional reservoir computing system which consists of an input layer, a reservoir block having recurrent connections among the constituent nodes and an output layer. The solid line arrows show the connections which are fixed and the dashed arrows are the connections which need to be trained. We propose to replace the reservoir block with three patterned and decoupled skyrmions whose magnetization dynamics we simulate. Each of the individual skyrmions is hosted in the ferromagnetic thin films with perpendicular magnetic anisotropic (PMA) as shown in Fig. 5-1b. A ferromagnetic reference layer, a tunnel barrier (MgO) and a synthetic antiferromagnetic (SAF) layer are patterned on top of the ferromagnetic free layer (that hosts the skyrmion) to create the MTJ as shown in Fig. 5-1c. This facilitates the read and write operation. The temporal inputs are linearly mapped into a voltage pulse and applied across the MTJs to modulate the PMA using the voltage-controlled magnetic anisotropy (VCMA) effect [52-54]. All the patterned skyrmions are subjected to the same set of inputs. When the PMA is modulated within a certain range, the skyrmions generate oscillatory response (skyrmion breathing) as shown in Fig. 5-2. The responses of the skyrmions are read with MTJs and processed with linear regression to compute the predicted values of the temporal time series.

For a typical reservoir consists of N number of nodes as seen in Fig. 5-1a, the time discretized states of the nodes,  $r_i^n$ , can be represented as follows:

$$r_i^{n+1} = f\left(\sum_{j=0}^{N-1} p_{ij}r_j^n + q_i u^n\right) \quad (1)$$

Here, the  $p_{ij}$ ,  $q_i$  are time-independent coefficients that are drawn from a random distribution having a mean of 0 and the standard deviations are adjusted for optimal performances. Also,  $f$  is the activation function which can be linear or non-linear and  $u^n$  is the input. Here, the “fading memory” or the short-term memory (an essential property of the reservoir) is achieved by using large number of nodes and their recurrent connections. In comparison, the skyrmion systems have inherent memory effect in their responses, thus instead of using several nodes only one skyrmion device can work as a reservoir. However, to increase the dimensionality, the states of this reservoir can be read at regular interval for a particular input, which acts as the virtual nodes of the reservoir. The concept is originally developed in reservoir with delayed feedback where a nonlinear node subjected to an input and delayed feedback acts as a chain of virtual nodes and provides performance similar to a typical reservoir [45]. Later, it has been shown the virtual nodes derived from reservoir responses subjected to only the input signal can provide optimal performance [17,18]. In such a scenario, if the virtual node interval (as shown by  $\theta$  in Fig. 5-1d) is lower than the characteristic time (relaxation time) of the reservoir, the node states are not only influenced by their own previous states, but also the neighboring node states and the input excitation. This allows for non-linear coupling among the nodes.

Thus, the interconnection matrix in Eq. 1 is simplified in the skyrmion reservoir case where the virtual nodes are assumed to be connected in ring topology (as seen in Fig. 5-1d). The resulting node states of the reservoir can be expressed as:

$$\begin{aligned} r_0^{n+1} &= f(r_0^n + r_{N-1}^{n-1} + u^n) \\ r_i^{n+1} &= f(r_i^n + r_{i-1}^n + u^n) \end{aligned} \quad (2)$$

Here, we have used linear activation function,  $f(w)=w$ , thus the read-out reservoir states are used for training without any post-processing or non-linear transformation.

The node states of the reservoir act as features, which are used to generate the output. Since we are using only linear activation, we also include several previous responses of the reservoir for generating the output. Due to the short term-memory effect inherent to the skyrmion dynamics, these previous states provide additional non-linear effects. The output of the reservoir can be expressed as follows:

$$y^n = \sum_{j=n-d}^n \sum_{i=0}^{N-1} w_{ij} r_i^j \quad (3)$$

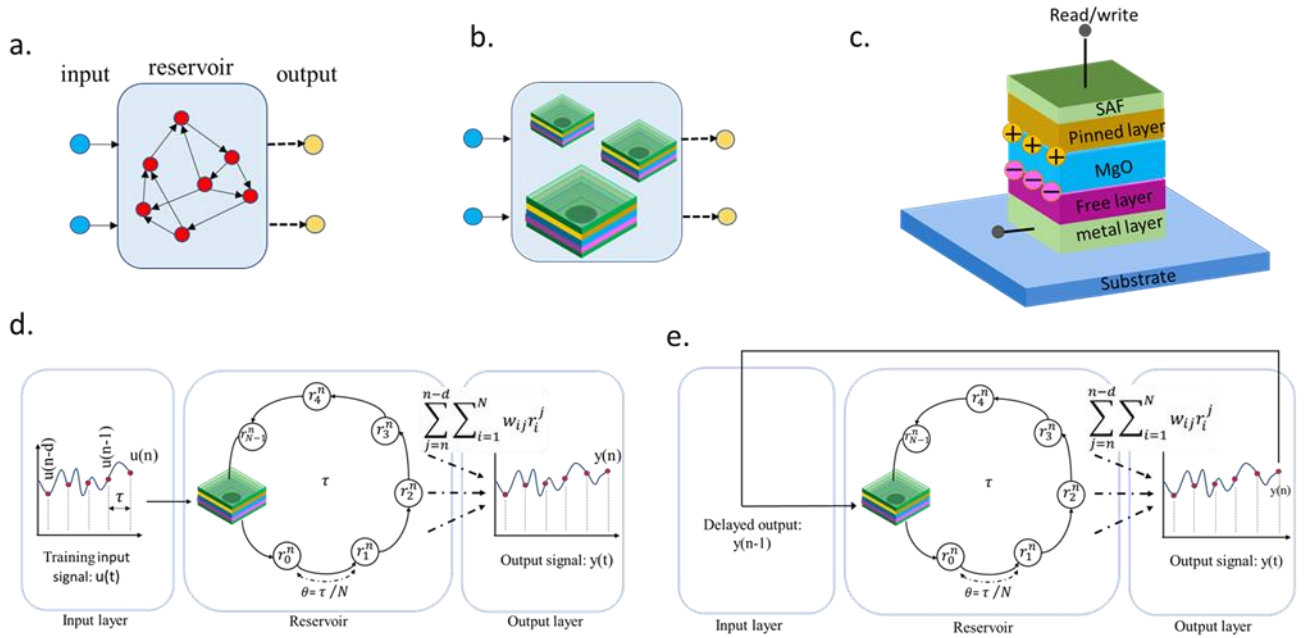


Figure 5-1 a. A conventional reservoir computing system with input layer, reservoir block with recurrent connections among nodes and the output layer. b. The reservoir block is replaced by a set of patterned skyrmion devices where each of the ferromagnetic films with PMA host a single skyrmion. c. Stacks of a skyrmion device with metallic electrode and MTJ. d. Training of a skyrmion reservoir for autonomous prediction task. The temporal input data is mapped into voltage values which are applied to each of the skyrmion devices and the responses are collected. The responses are read at regular interval and the read-out values act as the virtual node as represented by  $r_i^j$ . The states of the nodes (or reservoir responses) are used to predict the next time step value of the input time series. The weights are trained by computing the error of the predicted and target values and accomplished with simple pseudoinverse operation. e. During testing, the predicted output value is directly fed as input to the reservoir in order to perform multi-step autonomous prediction.

Where,  $d$  represents the number of time-steps for which the previous reservoir responses are included. The optimal weights,  $w_{ij}$  can be obtained by optimizing a cost function. We use mean squared error as our cost functions:

$$c = \langle (y^n - t^n)^2 \rangle \quad (4)$$

Where  $t^n$  represents the teacher or target output of the system at  $n^{\text{th}}$  time step. The optimization can be performed off-line using linear regression with regularization (ridge regression) or on-line using gradient descent optimizer.

During the training or weight optimization stage, the teacher input,  $I^n = u^n$  is applied to the reservoir to predict the next time step value  $y^n = u^{n+1}$  as seen from Fig. 5-1d. Once the optimized weights are obtained, the testing phase begins, where the inputs are disconnected and the output of the reservoir is connected directly to the input,  $I^n = y^{n-1}$  as can be seen in Fig. 5-1e.

### Magnetization Dynamics:

Magnetization dynamics was simulated by solving the Landu-Lifshitz-Gilbert (LLG) equation in micromagnetic framework MUMAX3 discussed in detail in section 1.2.7.

The magnetic films are discretized into cells with dimensions of  $2 \text{ nm} \times 2 \text{ nm} \times 1 \text{ nm}$ , which are much shorter than the exchange length ( $\sqrt{\frac{2A_{ex}}{\mu_0 M_S^2}}$ ). The simulations have been carried out without the thermal noise (T=0 K), however, from our previous study it has been shown that the short-term memory property of the skyrmion reservoir does not degrade much in the presence of room temperature thermal noise [31]. The simulation parameters are listed in table 5-1.

Table 5-1: Simulation parameter

Parameters	Values
DMI constant (D)	$0.0006 \text{ Jm}^{-2}$
Gilbert damping ( $\alpha$ )	0.015
Saturation magnetization ( $M_S$ )	$10^6 \text{ Am}^{-1}$
Exchange constant ( $A_{ex}$ )	$2 \times 10^{-11} \text{ Jm}^{-1}$
Perpendicular Magnetic Anisotropy ( $K_u$ )	$7.5 \times 10^5 \text{ Jm}^{-3}$

### Dataset:

We evaluated the long-term prediction performance of our proposed reservoir on two different time series forecasting datasets.

#### 1) Mackey-Glass Time Series

The MG time series can be expressed as non-linear time-delay differential equation as follows:

$$\frac{dx}{dt} = \alpha \frac{x(t - \tau)}{1 + (x(t - \tau))^n} - \beta x(t) \quad (9)$$

where  $x(t)$  is the MG time series value at time  $t$ , and  $\alpha=0.2$ ,  $\beta=0.1$ ,  $n=10$ ,  $\tau=17$  and  $x(0)=1$ . The equation is solved using the Runge-Kutta method with integration time step,  $dt=0.1$ . The time series is down sampled to 10 and normalized between  $[-1,1]$ . Despite the deterministic form, forecasting this chaotic series is challenging, thus the system is used for benchmark forecasting tasks in literature [42-44]. 431-time steps data are considered where the first 30-time steps data (1-30) are discarded for the training, next 370-time steps data (31-400) are trained and the next 30-time steps data (402-431) are predicted with autonomous prediction.

## 2) Individual Household Power Consumption

The other dataset we worked on is a benchmark dataset of electricity consumption for a single residential customer, named “Individual household electric power consumption” [61]. The data set contained power consumption measurements gathered between December 2006 and November 2010 with a one-minute resolution. The dataset contained aggregate active power load for the whole house and three sub-metering for three sections of the house. In this paper, only the aggregate active load values for the whole house are used. The dataset contained 2075259 measurements. The hourly resolution data were obtained by averaging the one-minute resolution data. Multistep autonomous prediction is especially challenging for these types of datasets due to the stochasticity in the data that arises from erratic human behavior or seasonal change. 284 hours of data are considered where, the first 20 hours data (1-20) are not trained, the next 220 hours of data (21-260) are used for training and the final 23 hours of data (262-284) are predicted with autonomous prediction.

## 5.2. Reservoir Setup

### Skymion Reservoir:

Three patterned ferromagnetic thin films with square geometry having the side lengths of 1000 nm, 800 nm and 700 nm are considered as the reservoir for MG time series prediction tasks as shown in Fig. 5-2. 15 nm thick slices are etched from all sides in the middle region, which leaves a 500 nm square block hosting the skyrmions. The etched block is used to prevent the propagation of spin waves, which is shown to negatively impact the short-term memory capacity of the skyrmion reservoir [31]. Moreover, the etched region will provide a boundary so that the skyrmion cannot be annihilated easily. In addition, different length scales of the periphery will provide different strength of dipole coupling to the skyrmions and create variability in the reservoir response thus enhancing the robustness of the reservoir. We note, input multiplexing is used in previous studies to incorporate diverse response of a single reservoir node. However, we do not use input



multiplexing and the variation of responses is incorporated by using different geometry nodes. The input time series values are transformed into voltage pulses with linear conversion between input magnitude and voltage applied. This voltage pulse translates to change in the perpendicular magnetic anisotropy using the VCMA coefficient,  $\varepsilon = \frac{\Delta K_{si}}{\Delta V/t_{MgO}}$  (described later in section 5.4) in the skyrmion devices. For modeling the response, we use voltage pulses which are applied sequentially with a 2 ns duration. After applying each input pulse, the system is relaxed for 16 ns. We note that instead of applying the pulse for 18 ns we opt for a shorter write pulse which not only saves energy but also triggers rich dynamics that occurs during the relaxation phase of the skyrmion device. Moreover, the relaxation offers flexibility in terms of post processing time required for multi-step autonomous prediction (current prediction is provided as input for the next prediction). Fig. 5-2 shows the magnetization responses of the different reservoirs during the training phase of the MG time series at time step 231-235. The reservoir responses for a single period (18 ns) is read at 3 ns interval (6 times). These 6 values act as the virtual nodes of the reservoir (see in Fig. 5-2, the red diamond marks). Nodes more than 6 can be selected; however, this does not improve the performance as 6 nodes can adequately capture the amount of information in one period. Next, instead of using only the states in the current period (as been done in our previous study for one step ahead prediction [27,31]), we also include reservoir states from the previous 30 periods of data (total 31 periods, 31\*6=186 states for one skyrmion device) for the autonomous long-term prediction of the MG series. The short-term memory capacity of the patterned skyrmion is shown to be  $\sim 4$  bits [31] (up to 4 periods). Thus, the reservoir is expected to remember inputs from the previous 4 periods and non-linearly transform its states based on the memorized inputs. However, when tasked with autonomous prediction using only the current states, the prediction quickly diverges (large prediction errors after third time-step). Thus, for multi-step prediction, previous responses are included to enable the RC to utilize important contexts from the past few observations. Once the reservoir states are obtained from all three skyrmion devices, ridge regression (Tikhonov regularization) is performed for training and computing the optimal weights. The mean squared error is considered as the cost function (shown in Eq. 4) and the activation functions for the reservoir nodes are considered to be linear. With these assumptions, the optimal weights can be found using the following:

$$w_{ij}^{opt} = (AA^T + \lambda I)^{-1} A^T B \quad (10)$$

where,  $A$  is the reservoir response vector including the present and past observation for all the training inputs,  $\lambda$  is the regularization coefficient,  $B$  is the vector of labels containing the target output. For MG series forecasting we choose the regularization coefficient to be,  $\lambda = 10^{-8}$ .

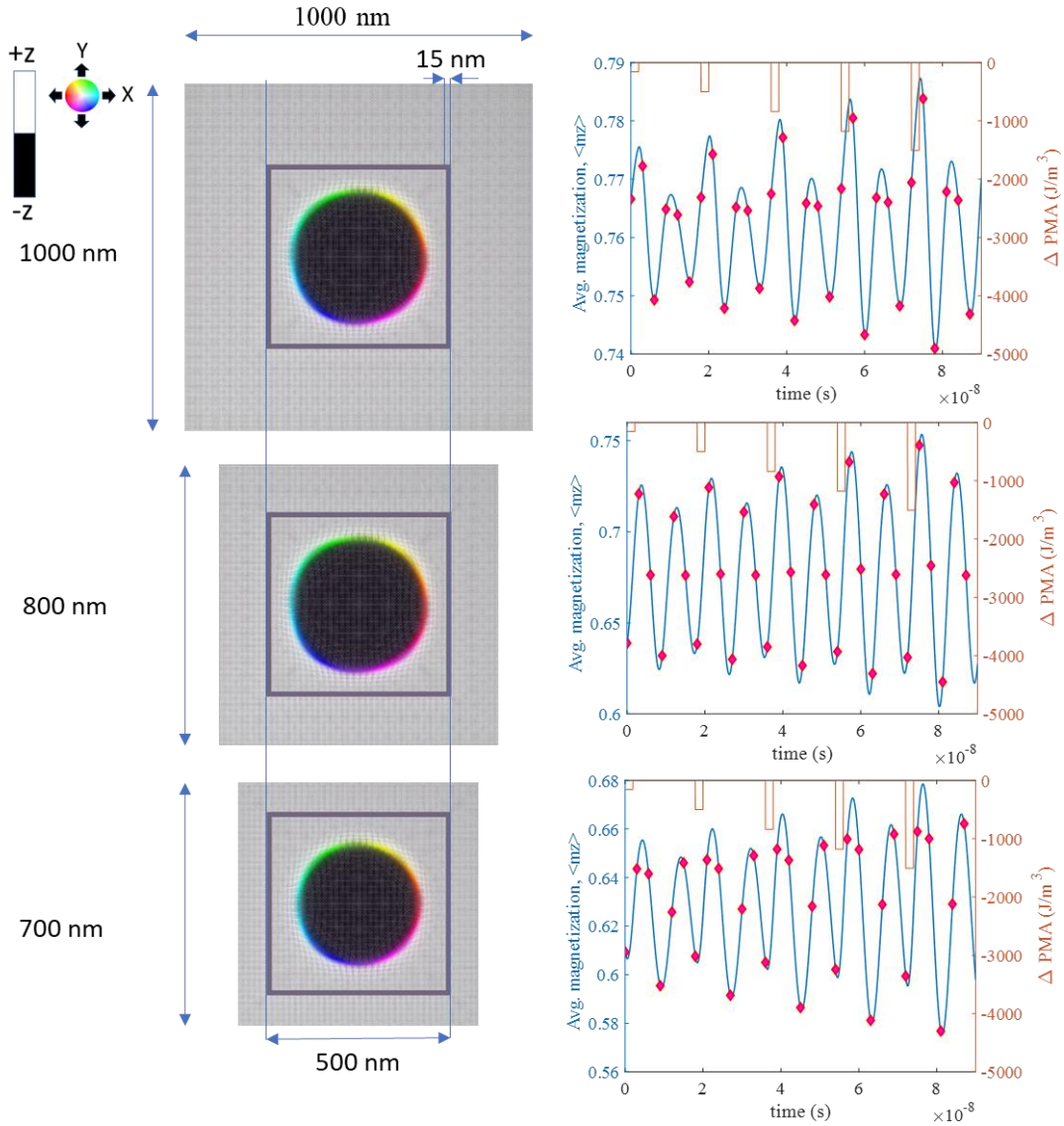


Figure 5-2 a. Three ferromagnetic thin films each hosting a magnetic skymion worked as the RC. The responses of the respective skymion devices are shown side by side when the thin films are perturbed by the inputs of MG time series from time-step 231 to 235 (inputs are mapped into voltage pulse amplitudes). The PMA modulation by the input voltage pulses is shown in orange color. The virtual nodes are marked in red diamond.

For household power prediction tasks, slightly modified geometries are used. The ferromagnetic thin films of 1050 nm, 850 nm and 750 nm are used. The additional side lengths of the ferromagnetic regions provide stability to the skymions to the stochastic changes presented in the household load data. The etched region of 15 nm and 500 nm middle regions for hosting the skymions remain the same. The voltage pulse duration of 2 ns and relaxation period of 16 ns are kept the same. Here, the reservoir states for a total of 21 periods (current period, and previous 20 periods) are used for forecasting tasks. The optimal weights are computed using ridge regression where the regularization parameter is chosen to be,  $\lambda = 10^{-1}$ . Due to the nonvolatile

and stochastic nature of the household load data, it is difficult to train the reservoir as accurately as possible without overfitting, thus a large regularization coefficient is required.

### **Training and Testing the Reservoir:**

At first, for each input, the reservoir states are read at 3 ns interval up to 18 ns. The state vector can be expressed as  $R_n = \{r_0^n, r_1^n, \dots, r_5^n\}$ , where the superscript  $n$  in  $r$  represents the  $n^{\text{th}}$  input of the temporal series and the subscript represents the virtual node number. For MG series prediction task, all 400 training inputs are applied sequentially to all of the skyrmion devices and the corresponding  $R_n$  are collected. The label,  $t^n$  for the prediction task is the next time-step MG function value,  $t^n = u^{n+1}$ , where  $u^n$  is the MG function value at  $n^{\text{th}}$  time step. The training could be performed using the reservoir response vector,  $A = [(R_1, \dots, R_d, R_{d+1})', \dots, (R_{n-d-1}, \dots, R_{n-2}, R_{n-1})', (R_{n-d}, \dots, R_{n-1}, R_n)']$  and the corresponding target output vector,  $B = [t^1, \dots, t^{n-1}, t^n]$  and using the ridge regression equation in Eq. 6. The “'” symbol denotes the transpose operation. Once the optimal weights are obtained after training the reservoir is ready for testing and performing autonomous prediction.

In an actual hardware implementation, the weight optimization (such as pseudoinverse) operation can take time. By the time the optimal weights are computed, the reservoir can be sufficiently relaxed and loses its memory. Thus, before starting the testing phase, the reservoir needs to warm up [17,42]. The same temporal series used in training can be used for warm-up or initializing the reservoir. This adds computational overhead to the model. However, during such initialization step the reservoir is only excited with actual training data and reading the states of the reservoir is not required, which saves read energy cost. The overhead of the reservoir initialization can be avoided by allocating some of the training data for initialization. Depending on the postprocessing optimization time, instead of using all the training data for weight optimization, some of the training data can be saved for reservoir initialization. Alternatively, the optimal weights can be obtained using simple gradient descent method, which can be performed at the same time the training data are supplied to the physical reservoir. This is demonstrated by authors in their optoelectronic reservoir implementation in FPGA [62]. Furthermore, the time complexity of matrix-vector multiplication (the most computationally expensive load of stochastic gradient descent optimization) could be further reduced to single time step using non-volatile computational memory device arranged in a crossbar. In the MG series prediction task, during the warm-up, the inputs are applied sequentially up to 401<sup>th</sup> input. Once the reservoir responses are collected, we predict the 402<sup>th</sup> time step value of the series and then this predicted output is fed directly to the reservoir as input as shown in Fig. 5-1e. We repeatedly performed these steps to autonomously predict the time series up to 431<sup>th</sup> time step. For autonomously predicting household power consumption, the reservoir is trained and initialized by providing input up to 261<sup>th</sup> hour data. Then, the reservoir responses are collected and using the optimal weights, the 262<sup>th</sup> hour

data is predicted. The predicted output is then directly fed as the input. Autonomous prediction up to 284<sup>th</sup> hour is performed following these steps.

### Sequence-To-Sequence LSTM Architecture:

The performance of the reservoir is also compared with state-of-the-art sequence-to-sequence (S2S) LSTM. S2S is an architecture that was proposed to map sequences of different lengths [63]. The architecture consists of two LSTM networks: an encoder and a decoder. As the input state for the decoder, the encoder's job is to transform input sequences of variable length into fixed-length vectors. Afterward, the decoder produces an output sequence with length  $n$ . In this case, the output is the energy load projection for the following  $n$  steps is the output sequence. This architecture's key benefit is that it accepts inputs of any length. In other words, the load for an arbitrary number of future time steps can be predicted using any number of available load measurements from previous time steps as inputs. The electricity load (active power) for a time step or multiple time steps in the future, given historical electricity load data, i.e.,  $M$  load measurements available, can be expressed as:

$$y = \{y[0], y[1], \dots, y[M-1]\} \quad (11)$$

Where,  $y[t]$  is the actual load measurement for time step  $t$ , the load for the following  $T - M$  time steps should be predicted. The predicted load values can be expressed as:

$$\hat{y} = \{\hat{y}[M], \hat{y}[M+1], \dots, \hat{y}[T]\} \quad (12)$$

For training, the encoder network is pre-trained to minimize the following error:

$$LE = \sum_{i=1}^M (y[i] - \hat{y}[i])^2 \quad (13)$$

Then the encoder is plugged into the decoder network and we train the two networks to reduce the objective function:

$$LD = \sum_{i=M+1}^T (y[i] - \hat{y}[i])^2 \quad (14)$$

The error of network is minimized using the backpropagation algorithm. Back-propagation signals are allowed to flow from the decoder to the encoder. Therefore, weights for both the encoder and decoder are updated in order to minimize the objective function expressed in Eq. 14. Both decoder and encoder are updated because the pre-training of the encoder alone is insufficient to achieve good performance. In this

paper, we tested multiple layers with different numbers of neuron units per layer and we found that the training dataset using a 2-layer network with 50 units in each layer gave the best performance. Increasing the capacity of the network did not improve performance on the testing data.

### **5.3. Results and discussions**

#### **Autonomous Prediction with Reservoir:**

Long-term prediction results of the proposed reservoir for MG time series prediction are shown in Fig. 5-3a. After training the reservoir up to 400 time-step data, the input is disconnected and the predicted output is connected to the reservoir input. The reservoir is able to predict the next 30 time-step output with very good accuracy and with a root mean squared error (RMSE) of 0.0015. As the errors after 30-time steps prediction is extremely small, further prediction is possible. However, we restrict our effort due to the simulation complexity and limited hardware resources. Fig. 5-3c and Fig. 5-3d show the phase plot of the training and testing data of the chaotic MG attractor. The overlapping plots in Fig. 5-3c and 5-3d show that the reservoir was able to accurately predict both of the training and test data. When the same task is given to the LSTM as shown in Fig. 5-3b, it performs well and the resulting RMSE was 0.000013. This performance improvement can be due to the dependencies that arise among the LSTM cell states in the 2-layer deep architectures because of the use of forget gates (control how many previous states to remember). In addition, the non-linear activation functions are used for the input, output and forget gates which provide the non-linear transformation effect. Despite using much simpler architectures and linear activations, the reservoir is able to predict the chaotic trend with competitive accuracy. In our reservoir, during testing, the output is directly fed as the input, and this connection is not scaled or optimized (as has been done in [42]). Moreover, we did not use any non-linear transformation of the reservoir states, rather use the states as it is, and included several previous states. Thus, the success of the reservoir for long term prediction can be attributed to the use of previous reservoir states during the training and testing which provides the necessary non-linearity. Interestingly, here the non-linearity arises from the physical reservoir's dynamic responses rather than using any external non-linear transformation (trigonometric or polynomial activation). Furthermore, we did not use any temporal mask to encode an input for generating diversified features, instead we used different geometry skyrmion devices for variability. The only parameter that is scanned and optimized during training is the number of previous reservoir states. For MG prediction, we used reservoir states for 30 previous inputs, which is shown to provide adequate non-linearity. For long term prediction with output feedback, it is extremely important to predict the immediate future steps as accurately as possible, as any error in the near future can accumulate fast and make the prediction divergent.

Previously, reservoir with delayed input is shown to improve the one-step ahead prediction of the reservoir [45], which demonstrates the importance of the non-linearity effect coming from the delayed input to the prediction performance. In our proposed reservoir, accuracy of the long-term prediction is maintained due to the inclusion of previous states (similar effect as of the delayed input).

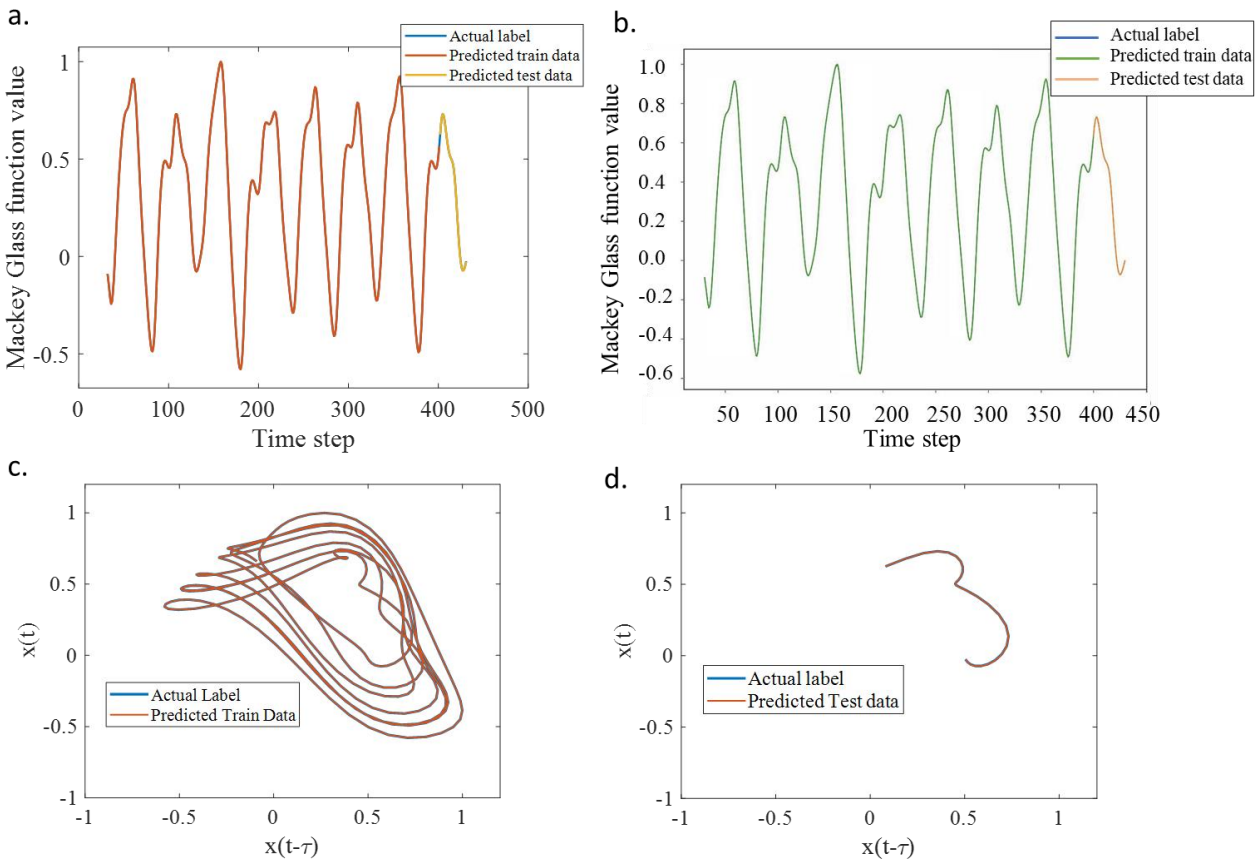


Figure 5-3 a. Long term autonomous prediction of chaotic MG time series with skyrmion reservoir. The dataset is trained with 31-400 time-step data. The reservoir is tasked to predict the next 30-time step data from 402-431. The overlapping of the predicted test data with actual label suggests accurate prediction b. prediction trend for MG time series with 2-layer deep sequence to sequence LSTM architecture. The LSTM is able to accurately predict the trend. Although, RMSE magnitude of the LSTM is lower than the reservoir, the prediction errors for both of the predictions remain extremely small. c. Phase diagram of the chaotic MG attractor during the training with reservoir. The predicted training data overlapped with the actual label implying the efficacy of the ridge regression training. d. Phase diagram of the reservoir for autonomous prediction. The superimposed plots suggest good prediction accuracy of the reservoir on test data.

After the successful performance of the proposed reservoir for the long-term chaotic time series prediction task, we focus on forecasting the long-term individual household power consumption. The task is challenging due to the non-volatile and univariate nature of the data (only the power value is readily

available, other parameters are unknown) and especially when the dataset is small. However, we find that the proposed skyrmion reservoir can achieve good accuracy when we use several previous reservoir responses. The total number of previous states that are included in the training are optimized and reservoir states for 20 previous inputs are used. The autonomous long-term prediction results are presented in Fig. 5-4a and 5-4b for the proposed reservoir and 2-layer deep S2S LSTM architecture respectively. From Fig. 5-4a, it is clear that the reservoir is able to predict the household power demand with good accuracy. The RMSE after 23 hours of prediction is calculated to be 0.0885. The prediction accuracy of the reservoir is good in the first several hours of the prediction (see the hourly RMSE plot in Fig. 5-5 labeled as reservoir: 262-284). In contrast, the accuracy of the LSTM is poor at the beginning of the prediction, however, regains accuracy in the next few predictions and the overall RMSE is calculated to be 0.0831. The accuracy degradation of the LSTM for the first few predictions could be due to lack of training data, as LSTM typically requires large numbers of observations to find the underlying dependencies in the data.

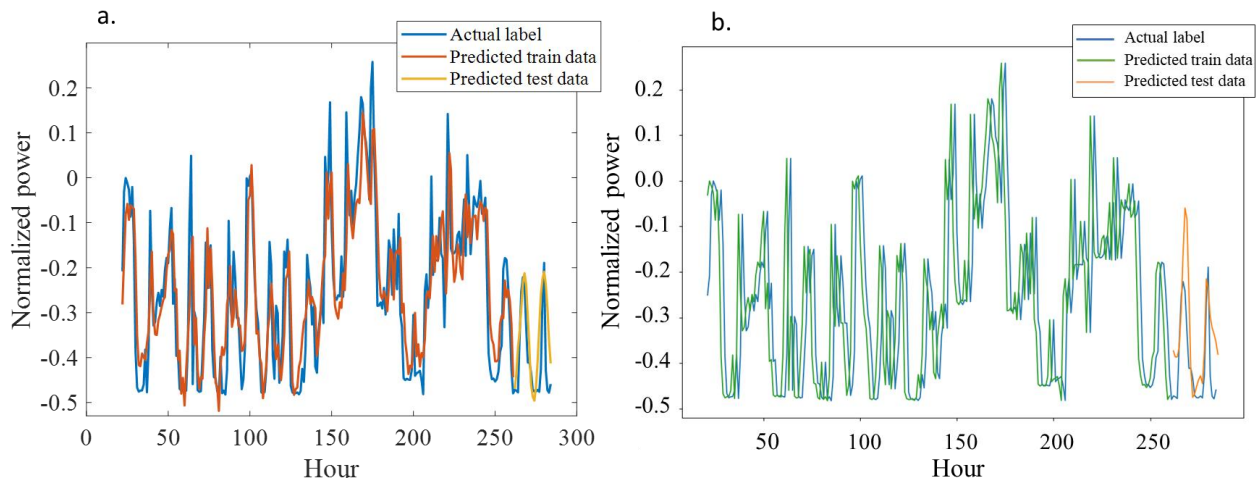


Figure 5-4 a. Long-term autonomous forecasting of individual household active power load with proposed reservoir. The reservoir is trained with 21-261 hours of data and tasked with predicting the next 23 hours of data. The predicted trend closely follows the actual load level suggesting good prediction accuracy with the skyrmion reservoir. b. The same task is performed with 2-layer sequence to sequence LSTM architecture. Although the LSTM is able to capture the trend, the prediction accuracy is less than the proposed reservoir for the first several hours of prediction.

Further, the reservoir is tasked to predict the next 286-308 hours of data where the load consumption trend is significantly stochastic. The hourly RMSE for both the reservoir and LSTM prediction for two different intervals: 262 to 284 hours and 286 to 308 hours are presented in Fig. 5-5. The training and testing trend of the reservoir for 286 to 308 hours interval is shown in the inset of Fig. 5-5. From the RMSE plot we can see that, in 286 to 308 hours interval, the initial prediction of the reservoir is beyond the magnitude of the regularization ( $\lambda=0.1$ ), however, the reservoir is able to minimize the error in the next few predictions and

follow the load trend. The large prediction error of the reservoir for the 286-308 hours of load compared to the previous set of prediction (262-284 hours) arises due to the significantly stochastic load behavior which is difficult to track with the reservoir. The limitation of the reservoir to track unstable stochastic trend is also observed in previous study [41].

For both of the prediction intervals, the reservoir prediction accuracy starts to diverge after several hours. As mentioned earlier, the long-term prediction with feedback depends on the accurate prediction of the immediate future. However, due to nonvolatile and stochastic trend of the household load data, the prediction accuracy degrades quickly, nonetheless good accuracy is found up to 20 hours. Compared to the reservoir, the LSTM prediction accuracy is poor for such stochastic load trends. The steady and higher prediction accuracy of the reservoir compared to the LSTM further proves the efficacy of the RC for the smaller dataset.

Finally, the RMSE of our proposed RC is compared with recent transformer [64] and generative recurrent unit (GRU) based optimization algorithms [65]. The RMSE of the RC for household power prediction task is computed to be  $\sim 0.6$  kW and the LSTM sequence to sequence model RMSE is 0.66 kW [51]. The best performance with regards to RMSE for the selective update and adaptive power optimization based GRU method is shown to be  $\sim 0.15$  kW [65]. On the other hand, the sparse transformer-based model with adversarial network learning-based approach [64] has RMSE  $\sim 0.30$  kW ( $\sim 50\%$  higher RMSE error compared to GRU based approach [65]). Although GRU and transformer-based approach shows higher prediction performance, they incur huge computational burden and associated energy cost for training. The computational complexity of the self-attention (most data intensive task in transformer) is quadratic as such the complexity for each position of a transformer can be expressed as  $O(d.N^2)$  where  $d$  is the dimension of each position of a sequence length of  $N$ . Additional complexity comes from feedforward linear operation, positional encoding, layer normalization. In GRU network based adaptive optimization approach, the hidden state information is compared at each time step which incurs an additional complexity of  $O(d)$  on top of the GRU unit's computational complexity of  $O(N.k.d)$ , where  $d$  is the dimension of hidden state vector and  $k$  is the number of GRU unit. Furthermore, to incorporate temporal dependency in gradient flow, the author proposes an additional hyperparameter called memory factor which requires additional multiplication and addition operation compared to conventional optimizer. Furthermore, the optimal value for the hyperparameter needs to be scanned using grid search, which is a computationally extensive process. In comparison, the reservoir states evolved by the internal dynamics of spintronic device. Thus, only a single feedforward layer needs to be trained to predict the future trajectories of the time-series. For a  $P \times 1$  layer feedforward network ( $P$  is the number of reservoir internal states that are read, and the RC output is a single prediction), the computational complexity can be expressed as  $O(P)$  which is significantly smaller



than the transformer, GRU and LSTM based networks. Thus, physical RC based approach could be a viable solution for edge intelligence especially in the scenario where we can trade-off accuracy in favor of limited resources.

#### 5.4. Energy Dissipation

There are two main contributions to the energy consumed for reservoir computing discussed here. First, energy is needed to modulate the PMA of the ferromagnetic layers. Maximum change in PMA coefficients is,  $\Delta PMA = 0.75 \times 10^3 \text{ J/m}^3$ . Thus, the maximum change in the surface anisotropy coefficients is estimated to be,  $\Delta K_{si} = \frac{\Delta PMA}{t_{CoFeB}} = \frac{0.75 \times 10^3}{1 \times 10^{-9}} = 0.75 \times 10^{12}$ . Assuming the VCMA coefficient of the MTJ to be,  $\varepsilon = \frac{\Delta K_{si}}{\Delta V / t_{MgO}} = 31 \text{ fJ}$  [66], the thickness of the tunneling barrier MgO to be,  $t_{MgO} = 1 \text{ nm}$ , the magnitude of voltage to perform the maximum PMA modulation can be calculated to be,  $\Delta V = 0.24 \text{ V}$ . Assuming the relative permittivity of MgO to be 7, the total capacitance can be calculated to be,  $C = \frac{\varepsilon_0 \varepsilon_r (L * L)}{t_{CoFeB}} \sim 62 \text{ fF}$ . Here, we assume  $L = 1050 \text{ nm}$  (so our estimate is conservative) for the length of the side of the square region of ferromagnetic layer. Thus, the total write energy to charge the capacitive tunneling region is estimated to be  $\frac{1}{2} CV^2 \sim 2 \text{ fJ}$ . Second, the reservoir responses that are read after certain interval (3 ns) known as virtual nodes also consume energy. The read energy can be estimated to be  $\sim 1.24 \text{ fJ}$  with a read delay of  $\sim 0.3 \text{ ns}$  [67] (which is well within 3 ns interval). In a period of 18 ns, the read operation is performed 6 times. Thus, the total energy for the write (PMA modulation) and read energy is calculated to be  $\sim 9.44 \text{ fJ}$ . For the household prediction task, a total of 21 hours of data (which translates to 21 discrete data points for the reservoir computing) are used to predict the next-hour household power. Thus, the total energy dissipation for one skyrmion reservoir is  $= 21 * 9.44 \sim 198 \text{ fJ}$ . During the reservoir initialization stage, the PMA is modulated. However, the states are not read. The total energy during the reservoir initialization stage is estimated to be  $\sim 440 \text{ fJ}$ . Including the reservoir initialization energy, the total energy consumption for 3 skyrmion reservoirs to predict one future value of individual household energy consumption is estimated to be  $\sim 651 \text{ fJ}$  (assuming worst case scenario). For LSTM implementation, the GPU energy is calculated to be  $\sim 0.68 \text{ J}$  per prediction. With reservoir implementation, the output layer is a feedforward layer that is implemented in a GPU as well for fair comparison. The GPU energy consumption for the reservoir feedforward layer is calculated to be  $0.043 \text{ J}$  per prediction. Thus, the reservoir is able to predict one instance with  $16 \times$  lower energy compared to the LSTM. We note that further reduction in energy consumption with reservoir can be achieved by implementing feedforward output layer with crossbar array of non-volatile memory using in-memory computing [68,69]. Thus, we could get an overall (at the system level which is the key metric of interest)  $10 \times$  to  $100 \times$  reduction in the computing energy needed to predict

building energy. At the individual RC MTJ device level, the energy savings would be enormous (pico-Joules instead of milli-Joules) but we do not use this metric as it is not a fair comparison if overall architecture and systems are not considered.

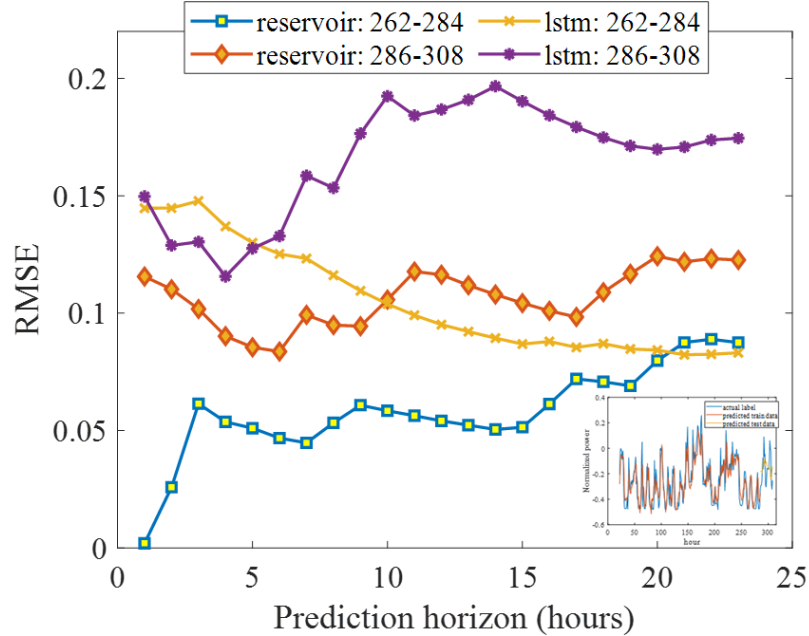


Figure 5-5 a. Hourly RMSE of the prediction accuracy for individual household load forecasting task for both of the proposed reservoir and LSTM. RMSE plots for two different long-term autonomous predictions, 262-284 hours and 286-308 hours are shown. The inset shows the prediction trend of the reservoir for prediction from 286 hour to 308 hours. The RMSE plots indicate higher prediction accuracy of the proposed reservoir compared to LSTM, even for much stochastic trend such as in 286-308 hours of data.

## 5.5 Conclusion

We have shown long-term autonomous prediction with a skyrmion reservoir. The reservoir is tasked with predicting the chaotic MG time series and real-world individual household load forecasting and is able to predict long-term trends with competitive accuracy. The proposed reservoir is set up using three patterned skyrmions having slightly different geometries. All the skyrmions are provided with the same temporal input series and the resulting skyrmions' oscillation is read at regular intervals and processed with simple linear regression. After training, the output is fed as the reservoir input to perform autonomous long-term prediction. The prediction performance greatly improves due to inclusion of previous states in addition to the current states of the reservoir as these previous states provide the non-linear effect necessary for accurate prediction. Furthermore, the physical reservoir does not consider masking the input and only the output weights and the number of previous states included during training are optimized. Energy consumption estimation shows that skyrmion reservoir can perform autonomous prediction with an energy consumption

of 0.043 J/per prediction which is at least  $16\times$  less than the LSTM based approach. In addition, we show that with our proposed physical RC one can achieve competitive accuracy with a much smaller dataset. Furthermore, with VCMA control, the skyrmion reservoir can be operated with ultra-low power as the anisotropy modulation is performed with voltage as opposed to energy hungry current control. Since in RC only the last layer is trained, thus our skyrmion reservoir provides a pathway to implement extremely energy efficient long-term prediction of real-world problem with high accuracy, which is specifically attractive in hardware and memory constraint edge computing platforms, where energy is at a premium.

## References:

- [1] P. J. Werbos, "Backpropagation through time: what it does and how to do it", Proc. IEEE, vol. 78, no. 10, pp. 1550 - 1560, Oct. 1990. DOI: 10.1109/5.58337.
- [2] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", Proc. Natl. Acad. Sci., vol. 79, no.8, pp. 2554-2558, Apr. 1982. DOI: <https://doi.org/10.1073/pnas.79.8.255>
- [3] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory", Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997. DOI: <https://doi.org/10.1162/neco.1997.9.8.1735>
- [4] M. Lukoševičius, H. Jaeger, "Reservoir computing approaches to recurrent neural network training", Comput. Sci. Rev., vol.3, pp. 127-149, Aug. 2009. DOI: 10.1016/J.COSREV.2009.03.005
- [5] G. Tanaka *et al.*, "Recent advances in physical reservoir computing: A review", Neural Netw., vol. 115, pp. 100-123, Jul. 2019. DOI: <https://doi.org/10.1016/j.neunet.2019.03.005>
- [6] F. Triefenbach, A. Jalalvand, B. Schrauwen, J.-P. Martens, "Phoneme Recognition with Large Hierarchical Reservoirs", in *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, Dec. 2010, pp. 1-9.
- [7] F. Wyffels, and B. Schrauwen, "A comparative study of Reservoir Computing strategies for monthly time series prediction", Neurocomputing, vol. 73, no. 10-12, pp. 1958-1964, Jun. 2010, doi: <https://doi.org/10.1016/j.neucom.2010.01.016>
- [8] R. Sacchi *et al.*, "Water Inflow Forecasting using the Echo State Network: a Brazilian Case Study", in *International Joint Conference on Neural Networks*, Orlando, FL, USA, Aug. 2007, pp. 1-6. DOI: 10.1109/IJCNN.2007.4371334
- [9] H. Jaeger, and H. Haas, "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication", Science, vol 304, no. 5667, pp. 78-80, Apr. 2004. DOI: 10.1126/science.10912

- [10] J. Misra, and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress", *Neurocomputing*, vol. 74, no. 1–3, pp. 239-255, Dec. 2010. DOI: <https://doi.org/10.1016/j.neucom.2010.03.021>
- [11] J. Hasler, and B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems", *Frontiers in Neuroscience*, vol. 7, Art. no. 118, pp. 1-29, Sep. 2013. DOI: <https://doi.org/10.3389/fnins.2013.00118>
- [12] M. C. Soriano *et al.*, "Delay-Based Reservoir Computing: Noise Effects in a Combined Analog and Digital Implementation", *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 2, pp. 388 - 393, Apr. 2014. DOI: 10.1109/TNNLS.2014.2311855
- [13] J. Li, K. Bai, L. Liu, Y. Yi, "A deep learning based approach for analog hardware implementation of delayed feedback reservoir computing system", in *19th International Symposium on Quality Electronic Design (ISQED)*, Santa Clara, CA, USA, Mar. 2018, pp. 1-6. DOI: 10.1109/ISQED.2018.8357305
- [14] Y. Paquot *et al.*, "Optoelectronic Reservoir Computing", *Sci. Rep.*, vol. 2, Art. no. 287, pp. 1-6, Feb. 2012. DOI: <https://doi.org/10.1038/srep00287>
- [15] K. Vandoorne *et al.*, "Experimental demonstration of reservoir computing on a silicon photonics chip", *Nat. Commun.*, vol. 5, Art. no. 3541, pp. 1-6, Mar. 2014. DOI: <https://doi.org/10.1038/ncomms4541>
- [16] C. Du *et al.*, "Reservoir computing using dynamic memristors for temporal information processing", *Nat. Commun.*, vol. 8, Art. no. 2204, pp. 1-10, Dec. 2017. DOI: <https://doi.org/10.1038/s41467-017-02337>
- [17] J. Moon *et al.*, "Temporal data classification and forecasting using a memristor-based reservoir computing system", *Nat. Electron.*, vol. 2, pp. 480–487, Oct. 2019. DOI: <https://doi.org/10.1038/s41928-019-0313-3>
- [18] J. Torrejon *et al.*, "Neuromorphic computing with nanoscale spintronic oscillators", *Nature*, vol. 547, pp. 428–431, Jul. 2017. DOI: <https://doi.org/10.1038/nature23011>
- [19] D. Pinna, G. Bourianoff, and K. Everschor-Sitte, "Reservoir Computing with Random Skyrmion Textures", *Phys. Rev. Appl.*, vol. 14, Art. no. 054020, pp. 1-12, Nov. 2020. DOI: <https://doi.org/10.1103/PhysRevApplied.14.054020>
- [20] R. Nakane, G. Tanaka, A. Hirose, "Reservoir Computing With Spin Waves Excited in a Garnet Film", *IEEE Access*, vol. 6, pp. 4462 - 4469, Jan. 2018. DOI: 10.1109/ACCESS.2018.2794584
- [21] V. Ababei *et al.*, "Neuromorphic computation with a single magnetic domain wall", *Sci. Rep.* vol.11, Art. no. 15587, pp. 1-13, Aug. 2021. DOI: <https://doi.org/10.1038/s41598-021-94975-y>

- [22] W. Jiang et al., “Physical reservoir computing using magnetic skyrmion memristor and spin torque nano-oscillator”, *Appl. Phys. Lett.*, vol.115, no. 19, pp. 1-6, Nov. 2019. DOI: <https://doi.org/10.1063/1.5115183>
- [23] H.-S. Philip Wong, and S. Salahuddin, “Memory leads the way to better computing, *Nat. Nanotechnol.*”, vol 10, pp. 191-194, Mar. 2015. DOI: <https://doi.org/10.1038/nnano.2015.29>
- [24] P. Chi et al., “Architecture design with STT-RAM: Opportunities and challenges”, in *21st Asia and South Pacific Design Automation Conference (ASP-DAC)*, Macao, China, Jan. 2016, pp. 1-6. DOI: [10.1109/ASPDAC.2016.7427997](https://doi.org/10.1109/ASPDAC.2016.7427997)
- [25] D. A. Allwood et al., “A perspective on physical reservoir computing with nanomagnetic devices”, *Appl. Phys. Lett.*, vol. 122, no. 4, pp. 1-11, Jan. 2023. DOI: <https://doi.org/10.1063/5.0119040>
- [26] A. J. Edwards et al., “Passive frustrated nanomagnet reservoir computing”, *arXiv:2103.09353*, Sept. 2022.
- [27] M. F. F. Chowdhury et al. “Focused surface acoustic wave induced nano-oscillator based reservoir computing”, *Appl. Phys. Lett.*, vol. 121, no. 10, pp. 1-7, Sept. 2022. DOI: <https://doi.org/10.1063/5.0110769>
- [28] M. Dale et al., “Computing with magnetic thin films: Using film geometry to improve dynamics”, in *International Conference on Unconventional Computation and Natural Computation*, Espoo, Finland, Oct. 2021, pp. 19–34. DOI: [https://doi.org/10.1007/978-3-030-87993-8\\_2](https://doi.org/10.1007/978-3-030-87993-8_2)
- [29] J. C. Gartside et al., “Reconfigurable training and reservoir computing in an artificial spin-vortex ice via spin-wave fingerprinting”, *Nat. Nanotechnol.*, vol.17, pp. 460–469, May 2022. DOI: <https://doi.org/10.1038/s41565-022-01091-7>
- [30] R. W. Dawidek et al, “Dynamically driven emergence in a nanomagnetic system”, *Adv. Funct. Mater.*, vol. 31, no. 15, pp. 1-13, Mar. 2021. DOI: [10.1002/adfm.202008389](https://doi.org/10.1002/adfm.202008389)
- [31] M. M. Rajib, W. A. Misba, M. F. F. Chowdhury, M. S. Alam, and J. Atulasimha, "Skyrmion based energy-efficient straintronic physical reservoir computing", *Neuromorph. Comput. Eng.*, vol. 2, no. 4, pp. 1-12, Nov. 2022. DOI: [10.1088/2634-4386/aca178](https://doi.org/10.1088/2634-4386/aca178)
- [32] O. Yokouchi et al., “Pattern recognition with neuromorphic computing using magnetic field–induced dynamics of skyrmions”, *Science Adv.*, vol. 8, no. 39, pp. 1-7, Sept. 2022. DOI: [10.1126/sciadv.abq5652](https://doi.org/10.1126/sciadv.abq5652)
- [33] S. Klingler et al., “Spin-Torque Excitation of Perpendicular Standing Spin Waves in Coupled YIG / Co Heterostructures”, *Phys. Rev. Lett.*, vol. 120, Art. no. 127201, pp. 1-6, Mar. 2018. DOI: <https://doi.org/10.1103/PhysRevLett.120.127201>

- [34] D. Bhattacharya, S. A. Razavi, H. Wu, B. Dai, K. L. Wang and J. Atulasimha, "Creation and annihilation of non-volatile fixed magnetic skyrmions using voltage control of magnetic anisotropy", *Nat. Electron.*, vol. 3, pp. 539–545, Jun. 2020. DOI: <https://doi.org/10.1038/s41928-020-0432-x>
- [35] D. Bhattacharya, and J. Atulasimha, "Skyrmion-mediated voltage-controlled switching of ferromagnets for reliable and energy-efficient two-terminal memory", *ACS Appl. Mater. Interfaces*, vol. 10, no. 20, pp. 17455–17462, Apr. 2018. DOI: <https://doi.org/10.1021/acsami.8b02791>
- [36] K. L. Wang, H. Lee, and P. K. Amiri, "Magnetoelectric Random Access Memory-Based Circuit Design by Using Voltage-Controlled Magnetic Anisotropy in Magnetic Tunnel Junctions", *IEEE Trans. Nanotechnol.*, vol. 14, no. 6, pp. 992 - 997, Nov. 2015. DOI: [10.1109/TNANO.2015.2462337](https://doi.org/10.1109/TNANO.2015.2462337)
- [37] D. Bhattacharya, M. M. Al-Rashid, and J. Atulasimha, "Voltage controlled core reversal of fixed magnetic skyrmions without a magnetic field", *Sci. Rep.*, vol. 6, Art. no. 31272, pp. 1-6, Aug. 2016, DOI: <https://doi.org/10.1038/srep31272>
- [38] N. D'Souza, M. S. Fashami, S. Bandyopadhyay, and J. Atulasimha, "Experimental clocking of nanomagnets with strain for ultralow power Boolean logic", *Nano Lett.*, vol. 16, no. 2, pp. 1069–1075, Jan. 2016. DOI: <https://doi.org/10.1021/acs.nanolett.5b04205>
- [39] A. K. Biswas, S. Bandyopadhyay, and J. Atulasimha, "Complete magnetization reversal in a magnetostrictive nanomagnet with voltage-generated stress: A reliable energy-efficient non-volatile magneto-elastic memory", *Appl. Phys. Lett.*, vol. 105, no. 7, Art. no. 072408, pp. 1-5, Jul. 2014. DOI: <https://doi.org/10.1063/1.4893617>
- [40] M. Inubushi, and K. Yoshimura, "Reservoir Computing Beyond Memory-Nonlinearity Trade-off", *Sci. Rep.*, vol. 7, Art. no. 10199, pp. 1-10, Aug. 2017. DOI: <https://doi.org/10.1038/s41598-017-10257-6>
- [41] P. Chen, R. Liu, K. Aihara, and L. Chen, "Autoreservoir computing for multistep ahead prediction based on the spatiotemporal information transformation", *Nat. Commun.*, vol. 11, pp. 1-15, Sept. 2020. DOI: <https://doi.org/10.1038/s41467-020-18381-0>
- [42] P. Antonik, M. Haelterman, and S. Massar, "Brain-Inspired Photonic Signal Processor for Generating Periodic Patterns and Emulating Chaotic Systems", *Phys. Rev. Applied*, vol. 7, Art. no. 054014, pp. 1-16, May 2017. DOI: <https://doi.org/10.1103/PhysRevApplied.7.054014>
- [43] L. Jaurigue, E. Robertson, J. Wolters, and K. Lüdge, "Reservoir Computing with Delayed Input for Fast and Easy Optimisation", *Entropy*, vol. 23, no. 12, pp. 1-13, Nov. 2021. DOI: <https://doi.org/10.3390/e23121560>
- [44] D. J. Gauthier, E. Bollt, A. Griffith, and W. A. S. Barbosa, "Next generation reservoir computing", *Nat. Commun.*, vol. 12, Art. no. 5564, pp. 1-8, Sept. 2021. DOI: <https://doi.org/10.1038/s41467-021-25801-2>

- [45] L. Appeltant et al., "Information processing using a single dynamical node as complex system", *Nat. Commun.*, vol. 2, Art. no. 468, pp. 1-6, Sept. 2011. DOI: <https://doi.org/10.1038/ncomms1476>
- [46] S. Naji et al., "Estimating building energy consumption using extreme learning machine method", *Energy*, vol. 97, pp. 506-516, Feb. 2016. DOI: <https://doi.org/10.1016/j.energy.2015.11.037>
- [47] "Building controls," *Energy.gov*. [Online]. Available: <https://www.energy.gov/eere/buildings/building-controls>. [Accessed: 21-Jan-2023].
- [48] C. Yuan, S. Liu, and Z. Fang, "Comparison of China's primary energy consumption forecasting by using ARIMA (the autoregressive integrated moving average) model and GM(1,1) model", *Energy*, vol. 100, pp. 384-390, Apr. 2016. DOI: <https://doi.org/10.1016/j.energy.2016.02.001>
- [49] R. Weron, *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*, 1st ed. Chichester: Wiley, 2006.
- [50] H. Shi, M. Xu, and R. Li, "Deep Learning for Household Load Forecasting—A Novel Pooling Deep RNN", *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5271 - 5280, Mar. 2017. DOI: 10.1109/TSG.2017.2686012
- [51] D. L. Marino, K. Amarasinghe, and M. Manic, "Building energy load forecasting using Deep Neural Networks", in *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, Florence, Italy, Oct. 2016, pp. 1-6. DOI: 10.1109/IECON.2016.7793413
- [52] T. Maruyama et al., "Large voltage-induced magnetic anisotropy change in a few atomic layers of iron", *Nat. Nanotechnol.*, vol. 4, pp. 158–161, Jan. 2009. DOI: <https://doi.org/10.1038/nnano.2008.406>
- [53] M. Weisheit et al., "Electric Field-Induced Modification of Magnetism in Thin-Film Ferromagnets", *Science*, vol. 315, no. 5810, pp. 349-351, Jan. 2007. DOI: 10.1126/science.1136629
- [54] M. M. Rajib, W. A. Misba, D. Bhattacharya, F. G.-Sanchez, and J. Atulasimha, "Dynamic Skyrmion-Mediated Switching of Perpendicular MTJs: Feasibility Analysis of Scaling to 20 nm With Thermal Noise", *IEEE Trans. Electron Devices*, vol. 67, no. 9, pp. 3883 - 3888, Sept. 2020. DOI: 10.1109/TED.2020.3011659
- [55] L. D. Landau, and E. M. Lifshitz, "Theory of the dispersion of magnetic permeability in ferromagnetic bodies", *Phys. Z. Sowietunion*, vol. 8, pp. 153–169, 1935.
- [56] T. L. Gilbert, "Lagrangian formulation of the gyromagnetic equation of the magnetization field," *Phys. Rev.*, vol. 100, pp. 1243–1243, 1955.
- [57] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. V. Waeyenberge, "The design and verification of MuMax3," *AIP Adv.*, vol. 4, no. 10, Art. no. 107133, Oct. 2014. DOI: 10.1063/1.4899186

- [58] M. J. Donahue, and D. G. Porter, "Exchange energy formulations for 3d micromagnetics", *Phys. B: Condens. Matter*, vol. 343, no: 1-4, pp. 177–183, Jan. 2004. DOI: <https://doi.org/10.1016/j.physb.2003.08.090>
- [59] A. N. Bogdanov, and U. K. Röbber, "chiral symmetry breaking in magnetic thin films and multilayers", *Phys. Rev. Lett.*, vol 87, Art. no. 3, pp. 1-4, Jul. 2001. DOI: 10.1103/PhysRevLett.87.037203
- [60] P. K. Amiri et al., "Electric-Field-Controlled Magnetoelectric RAM: Progress, Challenges, and Scaling", *IEEE Trans. Magn.*, vol. 51, no. 11, pp. 1-7, Nov. 2015. DOI: 10.1109/TMAG.2015.2443124
- [61] K. Bache, and M. Lichman, UCI Machine Learning Repository, University of California, School of Information and Computer Science, 2013, Available: <http://archive.ics.uci.edu/ml>
- [62] P. Antonik, M. Hermans, F. Duport, M. Haelterman, S. Massar, "Towards pattern generation and chaotic series prediction with photonic reservoir computers, Towards pattern generation and chaotic series prediction with photonic reservoir computers", *Proc. SPIE 9732, Real-time Measurements, Rogue Events, and Emerging Applications, 97320B*, San Francisco, California, United States, Mar. 2016, pp. 1-12. DOI: <https://doi.org/10.1117/12.2210948>
- [63] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks", in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, Dec. 2014, Cambridge: MIT Press, vol. 2, pp. 3104–3112.
- [64] S. Wu, X. Xiao, Q. Ding, P. Zhao, Y. Wei, and J. Huang, "Adversarial sparse transformer for time series forecasting," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, 2020, pp. 1–8.
- [65] W. Zheng, and G. Chen, "An Accurate GRU-Based Power Time-Series Prediction Approach With Selective State Updating and Stochastic Optimization", *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13902-13914, Dec. 2022. DOI: 10.1109/TCYB.2021.3121312.
- [66] J. G. Alzate et al., "Temperature dependence of the voltage-controlled perpendicular anisotropy in nanoscale MgO|CoFeB|Ta magnetic tunnel junctions", *Appl. Phys. Lett.*, vol. 104, no. 11, Art. no. 112410, pp. 1-5, Mar. 2014. DOI: <https://doi.org/10.1063/1.4869152>
- [67] W. Kang, L. Chang, Y. Zhang, and W. Zhao, "Voltage-Controlled MRAM for Working Memory: Perspectives and Challenges", in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Lausanne, Switzerland, Mar. 2017, pp. 1-6. DOI: 10.23919/DATE.2017.7927047
- [68] S. Ambrogio et al., "Equivalent accuracy accelerated neural-network training using analogue memory", *Nature*, vol. 558, pp. 60–67, Jun. 2018. DOI: 10.1038/s41586-018-0180-5.
- [69] W. A. Misba, M. Lozano, D. Querlioz, J. Atulasimha, "Energy Efficient Learning With Low Resolution Stochastic Domain Wall Synapse for Deep Neural Networks", *IEEE Access*, vol. 10, pp. 84946 - 84959, Aug. 2022. DOI: 10.1109/ACCESS.2022.3196688



## Chapter 6: Magnetic Anisotropy Modulation in Bismuth Substituted Yttrium Iron Garnet with Voltage Controlled Strain

In Chapter 2, we demonstrated energy-efficient magnetic memory devices by resonantly exciting magnetostrictive nanomagnets deposited on a piezoelectric substrate using voltage-induced strains. The efficiency of these devices can be further enhanced by utilizing magnetic materials with extremely low damping constants. As illustrated in the resonance characteristics of an oscillator in Fig. 6-1, damping is associated with the quality factor (Q-factor), which governs the rate at which the oscillator loses energy. Magnetic materials such as Bismuth-substituted Yttrium Iron Garnet (Bi-YIG), which exhibit extremely low damping constants ( $1.3\text{-}3 \times 10^{-4}$ ), can provide higher Q-factors and sustain higher amplitude magnetization precessions for longer periods of excitation. This motivates our study of the electric field tunability of Bi-YIG films deposited on a piezoelectric substrate.

In this chapter, we report magnetic anisotropy modulation in Bi-YIG thin films deposited on PMN-PT with the application of voltage-induced strain. The Bi content is selected for low coercivity and higher magnetostriction than that of YIG, yielding significant changes in the hysteresis loops through the magnetoelastic effect by application of voltage-induced stress. The piezoelectric substrate is poled along its thickness, which is the [011] direction, by applying voltage across the PMN-PT/SiO<sub>2</sub>/Bi-YIG/Pt heterostructure. In-situ magneto-optical Kerr microscopy (MOKE) reveals the modulation of magnetic anisotropy with voltage-induced strain. Furthermore, voltage control of the magnetic domain state of the Bi-YIG film at a fixed magnetic field produces a 90° switching of the magnetization easy axis above a threshold voltage. The magnetoelectric coefficient of the heterostructure is  $1.05 \times 10^{-7} \text{ sm}^{-1}$  which is competitive with that of other ferromagnetic oxide films on ferroelectric substrates. Our results demonstrate electric field control in a multiferroic heterostructure relevant to applications in energy efficient non-volatile memory and in-memory computing devices.

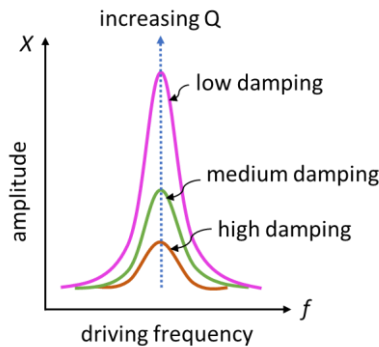


Figure 6-1. Sketch of resonance curves of an oscillator for different damping coefficients and increasing Q factors.

Electric field tunability of magnetization is particularly appealing for high density magnetic data storage with low energy dissipation [1,2] compared to current-based technologies [3,4]. In this regard, multiferroic structures with coupled ferroelectric (FE) and ferromagnetic properties have been well studied due to their ability to control the electric and magnetic ordering simultaneously through the converse magnetoelectric effect (CME) [5,6]. The electric current requirement is on the order of  $10^{11}$  A/m<sup>2</sup> with 10 fJ [7] dissipation compared to 1-100 aJ dissipation in capacitive multiferroic devices [7, 8]. Although single phase multiferroic materials [9,10] are desirable, composite heterostructures provide three to four orders of magnitude greater magnetoelectric effect and excellent stability at room temperature [11,12]. Several mechanisms have been explored for harnessing CME from composite heterostructures, such as transferring mechanical strain from the FE to the ferromagnet [13-15], modulation of the spin-up and spin-down densities of states at the FE-ferromagnet interface [16], and modification of an oxide ferromagnet through voltage driven migration of oxygen [17]. Strain transfer mechanisms demonstrate low heat dissipation per switching cycle and higher magnetoelectric coefficient compared to other mechanisms [18].

Relaxor ferroelectric materials such as PMN-PT show large piezoelectric coefficients when operated near the morphotropic phase boundary ( $x=0.3$ ) and are often used to transfer strain to a ferromagnetic material [19]. Ferromagnetic materials with low to moderate magnetostriction such as Ni [20], Co [13,14], CoFeB [15, 21], or FeGa [22] have been fabricated on top of PMN-PT to investigate magnetoelectric effects. In these composites, the magnetic films and patterned dots are often amorphous or polycrystalline in nature, thus electric field induced magnetoelastic anisotropy dominates magnetocrystalline anisotropy enabling 90° switching of the magnetic easy axis when a voltage is applied. Complete 180° switching was demonstrated in patterned Co/PMN-PT by sequentially applying voltages in the electrode pairs [23]. In contrast to ferromagnets, ferrimagnetic materials such as oxides offer more efficient and faster control of magnetization state due to their low damping and moderate saturation magnetization. Ferrimagnetic oxides such as yttrium iron garnet (YIG) and rare-earth iron garnet (REIG) have been used to demonstrate spin wave propagation and spin torque phenomena. In addition, the saturation magnetization, magnetostriction, anisotropy and Gilbert damping parameter can be modified by inserting rare earth ions [24]. Despite these advantages, the growth of ferrimagnetic garnets on piezoelectric compounds poses a significant challenge due to lattice incompatibility, thus limiting the potential to harness the benefit of electrical control. Previously, we have shown remnant strain induced anisotropy modulation of yttrium-substituted dysprosium iron garnet (YDyIG) film crystallized on PMN-PT [25].

In this study, we grow bismuth-substituted yttrium iron garnet (Bi-YIG) on PMN-PT and show magnetoelectric control by applying different electric fields across the heterostructure while characterizing their hysteresis loops in-situ using magneto-optical Kerr microscopy (MOKE). Bi-YIG has great

technological significance for magneto-optical and magnetoelectric applications such as microwave splitters, optical isolators [26], magnetic sensors and spin wave carriers with extremely low dissipation and eddy current loss. Magnetic domain wall propagation in a BiYIG racetrack was demonstrated recently using low power rf pulses with duration of 1 ns and above [27]. In this chapter we first discuss the growth and magnetic properties of the BiYIG film on a PMN-PT substrate, then show that the magnetic easy axis of Bi-YIG films can be reoriented by 90° degrees by applying an electric field across the heterostructure composite. Furthermore, MOKE shows domain wall nucleation and propagation in BiYIG films for varying amplitudes of the electric field. We then demonstrate voltage induced strain control of the magnetic domain at a fixed magnetic field and estimate the magnetoelectric coefficient. These results suggest a path towards control of magnetic bits in a BiYIG domain-wall device using a combination of voltage and current pulses [28,29,30].

### **6.1 Sample growth and characterization:**

The heterostructure samples employed in our experiments were grown by our collaborator at MIT. The (011)-oriented PMN-PT [ $\text{PbMg}_{0.33}\text{Nb}_{0.67}\text{O}_3$ ] $_{1-x}$ ( $\text{PbTiO}_3$ ) $_x$ ;  $x=0.29-0.33$  substrates (supplied by MTI Corp.) were coated with 2.4 nm  $\text{SiO}_2$  by magnetron sputtering. The 45.6 nm thick BiYIG films were grown via pulsed laser deposition (PLD) by co-deposition from stoichiometric YIG ( $\text{Y}_3\text{Fe}_5\text{O}_{12}$ ) and BFO ( $\text{BiFeO}_3$ ) targets. A 248 nm KrF excimer laser was used at an energy of 600 mJ, a repetition rate of 10 Hz, and was focused to a fluence of  $2 \text{ J cm}^{-2}$  at the target. The laser shots on each target were adjusted based on the calibrated growth rates. The chamber was pumped to a base pressure of  $1 \times 10^{-5}$  Torr and an oxygen pressure of 20 mTorr was maintained during the deposition. The films then underwent ex-situ anneal in a furnace for 72 hours at 600 °C in ambient temperature to crystallize the garnet. Fig. 6-2a shows the Grazing incidence x-ray diffraction (GIXD) images which show the growth of Bi-YIG on  $\text{SiO}_2/\text{PMN-PT}$  substrate. We note that, ~ 5 nm thick  $\text{SiO}_2$  diffusion barrier is deposited on top of PMN-PT before depositing BiYIG to avoid forming perovskite structure which is undesirable. Next, the hysteresis loops of the deposited heterostructures are measured for external fields along in-plane and out of plane directions. The sample hysteresis in Fig. 6-2b shows the magnetizations prefer to orient along an in-plane direction. The saturation magnetization is measured to be 101 kA/m and the in-plane coercivity is around 10 mT. The scanning electron microscopy (SEM) images are presented for the BiYIG deposited on Si and  $\text{SiO}_2/\text{PMN-PT}$  in Fig. 6-3. Although a few tiny amorphous regions are observed, overall, the BiYIG sample deposited on  $\text{SiO}_2/\text{PMN-PT}$  shows excellent crystallinity across the heterostructure.

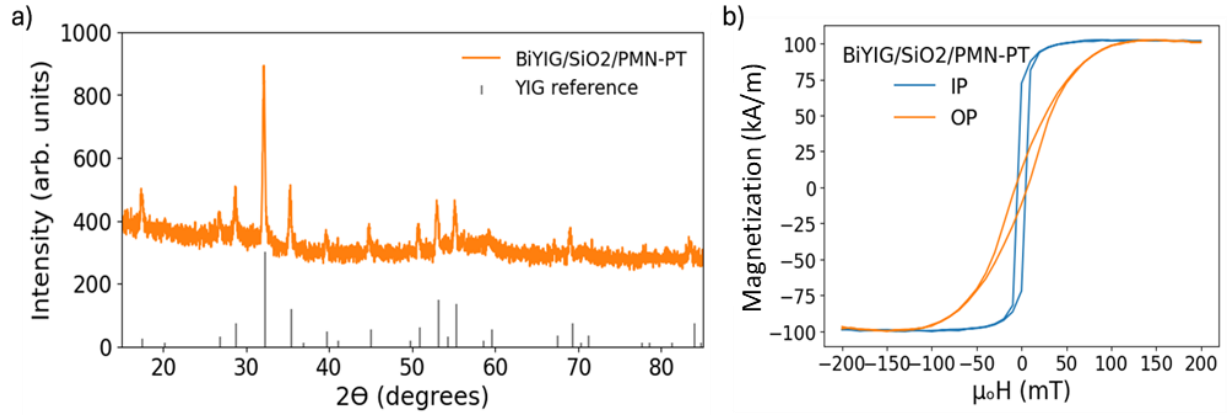


Figure 6-2 a. GIXD diffraction image shows Bi-YIG growth on SiO<sub>2</sub>/PMN-PT substrate. Data has been shifted vertically for clarity. b. Hysteresis loops taken via vibrating sample magnetometry of the BiYIG/SiO<sub>2</sub>/PMN-PT sample. The curves were measured out of plane (OP) and in plane (IP) to the sample surface.

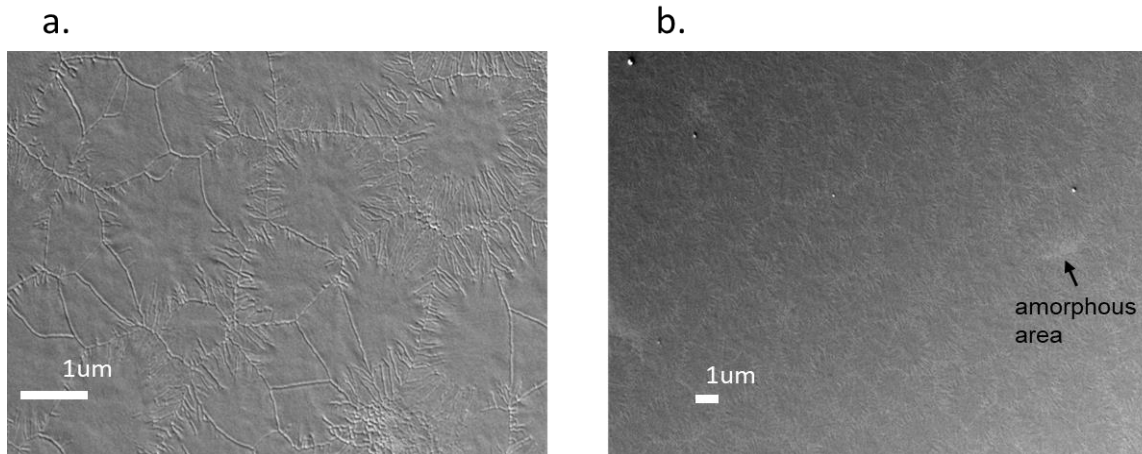


Figure 6-3 Top surface SEM images of a) BiYIG/Si and b) BiYIG/SiO<sub>2</sub>/PMN-PT.

## 6.2 Magnetic Hysteresis Modulation with Strain

The magnetic properties of the ferromagnetic material in a FE-ferromagnet heterostructure can be modulated by utilizing the piezoelectric properties of the FE crystal. Applying a voltage across the thickness of the PMN-PT (i.e. along the film normal, defined as  $\hat{z}$ , the [011] direction) generates an electric field  $E$  leading to a piezoelectric strain in the FE-crystal along the two orthogonal in-plane directions,  $\hat{x}$ , [100] and  $\hat{y}$ , the [01 $\bar{1}$ ] direction as shown in the heterostructure schematic in Fig. 6-4a. When the electric field is zero, the BiYIG shows isotropic magnetic behavior. An electric field along  $\hat{z}$  leads to compressive strain along  $\hat{x}$  and tensile strain along  $\hat{y}$  of the substrate, breaking the degeneracy of the BiYIG in-plane hysteresis loops.

To characterize the magnetoelectric behavior of the composite, we first poled the BiYIG/SiO<sub>2</sub>/PMN-PT by applying 450 V along  $\hat{z}$  ( $E = 9$  kV/cm) for 90 min, then set the voltage to zero. We then applied voltages in 50V increments, measuring the longitudinal and polar MOKE hysteresis loops at each voltage, Fig. 6-3a and Fig. 6-4b. Blue light (wavelength  $\sim 465$  nm) was used because BiYIG has a high MOKE response at this wavelength. The as-deposited sample is isotropic in plane and showed similar magnetic hysteresis loops along  $\hat{x}$  and  $\hat{y}$ . Poling and subsequent relaxation leads to a remnant strain in the PMN-PT, which is tensile along  $\hat{x}$  and compressive along  $\hat{y}$  [25,31]. Hence, after poling the Bi-YIG shows a harder direction (lower remanence and squareness) along  $\hat{x}$  and an easier direction along  $\hat{y}$  at 0 V compared with the as-grown state, consistent with a negative magnetostriction [32,33]. When voltage is subsequently applied to the poled sample, the PMN-PT experiences increasing compressive strain along  $\hat{x}$  due to the negative piezoelectric coefficient,  $d_{31}$  of PMN-PT [31], and tensile strain along  $\hat{y}$  due to the positive piezoelectric coefficient  $d_{32}$  of PMN-PT. This leads to an anisotropy reorientation in the Bi-YIG with  $\hat{x}$  becoming the easy in-plane direction for a sufficiently large voltage.

With reference to  $\theta$  and  $\varphi$ , the polar and azimuthal angles of magnetization shown in Fig. 6-4a, the magnetoelastic energy can be expressed as  $F_{me} = -\frac{3}{2}\lambda_s \frac{Y}{1+\vartheta} \varepsilon_{xx} \sin^2\theta \cos^2\varphi - \frac{3}{2}\lambda_s \frac{Y}{1+\vartheta} \varepsilon_{yy} \sin^2\theta \sin^2\varphi$ , where  $\varepsilon_{xx}$  and  $\varepsilon_{yy}$  are the strains along  $\hat{x}$  and  $\hat{y}$ ,  $Y$  is the Young's modulus, and  $\vartheta$  is the Poisson's ratio of BiYIG. There is no stress along  $\hat{z}$  due to the free boundary condition at the top surface. The saturation magnetostriction coefficient of BiYIG is,  $\lambda_s \approx -4 \times 10^{-6}$  [32,33], and the energy is lowered when the magnetization is aligned along a compressively strained direction.

Fig. 6-4a shows the hysteresis loops become squarer along  $\hat{x}$  as the voltage is increased from 0 to 450 V. The coercive field increases from 24.8 mT at 0 V to 27.1 mT at 450 V, and the field at which the loop closes decreases to  $55 \pm 0.5$  mT field at 450 V compared to  $70 \pm 0.5$  mT at 0 V. Opposite trends are observed along  $\hat{y}$ : the loop becomes less square with increasing voltage and the coercivity decreases from 30 mT at 0 V to 24.2 mT at 450 V, Fig. 6-4b. A high squareness ratio, defined as  $Sq = M_r/M_s$  where  $M_r$  and  $M_s$  are the remanent and saturation magnetization respectively, indicates the easy axis.  $Sq$  increases (decreases) along  $\hat{x}$  ( $\hat{y}$ ) with increasing voltage, Fig. 6-4a, 6-4b. A butterfly-like hysteresis loop is observed for  $M_r$  vs. V, Fig. 6-4c, illustrates the magnetoelectric coupling between the applied voltage and remanent magnetization. The loop measured along the poling direction,  $\hat{z}$ , shows little change with voltage, Fig. 6-4d.

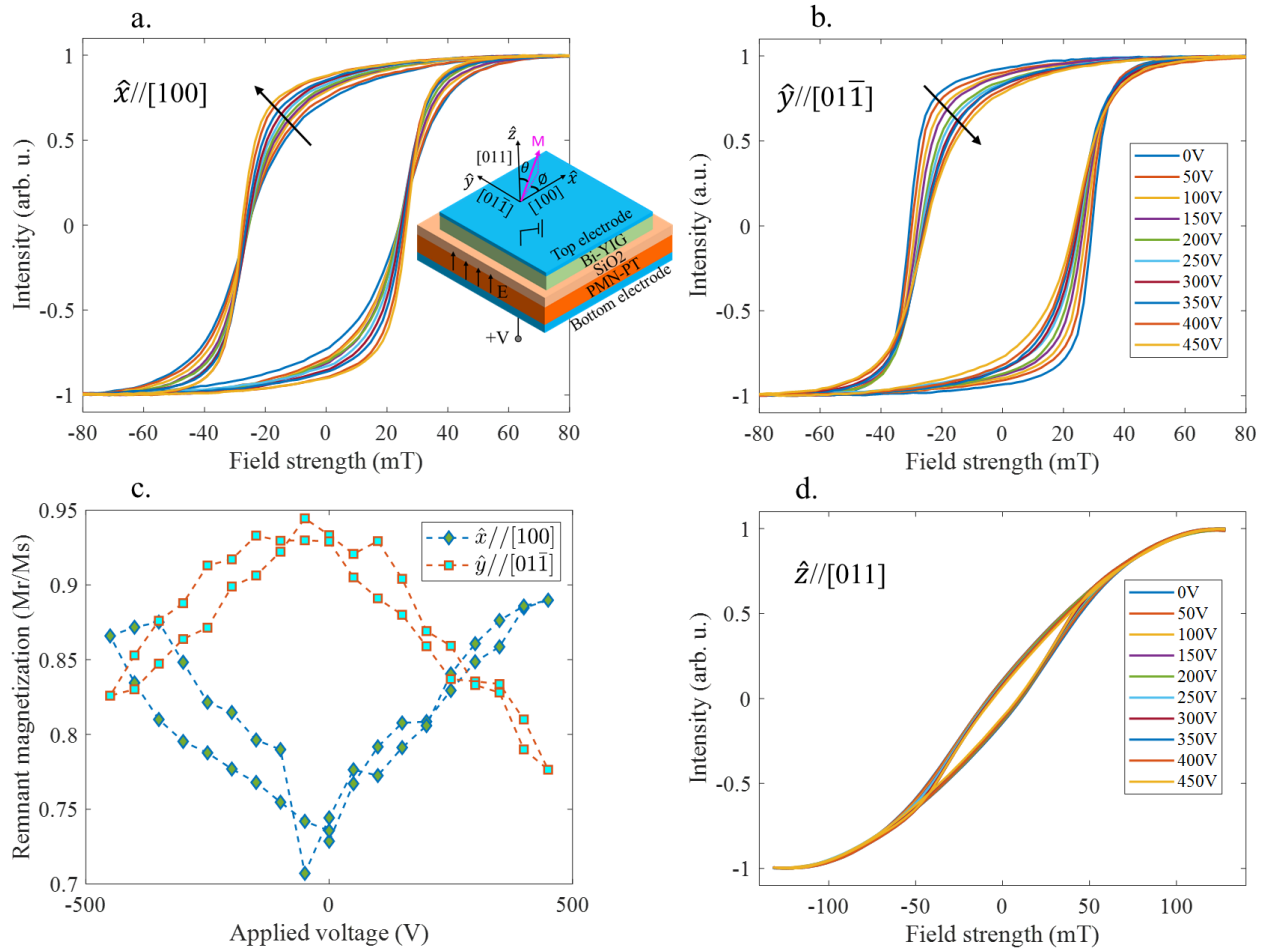


Figure 6-4 a,b. Hysteresis loops obtained from MOKE magnetometry for different voltages applied along the thickness of the heterostructure, PMN-PT/SiO<sub>2</sub>/BiYIG when the magnetic field is applied along the in-plane direction a.  $\hat{x}$ . b.  $\hat{y}$ . Black arrows indicate the trend for increased voltage. The inset in (a) shows a schematic of the heterostructure with the direction of the applied voltage, principal axes and the polar angle,  $\theta$  and azimuthal angle,  $\varphi$  of the BiYIG film magnetization,  $M$ . c. ratio of remnant and saturation magnetization vs the applied voltage for both in-plane directions,  $\hat{x}$  and  $\hat{y}$ . d. hysteresis loops as a function of voltage obtained from polar MOKE for the out of plane direction,  $\hat{z}$ .

In Fig. 6-5 we analyze the magnetization switching of the poled heterostructure for two cases, 0V and 450V for the in-plane directions  $\hat{x}$  and  $\hat{y}$ . Initially, a reference background image is taken from which the images acquired at different magnetic fields are subtracted. At positive saturation, +88 mT, predominant white contrast domains are observed. The arrows in Fig. 6-5b and 6-5d indicate specific domains, with upward (downward) arrows representing magnetization along the  $+\hat{x}$  ( $-\hat{x}$ ) or  $+\hat{y}$  ( $-\hat{y}$ ) direction. The images in Fig. 6-5b and 6-5d correspond to the positions marked on the hysteresis loops in Fig. 6-5a and 6-5c respectively.

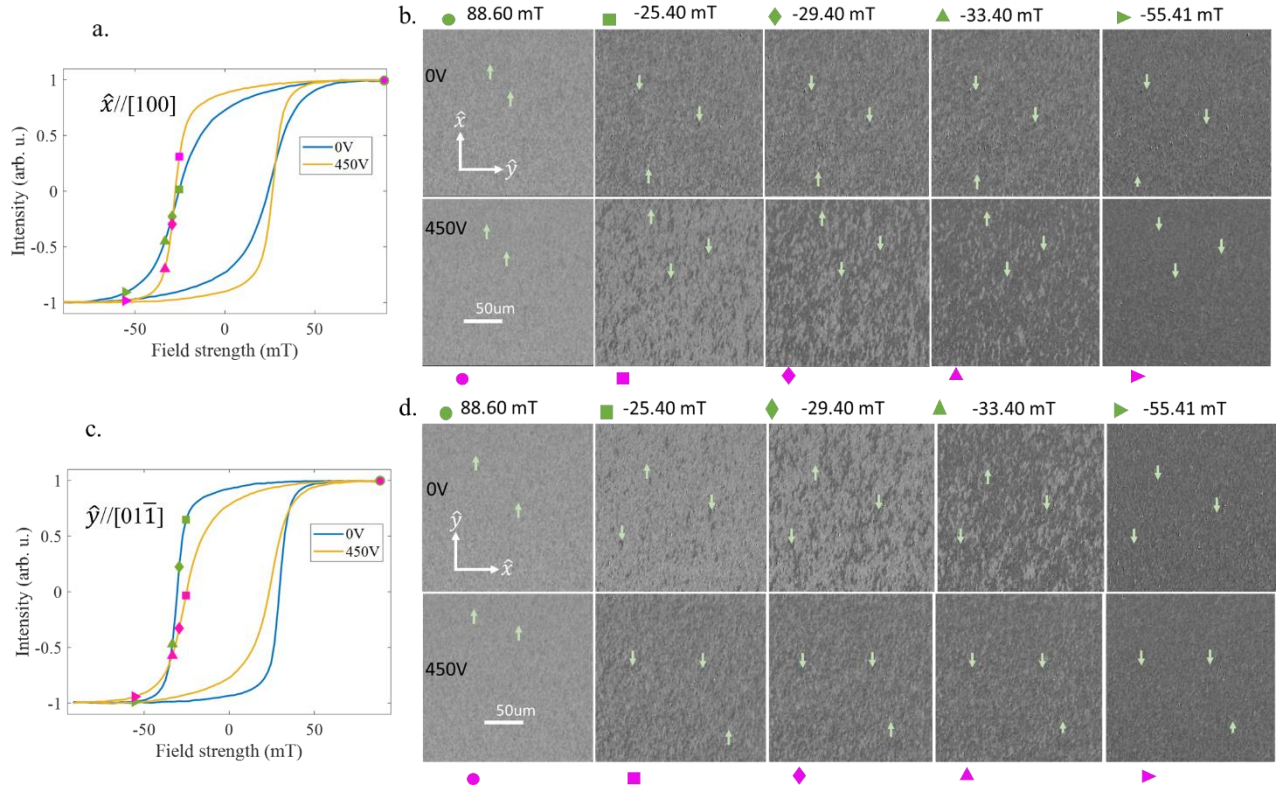


Figure 6-5 a. Hysteresis curves with external fields applied along the in-plane direction  $x$  when the heterostructure is subjected to an applied voltage of 0 V and 450 V. b. longitudinal MOKE images showing magnetization reversal process. The corresponding field values for which the images are taken are also marked in the hysteresis loops. The upward (downward) arrows represent domain magnetizations that are pointed along the  $+\hat{x}$  ( $-\hat{x}$ ) directions. c. hysteresis loops for in-plane direction  $y$  for 0 V and 450 V and d. corresponding magnetization reversal images. The upward (downward) arrows represent domain magnetizations oriented along the or  $+\hat{y}$  ( $-\hat{y}$ ) directions.

As the external field is increased in the negative direction, reversal is indicated by the black contrast. For fields applied along  $\hat{x}$ , when the voltage is 0 V, the switching corresponds to a gradual change in contrast, and no significant domain wall propagation is observed (compare the images at -29.4 mT and -33.4 mT in bottom panels of Fig. 6-5b). Also, the change in contrast starts to appear early at 0 V and the sample is not saturated at -55 mT. However, at 450V, abrupt switching is accomplished by domain wall nucleation and growth and the sample saturates at -55mT. This behavior is consistent with  $\hat{x}$  being the easy axis at 450 V but a hard axis at 0 V. Along  $\hat{y}$  direction the easy axis switching process occurs with domain wall nucleation and propagation at 0 V, but a gradual contrast change is observed at 450 V, consistent with  $\hat{y}$  becoming a hard axis with increasing voltage.

### 6.3 Magnetization evolution with electric field:

Fig. 6-6 shows the effect of voltage on the magnetic domain pattern for a fixed magnetic field. The sample was first saturated at  $-70$  mT applied field, then the field was set to  $+27$  mT and the domains were imaged as a function of voltage. The change in the domain formation pattern is most significant at  $27$  mT, which is close to the coercive fields for both of the in-plane directions.

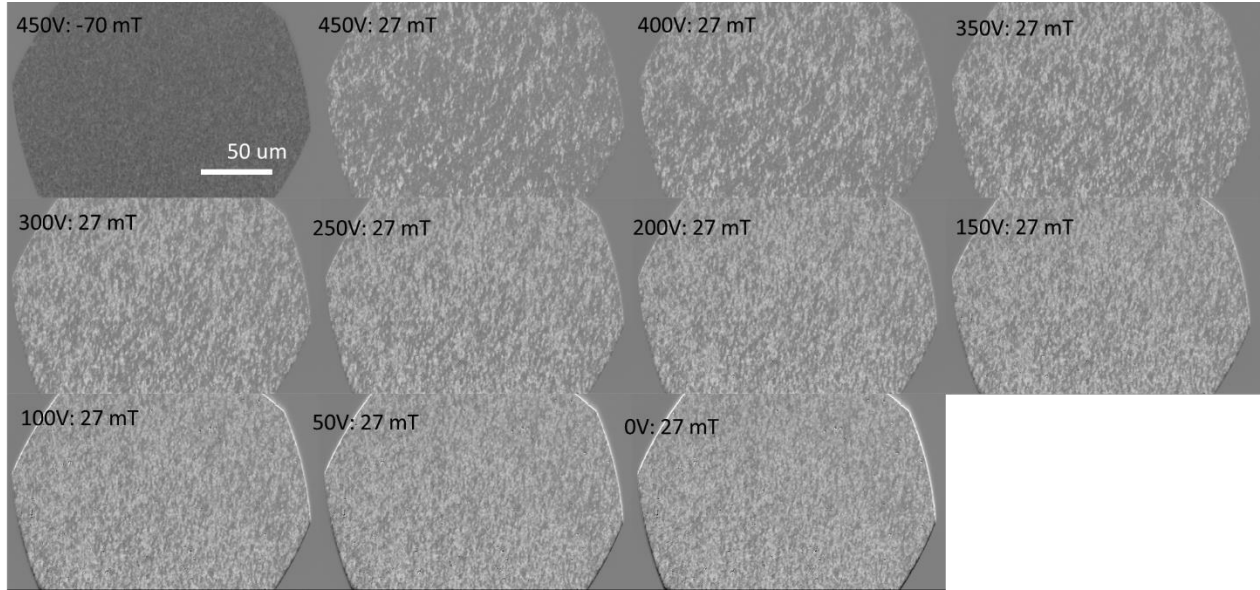


Figure 6-6 MOKE images showing domain reversal along in-plane direction  $\hat{x} // [100]$  for varying voltages. The sample is first saturated with a  $-70$  mT field and the reversal field is set to  $+27$  mT, then a voltage of  $450$  V was applied and reduced in steps of  $50$  V.

The sample was poled by applying  $450$  V and subsequently relaxed to  $0$  V. For image acquisition in the  $\hat{x}$ -direction, the voltage is then reset to  $450$  V, and a magnetic field of  $-70$  mT was applied. In this configuration, domains with black contrast are predominant. The external field is then increased to  $27$  mT. White-contrast domains indicate the onset of reversal, which increases as the voltage is stepped downwards in increments of  $50$  V while keeping the magnetic field at a constant  $27$  mT. The domain pattern shows little change for voltages below  $100$  V. The behavior is explained by the  $\hat{x}$  axis being the easy axis at  $450$  V but becoming unfavorable as the voltage decreases. The reduction in anisotropy leads to the reorientation of the magnetization towards the  $\hat{y}$ -axis, reduction in domain sizes and a weakening of contrast. At low voltages,  $\hat{x}$  is a hard direction and the magnetization orientation within domains is governed by the balance between the magnetoelastic anisotropy and the Zeeman energy. An analogous but opposite trend is found



when the field is applied along  $\hat{y}$ , shown in Fig. 6-7. Thus, the preferred axis of magnetization shifts from  $\hat{y}$  to  $\hat{x}$  as the voltage is increased from 0 V to 450 V and a 90° switching of the easy axis is accomplished.

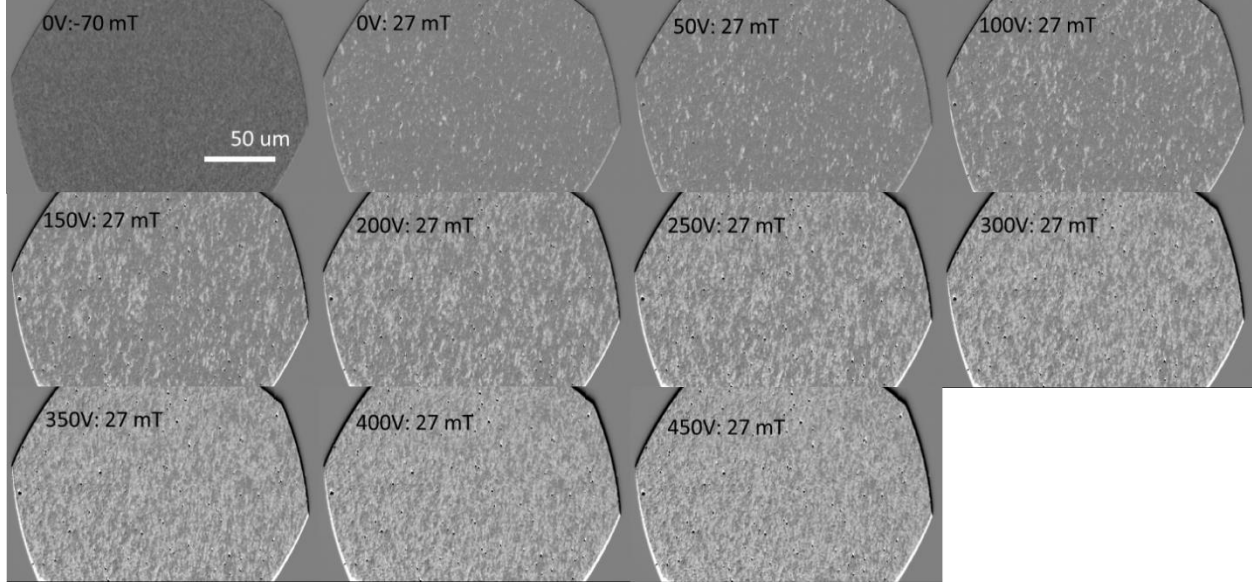


Figure 6-7 MOKE images showing domain reversal along in-plane direction  $\hat{y} // [01\bar{1}]$  for different voltages. The sample is first saturated with a -70 mT field and the reversal field is set to +27 mT, the voltage is increased in 50 V steps from 0 V to 450 V.

The magnetoelectric coefficient,  $\alpha_E = \mu_0 \frac{\Delta M}{\Delta E}$  [34] of the Bi-YIG/SiO<sub>2</sub>/PMN-PT is calculated from Fig. 6-4c (in-plane direction,  $\hat{x}$ ), where the applied field strength is 0 mT. Thus,  $\Delta M$ , is the change in the remnant magnetization,  $\Delta M_r$  and,  $\Delta E = \frac{\Delta V}{t}$ , where  $t$  is the thickness of the heterostructure,  $E$  is the electric field and  $V$  is the applied voltage across the heterostructure. The highest value of  $\alpha_E$  is determined to be  $1.05 \times 10^{-7} \text{ sm}^{-1}$  as illustrated in Fig. 6-8, which is achieved when the voltage is changed from 0 V to -50 V with a maximum change in remnant magnetization of  $\approx 8.36 \text{ kA/m}$ . This value is smaller than our previously studied yttrium substituted dysprosium iron garnet (YDyIG) which yielded a magnetoelectric coefficient of  $2.8 \times 10^{-7} \text{ sm}^{-1}$  [25]. However, the  $\alpha_E$  value is comparable to, or even surpasses, those obtained in previous studies involving magnetic oxide films on ferroelectric substrates. For instance, La<sub>0.7</sub>Sr<sub>0.3</sub> MnO<sub>3</sub>/PMN-PT exhibited  $\alpha_E = 6.4 \times 10^{-8} \text{ sm}^{-1}$  [35], while YIG/PMN-PZT achieved a maximum of  $\alpha_E = 1.8 \times 10^{-7} \text{ sm}^{-1}$  [36]. Nevertheless, this value is still lower than those recorded for metallic magnetic films such as, FeGa/PMN-PT ( $2.7 \times 10^{-6} \text{ sm}^{-1}$ ) [37], FeRh/BaTiO<sub>3</sub> ( $1.6 \times 10^{-5} \text{ sm}^{-1}$ ) [38], and Co<sub>2</sub>FeSi/PMN-PT ( $1 \times 10^{-5} \text{ sm}^{-1}$ ) [39]. Furthermore, the reported magnetoelectric coefficient exceeds that of the 4.9  $\mu\text{m}$  YIG/PMN-PT which yielded  $5.4 \times 10^{-9} \text{ sm}^{-1}$  [40], and the 10–40  $\mu\text{m}$  YIG-PZT, which reported values ranging from  $1 \times 10^{-9} \text{ sm}^{-1}$  to  $1.5 \times 10^{-9} \text{ sm}^{-1}$  [41]. Although the

reported magnetoelectric coefficient is lower than the metallic magnetic film, the damping coefficient of BiYIG is shown to be as low as  $1.3\text{-}3 \times 10^{-4}$  [42,43] compared to the damping of 0.042-0.2 for FeGa [44,45],  $2.3\text{-}3.5 \times 10^{-3}$  for FeRh [46,47] and  $8 \times 10^{-3}$  for Heusler alloy (Co<sub>2</sub>FeSi) [48]. Thus, the BiYIG with moderate magnetoelectric coefficient coupled with low damping coefficient can allow for faster magnetization dynamics and low energy dissipation in voltage controlled spintronic devices.

Finally, to check the reproducibility property of the studied heterostructure, we investigated another sample with 55 nm thickness. This sample shows similar trend in strain induced anisotropy modulation and the magnetoelectric coefficient is found to be  $0.9 \times 10^{-7} \text{sm}^{-1}$ .

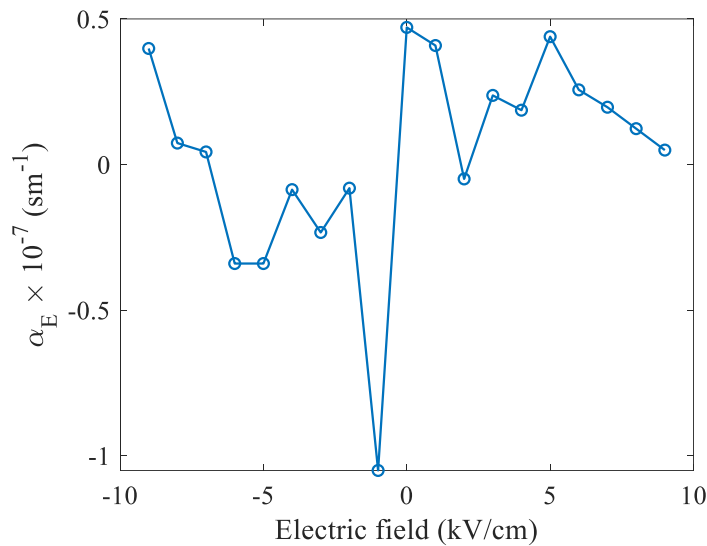


Figure 6-8 Magnetoelectric coefficient with respect to the electric field applied across the Bi-YIG/SiO<sub>2</sub>/PMN-PT heterostructure.

#### 6.4 Summary and conclusion

In summary, we have shown 90° switching of the magnetization easy axis of a multiferroic heterostructure, Bi-YIG/SiO<sub>2</sub>/PMN-PT by using voltage induced strain. The sample is fabricated using pulsed layer deposition and the ratio of Bi and YIG is optimized for lower coercivity and high saturation magnetization. An intermediate SiO<sub>2</sub> diffusion layer is deposited between PMN-PT and BiYIG for high quality crystal growth. The MOKE magnetometry and domain reversal image shows large changes in domain wall nucleation and propagation in BiYIG films with the change in electric field. Our study shows electric field control of a multiferroic structure incorporating a ferrimagnet by utilizing the magnetoelastic effect, which is significant in realizing energy efficient voltage controlled spintronic devices for storage and neuromorphic implementations. Although the magnetoelectric coefficient is moderate compared to

magnetostrictive metal/PMN-PT, a combination of moderate magnetoelectric coefficient and low damping can stimulate novel devices that use resonant effects [49-51] with high quality factors.

## References:

- [1] F. Matsukura, Y. Tokura, H. Ohno, Control of Magnetism by Electric Fields. *Nat. Nanotechnol.* 2015, 10, 209–220.
- [2] S. Bandyopadhyay, J. Atulasimha, *Nanomagnetic and Spintronic Devices for Energy-Efficient Memory and Computing*, 1st ed.; Wiley, 2016
- [3] J. C. Slonczewski, Current-driven excitation of magnetic multilayers, *J. Magn. Magn. Mater.* 159, L1 (1996).
- [4] H. Kubota, A. Fukushima, K. Yakushiji, T. Nagahama, S. Yuasa, K. Ando, H. Maehara, Y. Nagamine, K. Tsunekawa, D. D. Djayaprawira, N. Watanabe, and Y. Suzuki, Quantitative measurement of voltage dependence of spin-transfer torque in MgO
- [5] N. A. Spaldin and M. Fiebig, *Science* 309, 391 (2005).
- [6] R. Ramesh and N. A. Spaldin, *Nature Mater.* 6, 21 (2007)
- [7] N. A. Spaldin, and R. Ramesh, Advances in magnetoelectric multiferroics, *Nature Materials*, vol 18, 203–212 (2019)
- [8] J. Atulasimha, S. Bandyopadhyay, Bennett clocking of nanomagnetic logic using multiferroic single-domain nanomagnets, *Appl. Phys. Lett.* 97, 173105 (2010).
- [9] N. Hur, S. Park, P. A. Sharma, J. S. Ahn, S. Guha, S.-W. Cheong, Electric Polarization reversal and Memory in a Multiferroic Material Induced by Magnetic Fields. *Nature* 2004, 429, 392–395.
- [10] V. J. Folen, G. T. Rado, E. W. Stalder, Anisotropy of the Magnetoelectric Effect in Cr<sub>2</sub>O<sub>3</sub>. *Phys. Rev. Lett.* 1961, 6, 607–608
- [11] W. Eerenstein, M. Wiora, J. L. Prieto, J. F. Scott, N. D. Mathur, Giant Sharp and Persistent Converse Magnetoelectric Effects in Multiferroic Epitaxial Heterostructures. *Nat. Mater.* 2007, 6, 348–351.
- [12] J. T. Heron, J. L. Bosse, Q. He, Y. Gao, M. Trassin, L. Ye, J. D. Clarkson, C. Wang, J. Liu, S. Salahuddin, D. C. Ralph, D. G. Schlom, J. Iñiguez, B. D. Huey, R. Ramesh, Deterministic Switching of Ferromagnetism at Room Temperature Using an Electric Field. *Nature* 2014, 516, 370–373
- [13] N. D'Souza, M. S. Fashami, S. Bandyopadhyay, J. Atulasimha, Experimental Clocking of Nanomagnets with Strain for Ultra Low Power Boolean Logic, *Nano Letters*, vol. 16, pp. 1069–1075 (2016)

- [14] V. Sampath, N. D'Souza, D. Bhattacharya, G. M. Atkinson, S. Bandyopadhyay, and J. Atulasimha, Acoustic wave-induced magnetization switching of magnetostrictive nanomagnets from single-domain to nonvolatile vortex states, *Nano Lett.* 16, 5681 (2016).
- [15] S. Zhang, Y. G. Zhao, P. S. Li, J. J. Yang, S. Rizwan, J. X. Zhang, J. Seidel, T. L. Qu, Y. J. Yang, Z. L. Luo, Q. He, T. Zou, Q. P. Chen, J. W. Wang, L. F. Yang, Y. Sun, Y. Z. Wu, X. Xiao, X. F. Jin, J. Huang, C. Gao, X. F. Han, R. Ramesh, Electric-Field Control of Nonvolatile Magnetization in  $\text{Co}_{40}\text{Fe}_{40}\text{B}_{20}/\text{Pb}(\text{Mg}_{1/3}\text{Nb}_{2/3})_{0.7}\text{Ti}_{0.3}\text{O}_3$  Structure at Room Temperature. *Phys. Rev. Lett.* 2012, 108, 137203.
- [16] C.-G. Duan, J. P. Velev, R. F. Sabirianov, Z. Zhu, J. Chu, S. S. Jaswal, E. Y. Tsymbal, Surface Magnetoelectric Effect in Ferromagnetic Metal Films. *Phys. Rev. Lett.* 2008, 101, 137201.
- [17] U. Bauer, L. Yao, A. J. Tan, P. Agrawal, S. Emori, H. L. Tuller, S. van Dijken, G. S. D. Beach, Magneto-ionic Control of Interfacial Magnetism. *Nat. Mater.* 2015, 14, 174–181
- [18] J.-M. Hu, C.-G. Duan, C.-W. Nan, L.-Q. Chen, Understanding and Designing Magnetoelectric Heterostructures Guided by Computation: Progresses, Remaining Questions, and Perspectives. *npj Comput. Mater.* 2017, 3, 18
- [19] S. Zhang, F. Li, High performance ferroelectric relaxor-PbTiO<sub>3</sub> single crystals: Status and perspective. *J. Appl. Phys.* 111, 031301 (2012).
- [20] S. Lindemann, J. Irwin, G.-Y. Kim, B. Wang, K. Eom, J. Wang, J. Hu, L.-Q. Chen, S.-Y. Choi, C.-B. Eom, M. S. Rzechowski, Low-voltage magnetoelectric coupling in membrane heterostructures, *Sci. Adv.* 7, eabh2294 (2021)
- [21] Z. Zhao, M. Jamali, N. D'Souza, D. Zhang, S. Bandyopadhyay, J. Atulasimha, and J.-P. Wang, Giant voltage manipulation of MgO-based magnetic tunnel junctions via localized anisotropic strain: A potential pathway to ultra-energy-efficient memory technology, *Appl. Phys. Lett.* 109, 092403 (2016)
- [22] A. Begue and M. Ciria, Strain-Mediated Giant Magnetoelectric Coupling in a Crystalline Multiferroic Heterostructure, *ACS Appl. Mater. Interfaces* 2021, 13, 6778–6784
- [23] A. Biswas, H. Ahmad, J. Atulasimha, and S. Bandyopadhyay, “Experimental demonstration of complete 180° reversal of magnetization in isolated Co nanomagnets on a PMN-PT substrate with voltage generated strain”, *Nano Letters*, 17 (6), 3478–3484, 2017.
- [24] J. J. Bauer, E. R. Rosenberg, S. Kundu, K. A. Mkhoyan, P. Quarterman, A. J. Grutter, B. J. Kirby, J. A. Borchers, and C. A. Ross, Dysprosium Iron Garnet Thin Films with Perpendicular Magnetic Anisotropy on Silicon, *Adv. Electron. Mater.* 2020, 6, 1900820
- [25] M. J. Gross, W. A. Misba, K. Hayashi, D. Bhattacharya, D. B. Gopman, J. Atulasimha, C. A. Ross, Voltage modulated magnetic anisotropy of rare earth iron garnet thin films on a piezoelectric substrate, *Appl. Phys. Lett.* 121, 252401 (2022)

- [26] T. Fakhru, S. Tazlaru, L. Beran, Y. Zhang, M. Veis, C. A. Ross, Magneto-Optical Bi:YIG Films with High Figure of Merit for Nonreciprocal Photonics, *Adv. Optical Mater.* 7, 1900056 (2019)
- [27]. Y. Fan, M. J. Gross, T. Fakhru, J. Finley, J. T. Hou, S. Ngo, L. Liu and C. A. Ross, Coherent magnon-induced domain-wall motion in a magnetic insulator channel, *Nature Nanotechnology* volume 18, pages1000–1004 (2023)
- [28] Y. C. Wu, K. Garello, W. Kim, M. Gupta, M. Perumkunnil, V. Kateel, S. Couet, R. Carpenter, S. Rao, S. Van Beek, K. K. Vudya Sethu, F. Yasin, D. Crotti, and G. S. Kar, Voltage-Gate Assisted Spin-Orbit Torque Magnetic Random Access Memory for High-Density and Low-Power Embedded Application, *Physical Review Applied* 15, 064015 (2021).
- [29] M. A. Azam, D. Bhattacharya, D. Querlioz, C. A. Ross, J. Atulasimha, Voltage control of domain walls in magnetic nanowires for energy-efficient neuromorphic devices, *Nanotechnology* **31** 145201, 2020.
- [30] W. A. Misba, M. Lozano, D. Querlioz, J. Atulasimha, Energy Efficient Learning with Low Resolution Stochastic Domain Wall Synapse Based Deep Neural Networks, *IEEE Access*, 10, 84946, 2022
- [31] T. Wu, P. Zhao, M. Bao, A. Bur, J. L. Hockel, K. Wong, K. P. Mohanchandra, C. S. Lynch, and G. P. Carman, Domain engineered switchable strain states in ferroelectric (011)  $[\text{Pb}(\text{Mg}_{1/3}\text{Nb}_{2/3})\text{O}_3](1-x)-[\text{PbTiO}_3]x$  (PMN-PT,  $x \approx 0.32$ ) single crystals, *J. Appl. Phys.* 109, 124101 (2011)
- [32] P. Hansen, K. Witter, and W. Tolksdorf, Magnetic and magneto-optic properties of lead- and bismuth-substituted yttrium iron garnet films, *Physical Review B*, vol. 27, 11 (1983)
- [33] Y. Lin, L. Jin, H. Zhang, Z. Zhong, Q. Yang, Y. Rao, M. Li, Bi-YIG ferrimagnetic insulator nanometer films with large perpendicular magnetic anisotropy and narrow ferromagnetic resonance linewidth, *Journal of Magnetism and Magnetic Materials* 496, 165886 (2020)
- [34] Y. Zhang, Z. Wang, Y. Wang, C. Luo, J. Li, and D. Viehland, Electric-field induced strain modulation of magnetization in Fe-Ga/Pb(Mg<sub>1/3</sub>Nb<sub>2/3</sub>)-PbTiO<sub>3</sub> magnetoelectric heterostructures, *Journal of Applied Physics* 115, 084101 (2014)
- [35] D. Pesquera, E. Khestanova, M. Ghidini, S. Zhang, A. P. Rooney, F. Maccherozzi, P. Riego, S. Farokhipoor, J. Kim, X. Moya, M. E. Vickers, N. A. Stelmashenko, S. J. Haigh, S. S. Dhesi, and N. D. Mathur, Large magnetoelectric coupling in multiferroic oxide heterostructures assembled via epitaxial lift-off, *Nat. Commun.* 11, 3190 (2020).
- [36] H. Liuyang, P. Freddy, R. Denis, L. Tuami, T. Nicolas, W. Genshui, and P. Philippe, *Ferroelectrics* 557, 1 (2020).
- [37] W. Jahjah, J. P. Jay, Y. le Grand, A. Fessant, A. R. E. Prinsloo, C. J. Sheppard, D. T. Dekadjevi, and D. Spenato, Electrical Manipulation of Magnetic Anisotropy in a Fe<sub>81</sub>Ga<sub>19</sub>/Pb(Mg<sub>1/3</sub>Nb<sub>2/3</sub>)O<sub>3</sub>-Pb(ZrxTi<sub>1-x</sub>)O<sub>3</sub> Magnetoelectric Multiferroic Composite, *Phys. Rev. Appl.* 13, 034015 (2020).

- [38] R. O. Cherifi, V. Ivanovskaya, L. C. Phillips, A. Zobelli, I. C. Infante, E. Jacquet, V. Garcia, S. Fusil, P. R. Briddon, N. Guiblin, A. Mougin, A. A. Unal, F. € Kronast, S. Valencia, B. Dkhil, A. Barthelemy, and M. Bibes, Electric-field control of magnetic order above room temperature, *Nat. Mater.* 13, 345 (2014).
- [39] S. Fujii, T. Usami, Y. Shiratsuchi, A. M. Kerrigan, A. M. Yatmeidhy, S. Yamada, T. Kanashima, R. Nakatani, V. K. Lazarov, T. Oguchi, Y. Gohda, and K. Hamaya, Strain-induced specific orbital control in a Heusler alloy-based interfacial multiferroics, *NPG Asia Mater.* 14, 43 (2022).
- [40] G. Srinivasan, M. I. Bichurin, and J. V. Mantese, Ferromagnetic-ferroelectric layered structures: magnetoelectric interactions and devices, *Integrated Ferroelectrics* 71, 45 (2005).
- [41] A. S. Tatarenko, V. Gheevarghese, G. Srinivasan, O. V Antonenkov, and M. I. Bichurin, Microwave magnetoelectric effects in ferrite—piezoelectric composites and dual electric and magnetic field tunable filters, *J. Electroceram.* 24, 5 (2010).
- [42] T. Fakhrul, B. Khurana, H. T. Nembach, J. M. Shaw, Y. Fan, G. A. Riley, L. Liu, and C. A. Ross, Substrate-Dependent Anisotropy and Damping in Epitaxial Bismuth Yttrium Iron Garnet Thin Films, *Adv. Mater. Interfaces* 2023, 10, 2300217
- [43] L. Soumah, N. Beaulieu, L. Qassym, C. Carrétéro, E. Jacquet, R. Lebourgeois, J. Ben Youssef, P. Bortolotti, V. Cros and A. Anane, Ultra-low damping insulating magnetic thin films get perpendicular, *Nature Communications*, vol. 9, 3355 (2018)
- [44] D. B. Gopman, V. Sampath, H. Ahmad, S. Bandyopadhyay, J. Atulasimha, Static and dynamic magnetic properties of sputtered Fe–Ga thin films, *IEEE Transactions on Magnetics*, vol. 53, no. 11 (2017)
- [45] D. Cao, X. Cheng, L. Pan, H. Feng, C. Zhao, Z. Zhu, Q. Li, J. Xu, S. Li, Q. Liu, and J. Wang, Tuning high frequency magnetic properties and damping of FeGa, FeGaN and FeGaB thin films, *AIP Advances* 7, 115009 (2017)
- [46] Y. Wang, M. M. Decker, T. N. G. Meier, X. Chen, C. Song, T. Grünbaum, W. Zhao, J. Zhang, L. Chen and C. H. Back, Spin pumping during the antiferromagnetic–ferromagnetic phase transition of iron–rhodium, *Nature Communications*, vol. 11, 275 (2020)
- [47] S. Mankovsky, S. Polesya, K. Chadova, H. Ebert, J. B. Staunton, T. Gruenbaum, M. A. W. Schoen, C. H. Back, X. Z. Chen, and C. Song, Temperature-dependent transport properties of FeRh, *Phys. Rev. B* 95, 155139(2017)
- [48] M. Oogane, R. Yilgin, M. Shinano, S. Yakata, Y. Sakuraba, Y. Ando, T. Miyazaki, Magnetic damping constant of Co<sub>2</sub>FeSi Heusler alloy thin film, *J. Appl. Phys.* 101, 09J501 (2007)
- [49] A. Roe, D. Bhattacharya, J. Atulasimha, Resonant acoustic wave assisted spin-transfer-torque switching of nanomagnets, *Appl. Phys. Lett.* 115, 112405 (2019)

- [50] W. A. Misba, M. M. Rajib, D. Bhattacharya, J. Atulasimha, Acoustic-wave-induced ferromagnetic-resonance-assisted spin-torque switching of perpendicular magnetic tunnel junctions with anisotropy variation, *Phys. Rev. Applied* 14, 014088 (2020)
- [51] W.-G. Yanga and H. Schmidt, Acoustic control of magnetism toward energy-efficient applications, *Appl. Phys. Rev.* 8, 021304 (2021)

## Chapter 7: Interfacial Exchange and Magnetostatic Coupling in a CoFeB/Perpendicular Ferrimagnetic Thulium Iron Garnet Heterostructure

In Chapter 6, we demonstrated the electric field tunability of heterostructures incorporating Bi-YIG thin films deposited on a piezoelectric substrate. Similar to Bi-YIG, rare earth iron garnets (REIGs) grown on piezoelectric substrates exhibit a low damping constant. Additionally, REIGs display perpendicular magnetic anisotropy (PMA), making them a highly attractive candidate for implementing racetrack memory devices, offering extremely efficient and rapid control of domain walls. However, due to the insulating nature of the iron garnet films, its incorporation to a magnetic tunnel junction (MTJs) is challenging. Coupling the iron garnet films to a ferromagnetic metal can address the issue.

In this chapter, we investigate the exchange coupling between a ferrimagnetic insulator (FI) Thulium Iron Garnet (TmIG) deposited on a gadolinium gallium garnet (GGG) substrate, which shows perpendicular magnetic anisotropy (PMA) of magnetoelastic origin and a ferromagnetic metal (FM) stack with oxide capping that consists of CoFeB( $x$ )/W(0.4 nm)/CoFeB(0.8 nm)/MgO(1 nm)/W(5 nm). Detailed vibrating sample magnetometry (VSM), magneto optical Kerr microscopy (MOKE) and first order reversal curves (FORCs) studies, coupled with micromagnetic simulation are used to analyze the extent of the coupling between these layers. Strong interlayer exchange coupling and magnetostatic coupling are observed in the samples where the relative strength between these interactions can be controlled by varying the thickness of the CoFeB layer. CoFeB with thickness  $x \leq 1$  nm is found to be strongly exchange coupled whereas the magnetostatic coupling dominates when the thickness  $x \geq 3$  nm. These findings have important implications towards realizing fast and energy efficient spintronic devices using a ferrimagnetic insulator, while its coupling to the ferromagnetic layer can be used for effective electrical read out of the magnetic state.

Rare earth iron garnets ( $\text{RE}_3\text{Fe}_5\text{O}_{12}$ ) including thulium, terbium, samarium, and other REIG films with perpendicular magnetic anisotropy (PMA), have been developed for spintronic applications [1,2]. Heterostructures such as Pt/TmIG and Pt/Bi:YIG have demonstrated spin orbit torque switching [1,3,4], chiral spin textures [3,5,6], and a relativistic domain wall velocity approaching the magnon group velocity [7], making these materials promising for memory or logic devices. Specifically, REIG materials exhibit ferrimagnetic characteristics due to the antiparallel superexchange coupling between the tetrahedral and octahedral Fe sublattices [8]. As mentioned earlier, the insulating nature of the REIG precludes its incorporation into a magnetic tunnel junction. Coupling the REIG to a magnetic metal would enable readout of the magnetization of the REIG, if the metal formed the free layer of a MTJ. FM metal/REIG multilayers



could also be useful in magnon-based spintronics for high frequency application and low power computing. This motivates study of the strength and mechanism of exchange coupling between thin films of magnetic metals and REIGs.

The ground states of a magnetic multilayers are dictated by the competition and interplay of long range magnetostatic or dipole interaction, short- range exchange interaction and local anisotropy variation that are often determined by the composition and thickness of the individual layers. Previously, exchange-coupled ferromagnetic multilayers in direct contact [9-11] and separated by non-magnetic spacer [12-16] and hybrid spacer [17] were studied. Dynamic coupling between the ferrimagnetic insulator (FI) and ferromagnetic metal (FM) in heterostructures due to exchange of non-equilibrium spin currents are identified by investigating broadband ferromagnetic resonance and spin torque ferromagnetic resonance [18-20]. However, few studies are performed on static interlayer exchange coupling between FI/FM heterostructures [21-24]. It has been reported that ion sputtered Fe of 5 nm to 10 nm on 100 nm thick YIG can replicate the stripe domain wall pattern of YIG due to exchange coupling [22]. In addition, static and dynamic magnetization response is characterized in heterostructures of 0.5  $\mu\text{m}$  thick  $(\text{YBiLu})_3(\text{FeAl})_5\text{O}_{12}$  and 30 nm thick permalloy film [23]. In this system, strong exchange coupling is attributed to the increased domain wall periodicity after permalloy deposition and no domain wall imprinting is reported. In another study, a 2 nm Co thin film is deposited on different  $\mu\text{m}$  scale thick YIG, which shows change in domain wall geometry of YIG films suggesting magnetostatic coupling [24].

In this study, we chose Thulium Iron Garnet (TmIG) as the FI layer. When deposited on a gadolinium gallium garnet (GGG) substrate, TmIG shows perpendicular magnetic anisotropy (PMA) of magnetoelastic origin. To make the FI/FM heterostructure, an FM compound consisting of  $\text{CoFeB}(x)/\text{W}$  (0.4 nm)/ $\text{CoFeB}$ (0.8 nm)/ $\text{MgO}$ (1 nm)/ $\text{W}$ (5 nm) is deposited on top of the FI using ion beam sputter deposition method. The thickness of  $\text{CoFeB}$  that is in direct contact with TmIG is varied to investigate the extent of exchange and magnetostatic interaction. Specifically, we chose TmIG due to its soft magnetic properties, which is useful for magnetic memory devices and  $\text{CoFeB}$  as it has been a standalone choice in current MRAM technologies. Fig. 7-1(a) shows the heterostructures investigated, where the magnetic ground state can be preferentially oriented along out of plane or in-plane depending on the thickness of the  $\text{CoFeB}$  overlayer. Detailed magnetometry analysis using vibrating sample magnetometry (VSM), and magnetic imaging with magneto-optical Kerr microscopy (MOKE) is performed to investigate the coupling phenomena. First order reversal curves (FORCs) are obtained to analyze the magnetic interactions in the heterostructure layers and detailed micromagnetic simulations are performed to quantify the extent of competing interactions from the inter-layer exchange and magnetostatic interaction.

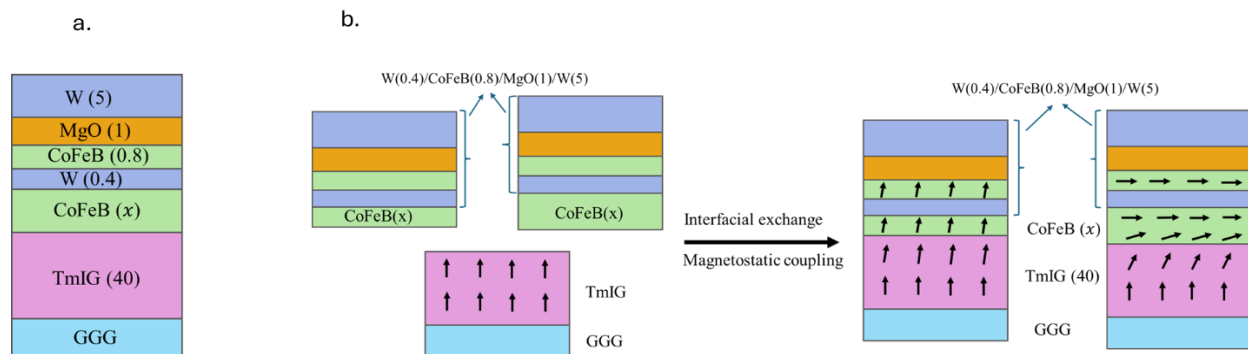


Figure 7-1 a. The prepared FM stack CoFeB(x)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5) are deposited on FI stack GGG/TmIG(40) to investigate FI/FM coupling. All the numbers in the parenthesis are thickness in nm. b. The FM stacks with CoFeB of  $x=1$  nm and lower thickness have magnetization predominantly oriented along the out of plane direction whereas in stacks with  $x=3$  nm and higher thickness the magnetizations are predominantly along in-plane. The resulting magnetizations in FI/FM stacks are canted with respect to out of plane and in-plane due to interfacial exchange coupling and magnetostatic interaction.

## 7.1 Sample preparation and characterization

The heterostructure samples used in the experiments were grown by our collaborators at MIT and NIST. A 40 nm thick TmIG was deposited on top of the GGG (111) substrate using a pulsed laser deposition (PLD) system with base pressure  $\leq 7 \times 10^{-4}$  Pa ( $5 \times 10^{-6}$  Torr). After loading the substrates, a 20 Pa (150 mTorr)  $O_2$  partial pressure environment is maintained through continuous  $O_2$  flow during substrate temperature ramp-up to 650 °C, during the deposition, and while cooling down post-deposition. A 248 nm KrF excimer laser focused to a fluence of 2 J/cm<sup>2</sup> on a stoichiometric TmIG target at a frequency of 10 Hz.

The GGG/TmIG samples are transferred to a 12-target, 200 mm ion beam sputtering cluster with base pressure  $\leq 3 \times 10^{-6}$  Pa ( $2 \times 10^{-8}$  Torr). In this chamber, the sample structure of CoFeB(x)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5), (the number in parentheses reflects thicknesses in nm) is grown in an Ar working pressure of 0.03 Pa ( $2 \times 10^{-4}$  Torr). Using a Co<sub>20</sub>Fe<sub>60</sub>B<sub>20</sub> stoichiometric target, a thickness series where  $x$  ranges across the nominal thicknesses: 0.84 nm, 1 nm, 3 nm and 4 nm. A control sample for each CoFeB thickness was deposited on Si substrate with a 300 nm thick thermally oxidized Si overlayer simultaneously with the TmIG/GGG samples. All samples were processed in a rapid thermal annealer under vacuum at 300 °C for 10 min. The W spacer layer is employed as a boron sink during the annealing and to achieve high perpendicular magnetic anisotropy in the CoFeB layer [25].

The control samples are characterized by measuring the hysteresis loops for in-plane and out-of-plane external fields using vibrating sample magnetometry (VSM) as shown in Fig. 7-2. The layer sequences of the samples and corresponding saturation magnetization, easy axis direction and the coercive field along

the easy axis directions are presented in Table 7-1. The uncertainty while estimating the coercive field is  $\sim 10\%$  and comes from measurement uncertainty and fitting error. For  $x = 1$  nm and below, the FM stack shows a predominantly out of plane easy axis. On the other hand, samples with  $x = 3$  nm or higher, the magnetic easy axis is in-plane. Independent of the lower CoFeB layer of variable thickness, the fixed 0.8 nm CoFeB layer above the few atomic layers of W, which is in contact with a top MgO, is expected to have perpendicular magnetic anisotropy (PMA) [26]. Furthermore, strong ferromagnetic exchange coupling is expected between these two CoFeB layers, which may further influence both of their magnetization states. When the variable CoFeB layer thickness is 1 nm or lower, PMA from the MgO as well exchange coupling from the top PMA CoFeB layer can force the magnetization to orient along perpendicular direction. For CoFeB of  $x \geq 3$  nm, the sample magnetization remains in-plane due to dominant magnetostatic effects. In this case, the surface anisotropy from MgO/CoFeB is not enough to orient the overall magnetization along the perpendicular direction.

Table 7-1: Magnetic properties of control samples

Layer sequence	Saturation magnetization (kA/m)	Easy axis	Coercive field (mT) along easy axis
Si/CoFeB(4)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5)	$970 \pm 7.5$	In-plane	0.33
Si/CoFeB(3)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5)	$940 \pm 5$	In-plane	0.25
Si/CoFeB(1)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5)	$545 \pm 9$	Out-of-plane	0.52
Si/CoFeB(0.84)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5)	$365 \pm 11$	Out-of-plane	1.54

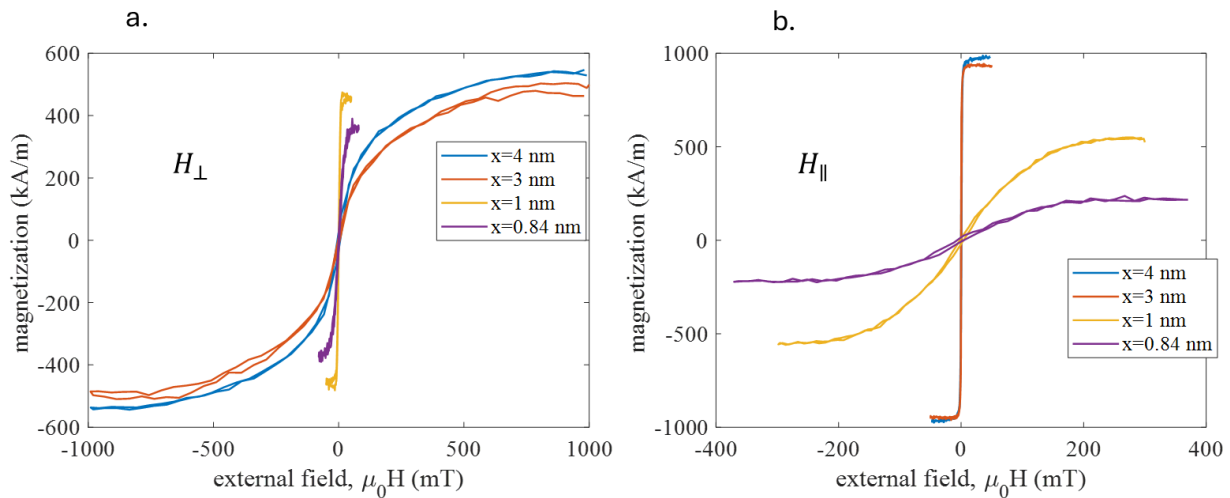


Figure 7-2 a. Out of plane and b. in-plane hysteresis loops for control samples, Si/CoFeB(x)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5) for variable thickness, x of CoFeB.

No appearance of multiple loops and or zero moment crossing is noted in the hysteresis loops, which suggests substantial ferromagnetic coupling between the CoFeB layers. The saturation magnetization of thickest,  $x = 4$  nm sample is measured to be  $970 \pm 7.5$  kA/m, which decreases with CoFeB layer thickness, and estimated to be  $365 \pm 11$  kA/m for the thinnest,  $x = 0.84$  nm sample. The out of plane coercivity of the CoFeB samples for  $x = 1$  nm and  $x = 0.84$  nm are estimated to be 0.52 mT and 1.54 mT respectively.

Next, the magnetic hysteresis loops of the FI/FM heterostructure are obtained for variable CoFeB thickness using the VSM measurements and shown in Fig. 7-3(a)-(d). The layer sequences of the FI-only stack and the FI/FM stacks along with the measured saturation magnetization, easy axis directions and the coercive field along the easy axis directions are presented in Table 7-2. Two distinct trends are seen between the samples with CoFeB thickness of  $x = 1$  nm or below (Fig. 7-3(a)-(b)) and 3 nm or above (Fig. 7-3(d)-(e)). For comparison, the hysteresis loops of pristine GGG/TmIG (FI only stack) are also presented. The out-of-plane coercivity of the FI /FM increases to 2 mT for low thickness CoFeB ( $x = 1$  nm or below), compared to the coercivity of the pristine FI of 0.22 mT and control FM stack which is 0.52 mT for  $x = 1$  nm and 1.54 mT for  $x = 0.84$  nm. This shows that the exchange coupling between the two layers modifies their PMA. However, the coupled samples do experience domain wall nucleation/annihilation during reversal (confirmed by FORC later in section III) and characterized by the more gradual transition towards saturation in Fig. 7-3(a). This suggests static exchange coupling between the FI and FM stacks. From in-plane hysteresis loops in Fig. 7-3(b) for samples with  $x \leq 1$  nm, we see that the loop shapes are less square, and the remanent magnetizations are much lower compared to the out of plane loops in Fig. 7-3(a). Also, the FI/FM loops are more canted, and the resulting magnetization saturates at higher fields,  $\mu_0 H_{\parallel} > 130$  mT compared to the FI-only case where the magnetization saturates at 100 mT. This suggests an increase in hard axis anisotropy and PMA of the FI/FM stack compared to the FI-only stack due to the exchange coupling between FI and FM stacks. Perpendicular magnetic anisotropy of the FI-stack, control FM stack and the FI/FM heterostructure can be estimated from the hard axis anisotropy,  $H_k$  obtained from the in-plane hysteresis loops [27] (see supplementary of Ref. 27). Specifically, the effective anisotropy,  $K_{u,eff} = H_k M_{s,FM} / 2$  is a measurable quantity as the  $H_k$  and  $M_s$  are known from VSM. The PMA coefficient,  $K_u$  can be estimated using the expression,  $K_u = K_{u,eff} + \frac{1}{2} \mu_0 M_s^2$ . From Fig. 7-2(b), for  $x=1$  nm the PMA energy of the FM stack is estimated to be  $244 \pm 25$  kJ/m<sup>3</sup>. Using Fig. 7-3(b), the PMA of FI-only stack and FI/FM are estimated to be  $7.8 \pm 0.8$  kJ/m<sup>3</sup> and  $10.74 \pm 1.1$  kJ/m<sup>3</sup> respectively. Thus, the exchange coupling in FI/FM heterostructure results in a PMA that is in between the lower PMA of TmIG and the higher PMA of CoFeB (for  $x \leq 1$  nm).

Table 7-2. Magnetic properties of the FI-only stack and FI/FM heterostructures

Layer sequence	Saturation magnetization (kA/m)	Easy axis	Coercive field (mT) along easy axis
GGG/TmIG(40)	$82 \pm 1.6$	Out-of-plane	0.22
GGG/TmIG(40)/CoFeB(0.84)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5)	$91.5 \pm 1.7$	Out-of-plane	2.03
GGG/TmIG(40)/CoFeB(1)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5)	$95 \pm 1.4$	Out-of-plane	1.81
GGG/TmIG(40)/CoFeB(3)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5)	$143 \pm 2.5$	In-plane	9.60
GGG/TmIG(40)/CoFeB(4)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5)	$169 \pm 2.31$	In-plane	3.35

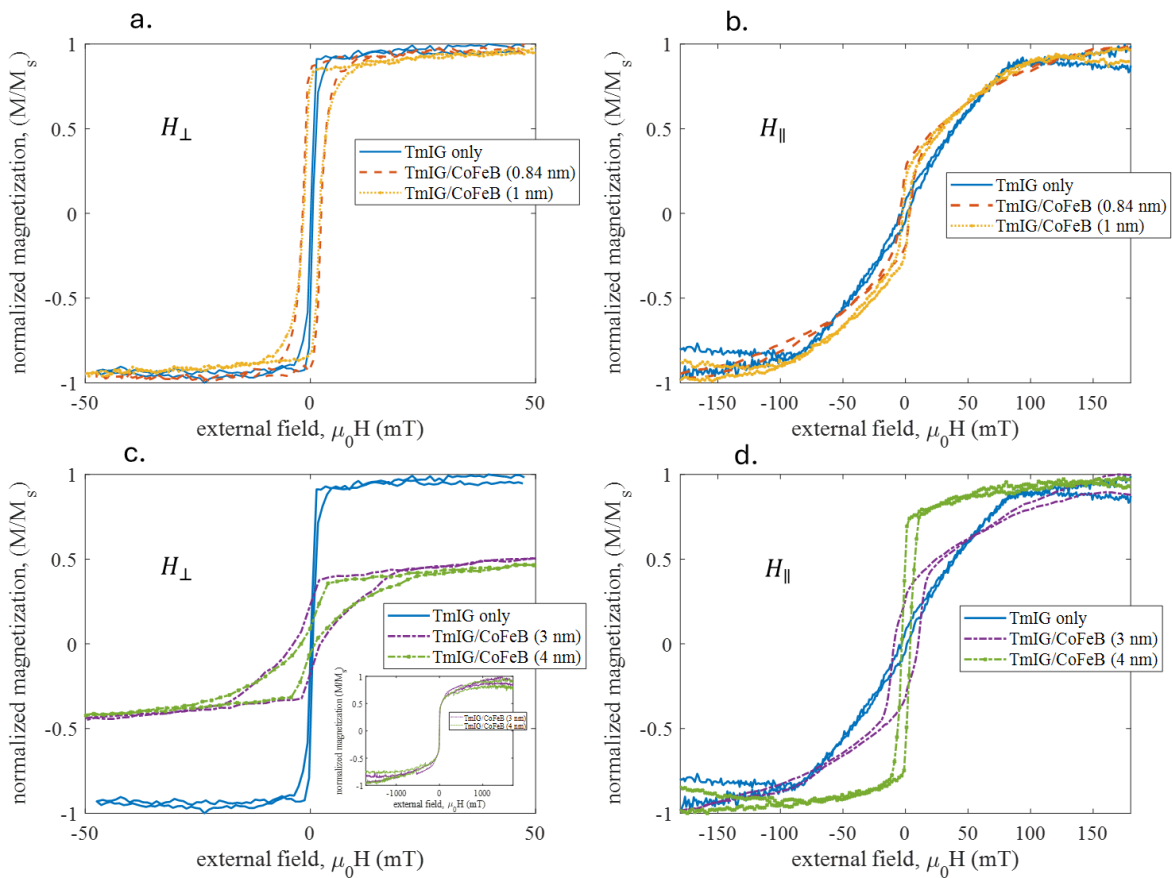


Figure 7-3 a. Out of plane and b. in-plane hysteresis loops of FI/FM heterostructure samples, GGG/TmIG(40)/CoFeB(x)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5), with CoFeB thickness,  $x \leq 1$  nm. c. out of plane and d. in-plane hysteresis loops with CoFeB thickness  $x \geq 3$  nm. The out of plane and in-plane hysteresis loops of the pristine GGG/TmIG(40) sample are presented for comparison.

As we increase the CoFeB thickness, such as  $x \geq 3\text{nm}$ , the out of plane loops become more canted and the CoFeB layer magnetization is now forced to the perpendicular direction in a reversible manner as an increasing portion of the CoFeB film (which is beyond the interlayer exchange length, estimated later using micromagnetic simulation) is not strongly coupled to the PMA TmIG and tends to be in-plane. Indeed, a large saturation field,  $\mu_0 H_{\perp} > 1000\text{ mT}$  (see inset of Fig. 7-3(c)) is required to saturate the heterostructure along the perpendicular direction. Furthermore, the in-plane loops of the FI/FM become squarer, particularly for larger thickness,  $x = 4\text{ nm}$  CoFeB. Also, the coercive field of the FI/FM along in-plane direction increases to 3.35 mT and 9.60 mT compared to the control FM stacks with  $x = 4\text{ nm}$  and  $x = 3\text{ nm}$ . This suggests, in thick CoFeB heterostructure ( $x \geq 3\text{nm}$ ) the exchange interaction from perpendicular TmIG bottom layer is not strong enough to orient the entire CoFeB magnetization along perpendicular direction and exhibit PMA.

Thus, investigation on the magnetometry loops of different FI/FM samples with varying CoFeB thickness suggests the samples are coupled and switched together, however, the strength of exchange interactions varies with CoFeB thickness.

## 7.2. Interlayer exchange and magnetostatic coupling

Polar magneto-optical Kerr microscopy (MOKE) is carried out to observe the domain pattern for the heterostructure sample during the switching process to investigate the extent of the coupling in the FI/FM heterostructures. The representative MOKE images of the samples are shown in Fig. 7-4 for external fields applied along the out-of-plane direction. The samples are first demagnetized, and a reference image is taken at 50 mT amplitude and 5 Hz ac applied field. All the subsequent images are then subtracted from the reference image for each of the samples. Then the samples are saturated by applying a positive 50 mT external field and the images are taken by reducing the fields from positive saturation. Clearly distinguishable and regularly spaced labyrinth domain patterns are observed for the TmIG sample (FI-only stack) when the applied field is reduced to -1mT. Domains with black contrast are seen to be nucleated in Fig. 7-4, which expands further with the decreasing field and saturates the sample around -10 mT. To extract the domain wall periodicity, the radial FFT intensity profile of the images is measured as shown in Fig. 7-5 where the peak location of the Gaussian FFT intensity denotes the most predominant domain wall periodicity and approximately the mean domain wall periodicity (as the distribution is almost symmetric, i.e. the distribution is almost Gaussian). The domain wall periodicity of the TmIG only sample is estimated to be 4.90  $\mu\text{m}$ . For TmIG/CoFeB ( $x$ ) with  $x=1\text{ nm}$ , significant domain wall nucleation is not observed up to -2 mT which suggests increased coercivity stemmed from strong interlayer exchange coupling and subsequent pinning of TmIG magnetization to the CoFeB magnetization. Around -3 mT, dendritic domain patterns with black contrast are observed, which nearly overtake the regions of the sample with white

domains. The black domains expand quickly and saturate the regions around -11 mT applied field similar to what we observe for the TmIG only sample. The periodicity of the domain wall reduces to  $3.2 \mu\text{m}$  for 1 nm CoFeB overlayer, which suggests increased resultant magnetic moment of the sample compared to TmIG sample. We note that the sample with  $x = 0.84 \text{ nm}$  CoFeB behaves similarly to that of  $x = 1 \text{ nm}$  CoFeB sample.

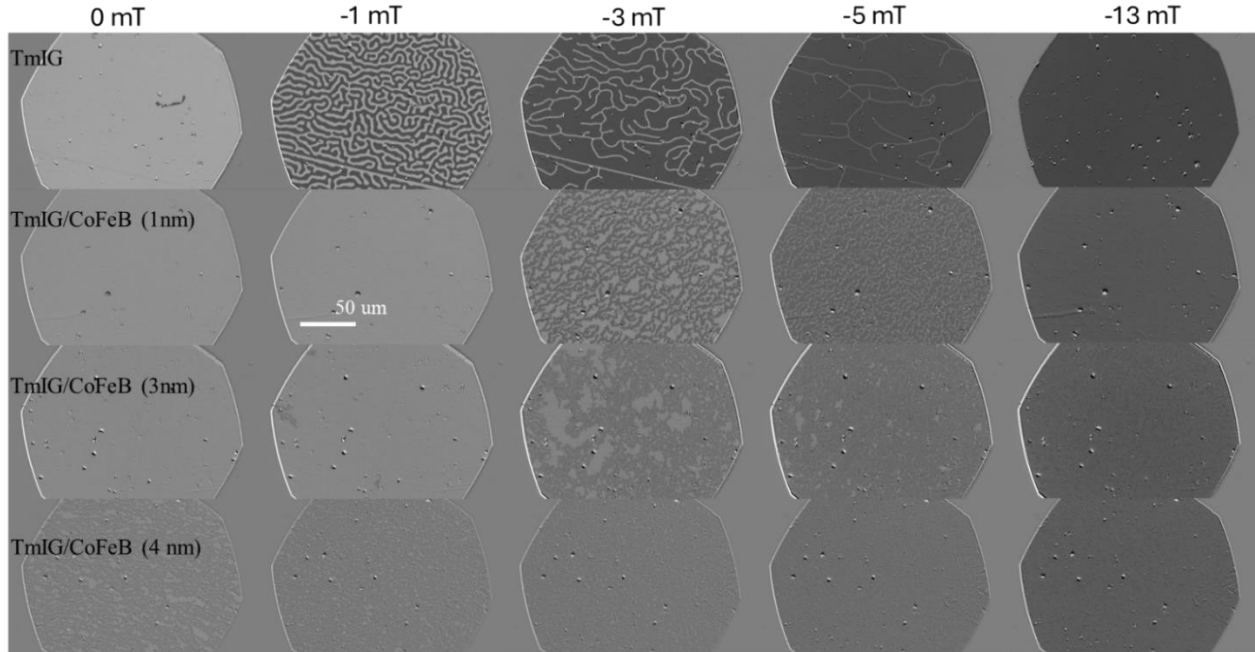


Figure 7-4 Polar MOKE magnetometry for FI only stack and FI/FM samples where the CoFeB thickness is varied to  $x = 1 \text{ nm}$ ,  $3 \text{ nm}$  and  $4 \text{ nm}$ . All the samples are first saturated to  $+50 \text{ mT}$  and then the out of plane external fields are decreased. The snapshots are taken at fields of  $0 \text{ mT}$ ,  $-1 \text{ mT}$ ,  $-3 \text{ mT}$ ,  $-5 \text{ mT}$  and  $-13 \text{ mT}$ . Labyrinthine domains with regular periodicity are observed for FI only stack, GGG/TmIG (40). Labyrinthine domains with dendritic patterns are observed with CoFeB overlayers. The periodicity of the domain wall periodicity decreases with increased CoFeB overlayer thickness implying higher magnetostatic coupling.

The periodicity of the dendritic domain decreases with the increase of the CoFeB layer thickness. For TmIG/CoFeB ( $x$ ) with  $x = 4 \text{ nm}$ , the domain wall starts to nucleate well before  $0 \text{ mT}$  suggesting strong magnetostatic interaction accompanied by large increase in the resulting magnetic moment of the heterostructure. At  $0 \text{ mT}$ , domain walls with multiple periodicities are observed as seen from the mostly flattened gaussian distribution with one barely distinguishable peak around  $0.93 \mu\text{m}$ . Although the reversed black contrast domains are seen to nucleate early for the thicker samples ( $x \geq 3 \text{ nm}$ ), the domains expand gradually, and a large field is needed to saturate the film (around  $-25 \text{ mT}$  not shown in the images). Contrary to the TmIG only and heterostructure samples with thin CoFeB, in thicker CoFeB heterostructures tiny sized domains with white contrast can be still seen in the images at  $-13 \text{ mT}$ , which suggests dominant magnetostatic coupling in addition to the exchange coupling. As seen in the magnetometry loop and in the

FORC analysis later, once opposite domains are nucleated in the TmIG, it is significantly hard to saturate those domains due to in-plane magnetization component of the CoFeB which facilitates flux closure of anti-parallel TmIG domains, Fig. 7-7(c).

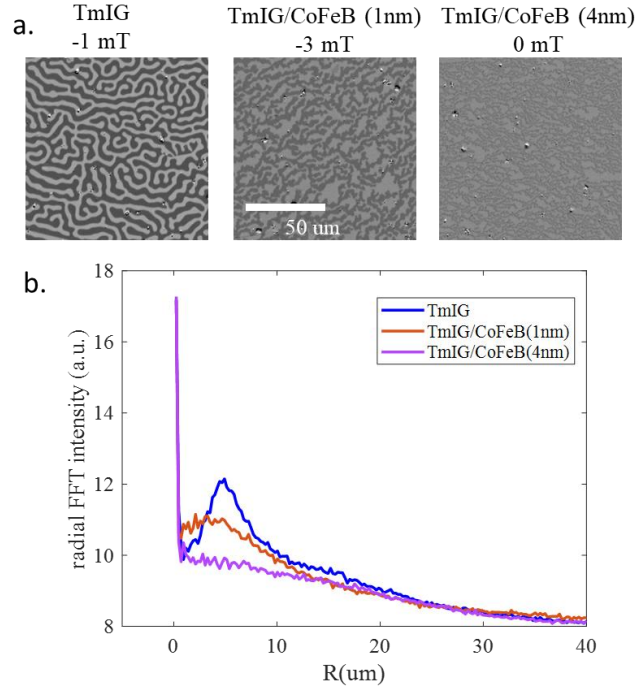


Figure 7-5 a. Domain wall pattern for TmIG, TmIG/CoFeB heterostructures with 1 nm and 4 nm variable CoFeB layer thickness at the corresponding fields. b. Radial FFT intensity to determine the periodicity of the domain wall for the respective samples.

The domain wall periodicity values serve as an important indicator to the competition between exchange interaction and magnetostatic coupling. Large exchange prefers a smaller number of domain walls due to the increased energy penalty for creating the domain walls. In contrast, increased dipolar coupling favors a greater number of domain walls and oppositely oriented lateral regions to minimize the dipolar energy. As a result, with increasing dipolar contribution, the domain wall periodicity decreases. The stripe domain width ( $L_s$ ), within the limit when the film thickness is much smaller than the domain period, can be found from the analytical model reported by Kaplan and Gehring [28],  $L_s = t e^{\left(\frac{\pi a}{2} + 1\right)} e^{\frac{\pi \sigma_w}{\mu_0 M_s^2 t}}$ , where  $t$  is the film thickness,  $M_s$  is the saturation magnetization and  $\sigma_w$  is the domain wall energy,  $a$  is the model dependent parameter with a value of -0.666. As the film thickness and saturation magnetization increases the model predicts lower value of domain wall periodicity. In our studied samples, the domain wall periodicity is the lowest in the heterostructure with thick CoFeB,  $x = 4$  nm. For  $x=1$  nm and 0.84 nm samples, the resulting saturation magnetization is smaller and the domain wall periodicity is higher than the samples with  $x=3$  nm and  $x = 4$  nm.



Thus, MOKE microscopy images of domain wall patterns show inserting a CoFeB layer alters the domain wall periodicity and geometry of the heterostructures which suggests coupling of TmIG magnetization.

Next, we carried out first order reversal curves (FORCs) experiments [29-34] to investigate further the interlayer exchange coupling that exists between the FI and FM layers. The FORC analysis is performed in collaboration with our collaborator in Georgetown University. VSM is used to measure the FORCs in the following manner. After driving the sample to positive saturation, the magnetic field is adjusted to a reversal field  $H_R$ . The magnetization is then measured as the magnetic field,  $H$  is gradually increased from  $H_R$  until the sample returns to positive saturation, forming a FORC. A series of FORCs is obtained by changing the  $H_R$  values in steps, as illustrated in Fig. 7-6(f). The FORC distribution is defined by a mixed second-order derivative:  $\sigma(H_R, H) = \frac{\partial^2 M}{\partial H_R \partial H}$ , which effectively removes the purely reversible components of magnetization. Attaining reversible features from FORC analysis requires special attention as the reversible feature first appear around  $H=H_R$ . Extended FORC method is used to obtain the reversible feature, where the data is extended by assuming  $M(H < H_R) \equiv M(H_R)$ . Thus, any nonzero value of  $\sigma$  in the  $H-H_R$  plane in the range  $H < H_R$  indicates irreversible switching processes, such as domain nucleation and annihilation and the reversible process at  $H \approx H_R$ .

Fig. 7-6(c) shows the FORC distribution  $\sigma$  corresponding to the  $M-H$  loops in Fig. 7-6(f) for the TmIG/CoFeB( $x=1$ nm)/W/CoFeB/MgO/W sample subjected to out-of-plane external fields. Three line-scans corresponding to the reference circles in Fig. 7-6(f) (indicating different reversal fields,  $H_R$ ) are marked in Fig. 7-6(c) with dashed lines. For  $\mu_0 H_R < 0$  mT, the sample remains mostly saturated, thus, appreciable FORC features are not observed. As  $\mu_0 H_R$  is decreased,  $-3.2$  mT  $< \mu_0 H_R < 0$  mT, the magnetization starts to switch by the nucleation and propagation of reversed domains. A horizontal FORC feature appears in Fig. 7-6(c) as denoted by is by region 1, indicates the propagation of reverse domains followed by steady domain evolution and then rapid growth of positively oriented domains. With further decrease in reversal field,  $-10.4$  mT  $< \mu_0 H_R < -3.2$  mT a vertical FORC feature is observed denoted by region 2. In this interval, the magnetization evolution remains steady for external field,  $\mu_0 H_R < \mu_0 H < 0.8$  mT, followed by rapid growth of positively oriented domains around  $0.8$  mT  $< \mu_0 H < 2.7$  mT. Thus, the vertical feature in FORC represents annihilation of reversed domain after large portions of the samples become negatively saturated. For  $\mu_0 H_R < -11$  mT, the sample remains negatively saturated and the successive FORC scans does not show any appreciable change in the magnetization reversal process.

The FORCs distribution for *out of plane fields* of FI-only (TmIG-only) sample, a representative control FM stack with  $x=1$  nm and FI/FM with  $x= 1$  nm, 3 nm and 4 nm are presented in Fig. 7-6(a)-(e). The FORC diagrams of both the FI-only and FI/FM heterostructure samples reveal two prominent features: a horizontal

ridge parallel to the  $H$  axis and a vertical ridge parallel to the  $H_R$  axis, as previously described and typically observed in films with PMA [30,35-37]. There is no clear splitting in the FORC ridges of the heterostructure samples in Fig. 7-6(c)-(e), that suggests there are no strongly separated modes of reversals for the individual FI and FM layers [38]. Thus, FI and FM layers are coupled and switched together although part of the thicker CoFeB layers switch at higher fields. However, the strength of exchange coupling and magnetostatic interaction also varies with the thickness of the CoFeB layer. We explain this as follows.

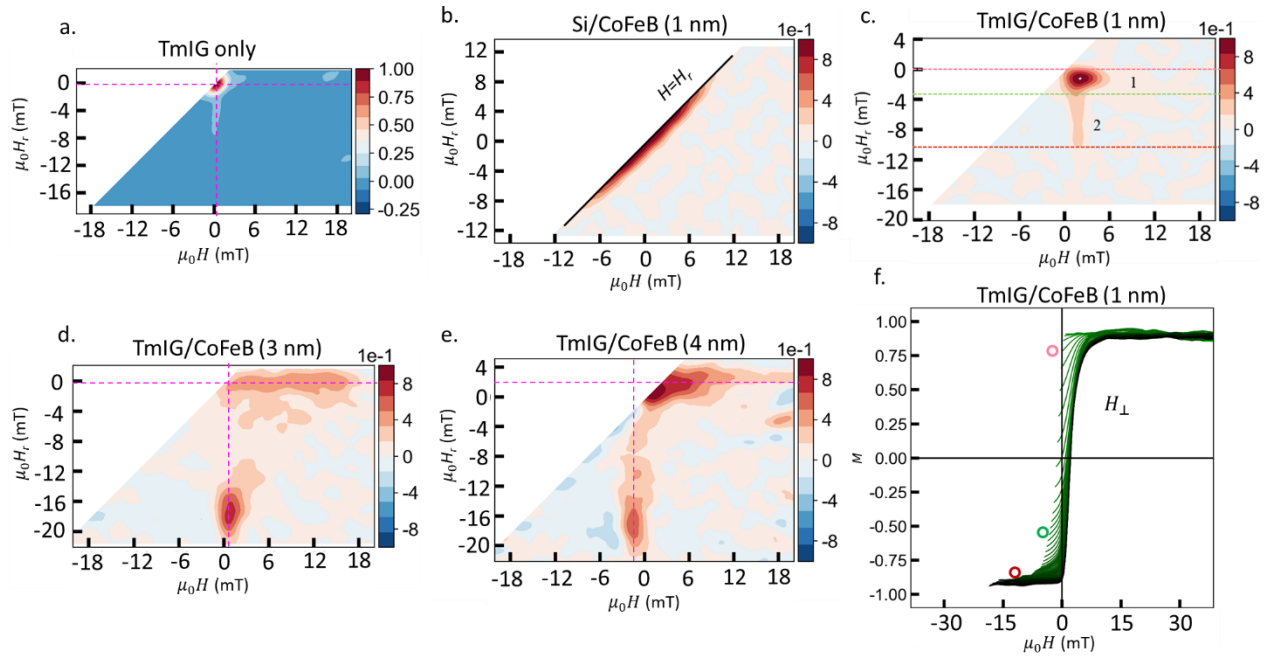


Figure 7-6 FORC distributions for samples with a. FI-only b. control FM stack with  $x=1$  nm CoFeB and FI/FM heterostructures with variable CoFeB layer thickness of c.  $x=1$  nm d.  $x=3$  nm e.  $x=4$  nm for out of plane external fields,  $H_{\perp}$ . f. Family of FORC curves determined for different reversal fields,  $H_R$  for FI/FM sample with  $x=1$  nm CoFeB.

For the pristine TmIG, horizontal ridge centers around  $\mu_0 H_R = -0.2$  mT and the vertical ridge around,  $\mu_0 H = 0.2$  mT as illustrated in Fig. 7-6(a) (see the horizontal and vertical dashed magenta lines). In contrast, the control FM stack with  $x=1$  nm, does not exhibit any horizontal or vertical FORC features. Rather the features are concentrated along the  $H=H_R$  line, indicating the magnetization switching is process predominantly reversible, Fig. 7-6(b).

For FI/FM with  $x=1$  nm the horizontal FORC ridge shifts to  $\mu_0 H_R = -1.2$  mT and the vertical ridge shifts to  $\mu_0 H = 1.8$  mT as can be seen in Fig. 7-6(c). The shift in FORC ridges compared to pristine TmIG sample indicates that large fields are required to either nucleate the reversed domains or coercively annihilate the domains of the heterostructures. This suggests pinning of the TmIG magnetization from the thin CoFeB

layer due to interlayer exchange coupling. Furthermore, no reversible FORC features around  $H=H_R$  line is observed indicating the CoFeB and TmIG magnetizations are coupled and switched together. Similar FORC features are also observed for FI/FM samples with  $x=0.84$  nm (not shown) suggesting strong interlayer exchange coupling in samples with  $x \leq 1$  nm.

In contrast, when the CoFeB thickness is increased to  $x=4$  nm, the vertical ridge shifts to negative  $\mu_0 H = -1.8$  mT and the horizontal FORC ridge shifts to positive values of  $\mu_0 H_R = 2$  mT, Fig. 7-6(e). This indicates that the nucleation and annihilation field occur sooner, as the in-plane anisotropy of the thick CoFeB layer promotes out-of-plane switching of FI/FM heterostructure at lower fields. This is consistent with increased dipolar interactions, which will promote early (lower field) domain nucleation [39] during reversal. In the PMA TmIG layer, lateral domains will have anti-parallel alignment due to uniaxial anisotropy, thus leading to strong dipolar fields. A portion of the thicker CoFeB layer that is beyond the interlayer exchange coupling length ( $\sim 1$ nm for CoFeB obtained from micromagnetic simulation in section 7.4) orients along the in-plane direction. Thus, the magnetic flux lines emanating from the oppositely oriented lateral domains are followed by the in-plane oriented CoFeB magnetization to form closed flux paths (see Fig. 7-7(c)) and minimizes the magnetostatic energy. In addition, the horizontal FORC ridge elongates significantly up to  $\mu_0 H = 20$  mT as seen in Fig. 7-6(e) compared to TmIG or heterostructures with thin CoFeB where the horizontal FORC ridge extends up to 6 mT. This suggests in heterostructures with thick CoFeB once reversed domains are formed in TmIG, they become magneto-statically coupled with in-plane CoFeB magnetization. As a result, large fields are required to bring the reversed domains to positive saturation.

The heterostructure with  $x=3$  nm behaves similarly to the heterostructure with  $x=4$  nm CoFeB excepts the horizontal FORC features shifted along negative value of  $\mu_0 H_R = -0.2$  mT and the vertical ridge along the positive value  $\mu_0 H = 0.6$  mT, Fig. 7-6(d). Although magnetostatic interaction is the dominating interaction, in the  $x = 3$  nm heterostructure sample, more of the CoFeB layer is exchange-coupled to TmIG layer compared to  $x=4$  nm sample. Additional features are observed in between the dominant horizontal and vertical FORC ridges for  $x \geq 3$  nm CoFeB heterostructures in Fig. 7-6(d)-(e). To confirm if those features are coming from any separate modes of reversal, we also performed FORC distribution for in-plane fields for  $x=3$  nm heterostructures as shown in Fig. 7-7(a)-(b). One single peak of the FORC distribution in Fig. 7-7(a) is observed suggesting the FM and FI layers are switching together for in-plane fields as well.

Thus, the FORC study on the heterostructure samples confirms that the individual FI and FM layers are coupled and switched together. The interaction is dominated by exchange coupling for sample with  $x \leq 1$  nm CoFeB, whereas magnetostatic coupling dominates in samples with  $x \geq 3$  nm CoFeB.

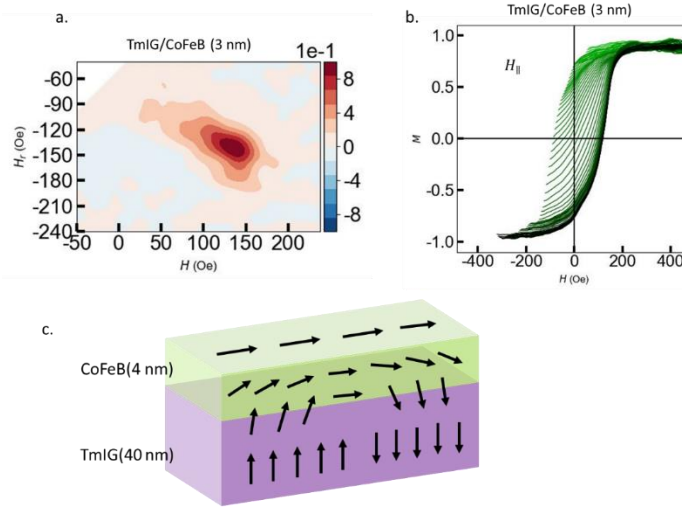


Figure 7-7 a. FORC distributions and b. family of FORC curves determined for different reversal fields for sample TmIG/CoFeB(3nm) with in-plane field direction. c. Schematic showing flux closure in heterostructure with CoFeB thickness,  $x=3, 4$  nm.

### 7.3 Micromagnetic simulation

Micromagnetic simulations are carried out using Mumax3 [40] to investigate the strength of interlayer exchange interaction for varying CoFeB overlayer thickness in the FI/FM heterostructures. We choose micromagnetic simulation over the macroscopic Stoner-Wolfarth model due to its ability to account for non-uniform magnetization, domain structures, and complex magnetic interactions in magnetic multilayers. Excellent qualitative match is found to the hysteresis loops from VSM, and the domain reversal obtained from MOKE experiments. A representative thin and thick CoFeB overlayer sample with  $x=4$  nm and 1 nm are studied considering the computational complexity of simulating all the samples. The simulation geometry is assumed to be  $2.048 \mu\text{m} \times 2.048 \mu\text{m} \times (40+x)$  nm with finite different discrete cells of  $4 \times 4 \times 1 \text{ nm}^3$  and  $8 \times 8 \times 1 \text{ nm}^3$  cells for  $x=4$  nm and 1 nm respectively. The small lateral dimensions were needed to ensure the micromagnetic simulations are computationally feasible. Furthermore, periodic boundary conditions are used to adequately capture the magnetostatic energy of the larger film geometry. The exchange stiffness of TmIG is considered to be  $A_{FI} = 2$  pJ/m and the uniaxial perpendicular anisotropy is  $K_{u,FI} = 7$  kJ/m<sup>3</sup> consistent with the values found in the literatures [41]. A 5% variation in  $K_{u,FI}$  across grains is considered with an average grain size of  $\sim 400$  nm. The saturation magnetization  $M_{S,FI}$  is 80 kA/m measured from the VSM. For the FM stack, CoFeB( $x$ )/W (0.4 nm)/CoFeB(0.8 nm)/MgO(1 nm)/W(5 nm) both of the CoFeB layers are considered with saturation magnetization of  $M_{S,FM}$  of 970 kA/m and 550 kA/m for  $x=4$  nm and 1 nm sample respectively. For  $x=1$  nm sample, the perpendicular anisotropy of MgO/CoFeB is considered to be  $K_u = 250$  kJ/m<sup>3</sup>. The exchange stiffness of the CoFeB is considered to be  $A_{FI} = 14$  pJ/m [42-43]. The exchange stiffness and saturation magnetization are kept the same for the other

CoFeB layer. The exchange interaction between the CoFeB layers separated by a very thin 0.4 nm W spacer is strongly ferromagnetic as seen from the VSM loops, thus assumed to be only reduced by 20%. The lateral cell size we considered is within the exchange length of TmIG,  $\sqrt{A_{FI}/0.5\mu_0M_{s,FI}^2} \sim 40$  nm and CoFeB,  $\sqrt{A_{FM}/0.5\mu_0M_{s,FM}^2} \sim 4.9$  nm and 8.6 nm for  $x=4$  nm and  $x=1$  nm CoFeB respectively. The interlayer exchange coupling (IEC) between TmIG/CoFeB is varied and a qualitative match is found with  $IEC=1.34$  mJ/m<sup>2</sup> and 0.25 mJ/m<sup>2</sup> for  $x=4$  nm and  $x=1$  nm sample respectively. The results are presented in Fig. 7-8 which shows the out of plane simulated hysteresis curves for total FM/FI heterostructures and the individual FM and FI layers. For  $x=4$  nm sample, the sample saturates at  $\mu_0H_{\perp} > 1000$  mT (not shown here) and the domains starts to nucleate around 5 mT (see green circle in the top panel in Fig. 7-8), qualitatively similar to what we obtained from VSM measurements where the saturation field is beyond 1000 mT and the domain nucleation starts around 4 mT (when the field is decreased from positive saturation). The pinched magnetometry loop observed in VSM is well predicted in micromagnetic simulation. Increasing the IEC from 1.34 mJ/m<sup>2</sup> results in much slower magnetization evolution towards saturation, while decreasing the IEC results in faster evolution, however the nucleation field becomes higher than 5 mT.

For  $x=1$  nm sample, with  $IEC=0.25$  mJ/m<sup>2</sup> the out of plane coercivity is predicted to be  $\approx 2.3$  mT which is comparable to the coercivity measured from VSM, 1.82 mT. Decreasing the IEC from 0.25 mJ/m<sup>2</sup> increases the coercive field from 2.3 mT. Although increasing the IEC decreases the coercive field, in these scenarios, when the magnetization reverses, it progresses much slowly towards saturation compared to what we observed in VSM loop in Fig. 7-3(a).

Micromagnetic snapshots with magnetization orientation in the finite difference cells for the  $2.048 \mu\text{m} \times 2.048 \mu\text{m}$  regions are presented for the total heterostructure, the top CoFeB layer and the bottom TmIG layers are also presented in Fig. 7-8. The black and white contrast in the snapshots represents the magnetization directions along the negative and positive out of plane directions respectively and the different colors show the magnetization orientations along in-plane directions. For  $x=4$  nm samples, the top layer CoFeB magnetization has smaller projections in the out of plane direction than the bottom TmIG layer magnetizations as seen from the hysteresis loops in the top panels of Fig. 7-8. This is due to the fact that, with the determined value of  $IEC=1.34$  mJ/m<sup>2</sup>, a small portion of the CoFeB that is within the exchange length is coupled with perpendicular TmIG bottom layer and large portions are oriented along in-plane to support flux closure as can be seen from the cross-sectional view of the micromagnetic spin configuration near the domain walls in Fig. 7-9. Thus, magnetostatic coupling dominates over exchange coupling for the heterostructure with thick CoFeB. Also, changes in magnetization values along the depths of the TmIG are observed due to the fact that interlayer exchange coupling is short range interaction, and the coupling

strength diminishes significantly outside the exchange length. The TmIG portion that is in direct contact with CoFeB aligns closely with the in-plane CoFeB magnetization while the TmIG magnetization that is further away from the CoFeB and in direct contact with substrate predominantly aligns to perpendicular to the sample plane (see Fig. 7-9(a) and different color contrasts between the two micromagnetic snapshots of bottom TmIG layer at different depth specifically that are in contact with top CoFeB layer and in contact with substrate GGG in Fig. 7-8). Moreover, the nucleation and annihilation of the domains and corresponding patterns during the switching process found from micromagnetic simulations are qualitatively the same as seen in MOKE microscopy. Weak imprinting of TmIG domains on thick CoFeB overlayer is predicted in micromagnetic simulation, which is similar to what we observed from MOKE images where the domains are seen to possess less sharp intensity contrast (bottom panels in Fig. 7-4). Furthermore, dendritic domains with maximum size of  $\sim 0.39 \mu\text{m}$  are observed at 0 mT applied field following positive saturation, which is close to what we found from domain reversal images from MOKE experiments ( $0.46 \mu\text{m}$ , half of the domain wall periodicity).

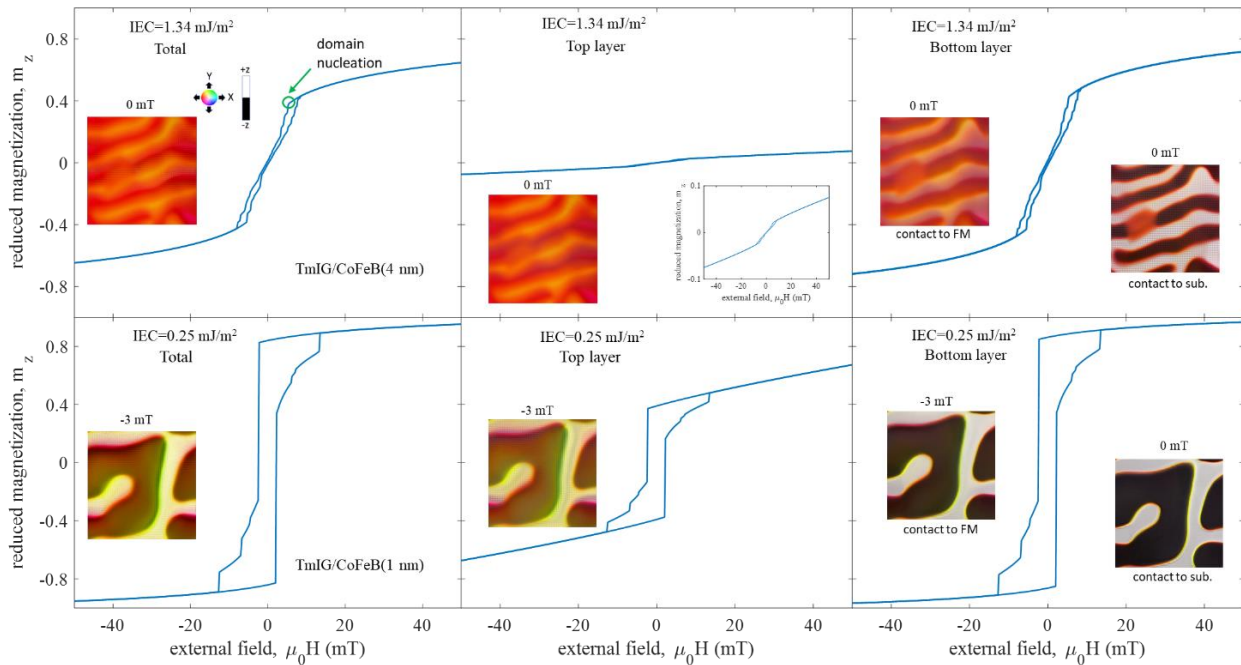


Figure 7-8 Hysteresis loops along the out of plane directions derived from micromagnetic simulations for the total heterostructure, top layer CoFeB and bottom layer TmIG are shown for samples with  $x=4 \text{ nm}$  and  $x=1 \text{ nm}$ . The micromagnetic snapshots for both of the samples for total magnetizations, top CoFeB layer and bottom TmIG layers magnetizations are shown. For the bottom TmIG layer, two snapshots taken at different depths of TmIG (contact to FM denotes the TmIG layer that is directly in contact with CoFeB and contact to sub. shows the TmIG layer that is in direct contact with the GGG substrate).

In contrast to the thick CoFeB, heterostructure with thin CoFeB sample ( $x=1$  nm) shows distinctive difference where the domains in TmIG are clearly seen to be strongly imprinted on the CoFeB layer as seen from Fig. 7-8. The CoFeB top layer magnetization exhibits large out of plane projections. With the determined interlayer exchange coupling,  $IEC= 0.25$  mJ/m<sup>2</sup>, it mostly couples the CoFeB layer. Also, the out of plane projections of TmIG magnetization and total heterostructure magnetization are larger than the thick CoFeB sample and qualitatively similar to what we have seen in VSM measurements as shown in Fig. 7-3(a). Depth dependent magnetization in the TmIG bottom layer is also simulated in the sample as the exchange length cannot encompass the whole extent of the 40 nm thick TmIG layer (see Fig. 7-9b). Qualitative matches between the domain wall patterns during the switching process are observed. Domain patterns at -3mT applied field following positive field saturation exhibits dendritic domains with maximum width of  $\sim 1.76$   $\mu$ m which is qualitatively same as the width of the domain wall observed in MOKE images.

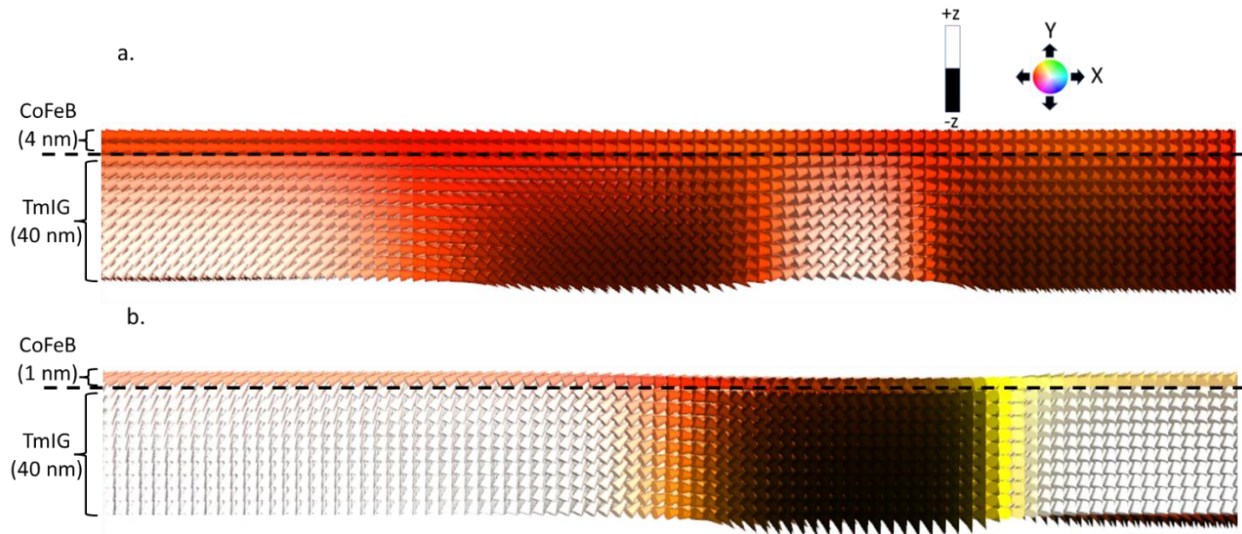


Figure 7-9 Micromagnetic spin orientation in FI/FM heterostructures with a.  $x=4$  nm and b.  $x= 1$  nm along the cross-section of domain walls for out of plane external field,  $\mu_0 H_{\perp}=0$  mT and  $\mu_0 H_{\perp} =-3$  mT respectively. The number of spins is down sampled from actual micromagnetic simulations for better visibility.

## 7.4 Summary and conclusion

Magnetometry loops from VSM, mean field interaction study using FORCs and micromagnetic simulation indicate coupling in the studied FI/FM heterostructure. The coupling is mediated by interlayer exchange and magnetostatic interactions with their relative strength determined by the thickness of the CoFeB overlayer. The resulting trend for different thickness CoFeB can be explained as follows. The exchange interaction prefers parallel alignments of the respective magnetization in the FI and FM layers. The exchange energy per unit area can be expressed as,  $IEC \cdot (\vec{m}_1 \cdot \vec{m}_2)$ , where the IEC is the interlayer exchange energy coefficient and  $\vec{m}_1$  and  $\vec{m}_2$  are the reduced average magnetization of the FI and FM layer

respectively. For TmIG/CoFeB (1 nm) sample, the  $IEC$  is determined to be in the range of  $0.25 \text{ mJ/m}^2$  from micromagnetic simulation, which shows a qualitative match in the hysteresis loop to that measured in VSM by considering practical material parameters. With  $IEC=0.25 \text{ mJ/m}^2$ , the exchange length for the CoFeB layer is  $l_{ex,CoFeB} \sim 0.88 \text{ nm}$  ( $l_{ex} \sim \sqrt{\frac{IEC \cdot c_z}{\frac{1}{2}\mu_0 M_{s,FM}^2}}$ , where,  $c_z$  is the cell size along film normal and  $M_{s,FM}$  is the saturation magnetization of the FM layer). This suggests that in the TmIG/CoFeB (1 nm) sample, most of the CoFeB layer is within the exchange length and coupled strongly with perpendicular TmIG. Extending the analysis for the thicker CoFeB sample with  $x=4 \text{ nm}$ , with  $IEC=1.34 \text{ mJ/m}^2$ , the exchange length for the CoFeB layer is determined to be  $l_{ex,CoFeB} \sim 1.1 \text{ nm}$ . This suggests that a large portion of the CoFeB is beyond the exchange length hence is not perpendicularly coupled to the TmIG bottom layer. Thus, the magnetizations in the CoFeB layer mostly align along in-plane and the resulting in-plane hysteresis shows square loop as seen in Fig. 7-3(d). Furthermore, the interaction between the FM and FI layers is now dominated by magnetostatic interaction where the in-plane CoFeB supports flux closure paths during the domain formation in the TmIG layer (Fig. 7-9(a)). Although the exchange coefficient determined from micromagnetic simulations is higher for heterostructures with thick CoFeB compared to the thin one, the interlayer exchange lengths in CoFeB are found to be similar for both samples. This is due to the fact that the saturation magnetization is higher in heterostructure samples with a thick CoFeB layer.

The interlayer exchange coupling between a ferrimagnetic insulator with PMA, GGG/TmIG (40) and a ferromagnetic compound CoFeB( $x$ )/W (0.4)/CoFeB (0.8)/MgO (1)/W (5) is studied by varying the thickness of the variable CoFeB layer. The relative strength of the competing interactions such as the interlayer exchange and magnetostatic coupling is found to be strongly dependent on the thickness of the CoFeB layer. When deposited on silicon, the ferromagnetic compound with CoFeB thickness  $x \leq 1 \text{ nm}$  exhibits PMA, however, the PMA is weak and the magnetization switching occurred predominantly in reversible manner as seen from the FORC features. When the same ferromagnetic compound with  $x \leq 1 \text{ nm}$  CoFeB is deposited on GGG/TmIG, the PMA of the heterostructure increased compared to pristine TmIG. Both of the CoFeB and TmIG magnetizations are observed to switch together with domain nucleation and propagation events as confirmed from the dominant FORC features. Thus, strong exchange coupling contributes to the modified PMA of the heterostructure samples, compared to pristine TmIG and control CoFeB sample with thickness  $x \leq 1 \text{ nm}$ . When ferromagnetic compound with  $x \geq 3 \text{ nm}$  thick CoFeB are deposited on PMA TmIG, magnetostatic coupling is observed to dominates over exchange coupling as confirmed from MOKE domain reversal images and FORC analysis which suggests increased saturation magnetization weakens the exchange interaction and facilitate flux closure paths through the thick CoFeB for minimizing magnetostatic energy. The findings of our study provide important insights towards realizing fast and efficient spintronic memory device controlled by voltage-induced strain [44-50], voltage-



controlled magnetic anisotropy (VCMA) [51-53], current [54-55], and combination of voltage and current [56-58] and consists of a ferrimagnetic insulator.

## References:

- [1] C. O. Avci, A. Quindeau, C.-F. Pai, M. Mann, L. Caretta, A. S. Tang, M. C. Onbasli, C. A. Ross and G. S. D. Beach, Current-induced switching in a magnetic insulator, *Nature Materials*, vol. 16, pp. 309–314 (2017).
- [2] C. O. Avci, E. Rosenberg, M. Baumgartner, L. Beran, A. Quindeau, P. Gambardella, C. A. Ross, and G. S. D. Beach, Fast switching and signature of efficient domain wall motion driven by spin-orbit torques in a perpendicular anisotropy magnetic insulator/Pt bilayer, *Appl. Phys. Lett.* 111, 072406 (2017)
- [3] C. O. Avci, E. Rosenberg, L. Caretta, F. Büttner, M. Mann, C. Marcus, D. Bono, C. A. Ross and G. S. D. Beach, Interface-driven chiral magnetism and current-driven domain walls in insulating magnetic garnets, *Nature Nanotechnology*, vol. 14, pp. 561–566 (2019)
- [4] S. Vélez, J. Schaab, M. S. Wörnle, M. Müller, E. Gradauskaite, P. Welter, C. Gutgsell, C. Nistor, C. L. Degen, M. Trassin, M. Fiebig and P. Gambardella, High-speed domain wall racetracks in a magnetic insulator, *Nature Communications*, vol. 10, Art. no.4750 (2019)
- [5] S. Ding, A. Ross, R. Lebrun, S. Becker, K. Lee, I. Boventer, S. Das, Y. Kurokawa, S. Gupta, J. Yang, G. Jakob, and M. Kläui, Interfacial Dzyaloshinskii-Moriya interaction and chiral magnetic textures in a ferrimagnetic insulator, *Phys. Rev. B* 100, 100406(R) (2019)
- [6] L. Caretta, E. Rosenberg, F. Büttner, T. Fakhru, P. Gargiani, M. Valvidares, Z. Chen, P. Reddy, D. A. Muller, C. A. Ross and G. S. D. Beach, Interfacial Dzyaloshinskii-Moriya interaction arising from rare-earth orbital magnetism in insulating magnetic oxides, *Nature Communications* volume 11, Art. no.: 1090 (2020)
- [7] L. Caretta, S. H. Oh, T. Fakhru, D. K. Lee, B. H. Lee, S. K. Kim, C. A. Ross, K. J. Lee, G. S. D. Beach, Relativistic kinematics of a magnetic soliton, *SCIENCE*, vol. 370,6523, pp.. 1438-1442 (2020)
- [8] K.-H. Hellwege, A. M. Hellwege, *Landolt-Börnstein , Group III Crystal and Solid State Physics Vol 12a*, Springer-Verlag, Berlin/Heidelberg 1978
- [9] L. Giovannini, S. Tacchi, G. Gubbiotti, G. Carlotti, F. Casoli and F. Albertini, Brillouin light scattering study of exchange-coupled Fe/Co magnetic multilayers, *J. Phys.: Condens. Matter* 17, 6483–6494 (2005)

- [10] A. Bollero, L. D. Buda-Prejbeanu, V. Baltz, J. Sort, B. Rodmacq, and B. Dieny, Magnetic behavior of systems composed of coupled ferromagnetic bilayers with distinct anisotropy directions, *Phys. Rev. B* 73, 144407 (2006)
- [11] T. N. Anh Nguyen, R. Knut, V. Fallahi, S. Chung, Q. Tuan Le, S. M. Mohseni, O. Karis, S. Peredkov, R. K. Dumas, Casey W. Miller, and J. Åkerman, Depth-Dependent Magnetization Profiles of Hybrid Exchange Springs, *Phys. Rev. Applied* 2, 044014 (2014)
- [12] A. Fert, P. Grünberg, A. Barthélémy, F. Petroff, W. Zinn, Layered magnetic structures: interlayer exchange coupling and giant magnetoresistance, *Journal of Magnetism and Magnetic Materials*, Volumes 140–144, Part 1, Pages 1-8 (1995)
- [13] M. D. Stiles, Interlayer exchange coupling, *Journal of Magnetism and Magnetic Materials*, Volume 200, Issues 1–3, Pages 322-337 (1999)
- [14] J. Choi, B.-C. Min, J.-Y. Kim, B.-G. Park, J. H. Park, Y. S. Lee, and K.-H. Shin, Non-collinear magnetization configuration in interlayer exchange coupled magnetic thin films, *Appl. Phys. Lett.* 99, 102503 (2011)
- [15] F. C. Ummelen; A. Fernández-Pacheco, R. Mansell, D. Petit; H. J. M. Swagten; R. P. Cowburn, Controlling the canted state in antiferromagnetically coupled magnetic bilayers close to the spin reorientation transition, *Appl. Phys. Lett.* 110, 102405 (2017)
- [16] A. Parente, H. Navarro, N. M. Vargas, P. Lapa, Ali C. Basaran, E. M. González, C. Redondo, R. Morales, A. Munoz Noval, Ivan K. Schuller, and J. L. Vicent, *ACS Appl. Mater. Interfaces*, 14, 49, 54961–54968 (2022)
- [17] Z. R. Nunn, C. Abert, D. Suess, E. Girt, Control of the noncollinear interlayer exchange coupling, Nunn et al., *Sci. Adv.* 6, eabd8861 (2020)
- [18] P. Hyde, L. Bai, D. M. J. Kumar, B. W. Southern, C.-M. Hu, S. Y. Huang, B. F. Miao, and C. L. Chien, Electrical detection of direct and alternating spin current injected from a ferromagnetic insulator into a ferromagnetic metal, *Phys. Rev. B* 89, 180404(R) (2014)
- [19] H. Qin, S. J. Hämäläinen and S. van Dijken, Exchange-torque-induced excitation of perpendicular standing spin waves in nanometer-thick YIG films, *Scientific Reports* 8, 5755 (2018)
- [20] S. Klingler, V. Amin, S. Geprägs, K. Ganzhorn, H. Maier-Flaig, M. Althammer, H. Huebl, R. Gross, R. D. McMichael, M. D. Stiles, S. T. B. Goennenwein, and M. Weiler, Spin-torque excitation of perpendicular standing spin waves in coupled YIG/Co heterostructures, *Phys. Rev. Lett.* 120, 127201 (2018)
- [21] P.-C. Changa, V. R. Mudinepallia, S.-Y. Liua, H.-L. Lina, C.-C. Hsua, Y.-T. Liaoa, S. Obinatab, T. Kimurab, M.-Y. Chernc, F.-Y. Loaa, W.-C. Lin, Interfacial exchange coupling-modulated

- magnetism in the insulating heterostructure of CoOx/yttrium iron garnet, *Journal of Alloys and Compounds* 875,159948 (2021)
- [22] Y. S. Chun and K. M. Krishnan, Interlayer perpendicular domain coupling between thin Fe films and garnet single-crystal underlayers, *Journal of Applied Physics* 95, 11 (2004)
- [23] N. Vukadinovic, J. Ben Youssef, V. Castel, and M. Labrune, Magnetization dynamics in interlayer exchange-coupled in-plane/out-of-plane anisotropy bilayers, *Phys. Rev. B* 79, 184405(2009)
- [24] M. Pashkevich, A. Stupakiewicz, A. Kirilyuk, A. Maziewski, A. Stognij, N. Novitskii, A. Kimel, and Th. Rasing, Tunable magnetic properties in ultrathin Co=garnet heterostructures, *JOURNAL OF APPLIED PHYSICS* 111, 023913 (2012)
- [25] G.-G. An, J.-B. Lee, S.-M. Yang, J.-H. Kim, W.-S. Chung, J.-P. Hong, Highly stable perpendicular magnetic anisotropies of CoFeB/MgO frames employing W buffer and capping layers, *Acta Materialia*, Volume 87, pp. 259-265 (2015)
- [26] Y. Yan, X. Lu, B. Liu, X. Zhang, X. Zheng, H. Meng, W. Liu, J. Wang, I. G. Will, J. Wu, P. K. J. Wong, J. Cai, J. Du, R. Zhang, Y. Xu, Element-specific spin and orbital moments and perpendicular magnetic anisotropy in Ta/CoFeB/MgO structures, *J. Appl. Phys.* 127, 063903 (2020)
- [27] D. M. Lattery, D. Zhang, J. Zhu, X. Hang, J.-P. Wang and X. Wang, Low Gilbert Damping Constant in Perpendicularly Magnetized W/CoFeB/MgO Films with High Thermal Stability, *Scientific Reports*, 8, 13395 (2018)
- [28] B. Kaplan, G. A. Gehring, The domain structure in ultrathin magnetic films, *J. Magn. Magn. Mater.* 1993, 128, 111
- [29] C. R. Pike, A. P. Roberts, K. L. Verosub, Characterizing interactions in fine magnetic particle systems using first order reversal curves, *J. Appl. Phys.* 85, 6660–6667 (1999)
- [30] J. E. Davies, O. Hellwig, E. E. Fullerton, G. Denbeaux, J. B. Kortright, and K. Liu, Magnetization reversal of Co/Pt multilayers: Microscopic origin of high-field magnetic irreversibility, *Phys. Rev. B* 70, 224434 (2004)
- [31] R. K. Dumas, C.-P. Li, I. V. Roshchin, I. K. Schuller, and K. Liu, Magnetic fingerprints of sub-100nm Fe dots, *Phys. Rev. B* 75, 134405 (2007)
- [32] B. J. Kirby, J. E. Davies, K. Liu, S. M. Watson, G. T. Zimanyi, R. D. Shull, P. A. Kienzle, and J. A. Borchers, Vertically graded anisotropy in Co/Pd multilayers, *Phys. Rev. B* 81, 100405(R) (2010)
- [33] J. E. Davies; O. Hellwig; E. E. Fullerton; M. Winklhofer; R. D. Shull; Kai Liu, Frustration driven stripe domain formation in Co/Pt multilayer films, *Appl. Phys. Lett.* 95, 022505 (2009)
- [34] A. Rotaru, J.-H. Lim, D. Lenormand, A. Diaconu, J. B. Wiley, P. Postolache, A. Stancu, and L. Spinu, Interactions and reversal-field memory in complex magnetic nanowire arrays, *Phys. Rev. B* 84, 134431 (2011)

- [35] C.-I. Dobrotă, A. Stancu, What does a first-order reversal curve diagram really mean? A study case: Array of ferromagnetic nanowires, *J. Appl. Phys.* 113, 043928 (2013)
- [36] T. Schrefl, T. Shoji, M. Winklhofer, H. Oezelt, M. Yano, G. Zimanyi, First order reversal curve studies of permanent magnets, *J. Appl. Phys.* 111, 07A728 (2012)
- [37] D. A. Gilbert, M.-Y. Im, K. Liu, P. Fischer, Element-specific first order reversal curves measured by magnetic transmission x-ray microscopy, *APL Mater.* 10, 111105 (2022)
- [38] J. E. Davies, D. A. Gilbert, S. M. Mohseni, R. K. Dumas, J. Åkerman, Kai Liu, Reversal mode instability and magnetoresistance in perpendicular (Co/Pd)/Cu/(Co/Ni) pseudo-spin-valves. *Appl. Phys. Lett.* 103, 022409 (2013).
- [39] D. A. Gilbert, J.-W. Liao, B. J. Kirby, M. Winklhofer, C.-H. Lai and K. Liu, Magnetic Yoking and Tunable Interactions in FePt-Based Hard/Soft Bilayers, *Scientific Reports*, vol. 6, Art. no. 32842 (2016)
- [40] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. V. Waeyenberge, The design and verification of MuMax3, *AIP Adv.*, vol. 4, no. 10, Oct. 2014, Art. no. 107133
- [41] E. R. Rosenberg, K. Litzius, J. M. Shaw, G. A. Riley, G. S. D. Beach, H. T. Nembach, and C. A. Ross, Magnetic Properties and Growth-Induced Anisotropy in Yttrium Thulium Iron Garnet Thin Films, *Adv. Electron. Mater.*, 7, 2100452 (2021)
- [42] J. Cho, J. Jung, K.-E. Kim, S.-I. Kim, S.-Y. Park, M.-H. Jung, C.-Y. You, Effects of sputtering Ar gas pressure in the exchange stiffness constant of Co<sub>40</sub>Fe<sub>40</sub>B<sub>20</sub> thin films, *Journal of Magnetism and Magnetic Materials* 339, 36–39 (2013)
- [43] G.-M. Choi, Exchange stiffness and damping constants of spin waves in CoFeB films, *Journal of Magnetism and Magnetic Materials* 516 167335, (2020)
- [44] V. Sampath, N. D'Souza, D. Bhattacharya, G. M. Atkinson, S. Bandyopadhyay, and J. Atulasimha, "Acoustic-wave-induced magnetization switching of magnetostrictive nanomagnets from single-domain to nonvolatile vortex states," *Nano Lett.*, vol. 16, no. 9, pp. 5681–5687, Sep. 2016.
- [45] N. D'Souza, M. Salehi Fashami, S. Bandyopadhyay, and J. Atulasimha, "Experimental clocking of nanomagnets with strain for ultralow power Boolean logic," *Nano Lett.*, vol. 16, no. 2, pp. 1069–1075, Feb. 2016.
- [46] A. K. Biswas, S. Bandyopadhyay, and J. Atulasimha, "Complete magnetization reversal in a magnetostrictive nanomagnet with voltage generated stress: A reliable energy-efficient non-volatile magneto-elastic memory," *Appl. Phys. Lett.*, vol. 105, no. 7, Aug. 2014, Art. no. 072408.
- [47] K. Roy, S. Bandyopadhyay, and J. Atulasimha, "Hybrid spintronics and straintronics: A magnetic technology for ultra low energy computing and signal processing," *Appl. Phys. Lett.*, vol. 99, no. 6, Aug. 2011, Art. no. 063108.

- [48] A. K. Biswas, S. Bandyopadhyay, and J. Atulasimha, “Energy-efficient magnetoelastic non-volatile memory,” *Appl. Phys. Lett.*, vol. 104, no. 23, Jun. 2014, Art. no. 232403.
- [49] X. Li et al., “Strain-mediated 180° perpendicular magnetization switching of a single domain multiferroic structure,” *J. Appl. Phys.*, vol. 118, no. 1, Jul. 2015, Art. no. 014101.
- [50] N. Lei et al., “Strain-controlled magnetic domain wall propagation in hybrid piezoelectric/ferromagnetic structures,” *Nature Commun.*, vol. 4, no. 1, p. 1378, Jan. 2013.
- [51] C. Grezes et al., “Ultra-low switching energy and scaling in electric-field-controlled nanoscale magnetic tunnel junctions with high resistance-area product,” *Appl. Phys. Lett.*, vol. 108, no. 1, Jan. 2016, Art. no. 012403.
- [52] D. Bhattacharya, M. M. Al-Rashid, and J. Atulasimha, “Voltage controlled core reversal of fixed magnetic skyrmions without a magnetic field,” *Sci. Rep.*, vol. 6, no. 1, Aug. 2016, Art. no. 31272.
- [53] D. Bhattacharya and J. Atulasimha, “Skyrmion-mediated voltage controlled switching of ferromagnets for reliable and energy-efficient two-terminal memory,” *ACS Appl. Mater. Interface*, vol. 10, no. 20, pp. 17455–17462, May 2018.
- [54] J. C. Slonczewski, “Current-driven excitation of magnetic multilayers,” *J. Magn. Magn. Mater.*, vol. 159, nos. 1–2, pp. 1–7, Jun. 1996.
- [55] K.-S. Ryu, S.-H. Yang, L. Thomas, and S. S. P. Parkin, “Chiral spin torque arising from proximity-induced magnetization,” *Nature Commun.*, vol. 5, no. 1, p. 3910, May 2014.
- [56] W. Al Misba, M. M. Rajib, D. Bhattacharya, and J. Atulasimha, “Acoustic-wave-induced ferromagnetic-resonance-assisted spin-torque switching of perpendicular magnetic tunnel junctions with anisotropy variation,” *Phys. Rev. Appl.*, vol. 14, no. 1, Jul. 2020, Art. no. 014088.
- [57] M. A. Azam, D. Bhattacharya, D. Querlioz, C. A. Ross, and J. Atulasimha, “Voltage control of domain walls in magnetic nanowires for energy-efficient neuromorphic devices,” *Nanotechnology*, vol. 31, no. 14, Apr. 2020, Art. no. 145201.
- [58] W. A. Misba, T. Kaiser, D. Bhattacharya, and J. Atulasimha, “Voltage controlled energy-efficient domain wall synapses with stochastic distribution of quantized weights in the presence of thermal noise and edge roughness,” *IEEE Trans. Electron Devices*, vol. 69, no. 4, pp. 1658–1666, Apr. 2022.

## Chapter 8: Dynamic Coupling in a CoFeB/Perpendicular Ferrimagnetic Thulium Iron Garnet Heterostructure

In addition to the static coupling observed in the FI/FM heterostructure GGG/TmIG(40 nm)/CoFeB(x)/W(0.4 nm)/CoFeB(0.8 nm)/MgO(1 nm)/W(5 nm) discussed in Chapter 7, we explored dynamic coupling arising from the exchange of non-equilibrium spin currents when the magnetic multilayers in these heterostructures are excited with broadband ferromagnetic resonance. When either of the ferrimagnetic (FI) layer (GGG/TmIG(40 nm)) or ferromagnetic (FM) layer (CoFeB(x)/W(0.4 nm)/CoFeB(0.8 nm)/MgO(1 nm)/W(5 nm)) is driven into resonance, the precessing magnetization generates spin currents that are absorbed by the non-resonating layer, dissipating spin momentum. This process induces a relaxation torque on the resonating layer, thereby altering its intrinsic damping properties. Consequently, dynamic coupling serves as a non-local medium for modulating magnetic properties such as damping.

### 8.1 Sample preparation:

The studied samples are the same as studied in chapter 7.

### 8.2 Results and discussion:

The magnetization dynamics of a uniform magnetizations,  $\mathbf{m}_i$  can be expressed using the LLG equation as follows:

$$\frac{d\mathbf{m}_i}{dt} = -\gamma\mathbf{m}_i \times \mathbf{H}_{eff} + \alpha_0\gamma\mathbf{m}_i \times \frac{d\mathbf{m}_i}{dt}$$

Where, the first term represents the torque induced by the effective field,  $\mathbf{H}_{eff} = \frac{\partial f}{\partial \mathbf{M}}$ , where the free energy  $f$  comprises of Zeeman energy, magnetostatic energy, magnetic anisotropy and exchange interaction. The second term is the Gilbert damping term which governs the relaxation to the equilibrium.

In heterostructures samples consisting magnetic multilayers, if magnetizations in any one of the layers undergoes resonant precession, it pumps spin current to the other magnetic layer which in in direct contact with it. The pumped spin current can be expressed as follows [1-3]:

$$\mathbf{I}_i^{pump} = \frac{\hbar}{4\pi} (g_{Re}^{\uparrow\downarrow}\mathbf{m}_i \times \frac{d\mathbf{m}_i}{dt} + g_{Im}^{\uparrow\downarrow}\mathbf{m}_i)$$

Where  $g_{Re}^{\uparrow\downarrow}$  and  $g_{Im}^{\uparrow\downarrow}$  are the real and imaginary part of the spin mixing conductance. Spin mixing conductance dictates the transport of spins that are non-colinear to the magnetization direction. This is also a measure of spin transfer across any interface such as FI/FM or FM/NM (NM is a non-magnetic metal). For most materials the imaginary part of the spin mixing conductance is small [1] and hence neglected in the ensuing discussion. There could be spin backflow due to the spin accumulation at the interface or due to the spin pumping from the other magnetic layers into the layer that is in consideration. The spin backflow reduces the efficiency of spin transfer across an interface. Thus, the net spin currents in the FI/FM can be expressed as,

$$\mathbf{I}_{si} = \mathbf{I}_i^{pump} + \mathbf{I}_i^{back}$$

When both of the magnetic layers' magnetizations undergo precession, both layers pump spins towards each other. In this scenario, the net spin current in any one of the magnetic layers can be expressed as [4]:

$$\mathbf{I}_{si} = \frac{\hbar}{4\pi} \frac{g_i^{\uparrow\downarrow} g_j^{\uparrow\downarrow}}{g_i^{\uparrow\downarrow} + g_j^{\uparrow\downarrow}} \left( \mathbf{m}_i \times \frac{d\mathbf{m}_i}{dt} - \mathbf{m}_j \times \frac{d\mathbf{m}_j}{dt} \right)$$

In FI/FM heterostructure,  $\mathbf{m}_i$  and  $\mathbf{m}_j$  are the FI and FM magnetizations and  $g_i^{\uparrow\downarrow}$  and  $g_j^{\uparrow\downarrow}$  are the real valued spin mixing conductance for FI/FM and FM/FI interfaces respectively.

The LLG equation of magnetization,  $\mathbf{m}_i$  considering the magnetization precession in both layers can be expressed as:

$$\frac{d\mathbf{m}_i}{dt} = -\gamma \mathbf{m}_i \times \mathbf{H}_{eff} + \alpha_0 \gamma \mathbf{m}_i \times \frac{d\mathbf{m}_i}{dt} + \frac{\gamma}{M_{si} V} \frac{\hbar}{4\pi} \frac{g_i^{\uparrow\downarrow} g_j^{\uparrow\downarrow}}{g_i^{\uparrow\downarrow} + g_j^{\uparrow\downarrow}} \left( \mathbf{m}_i \times \frac{d\mathbf{m}_i}{dt} - \mathbf{m}_j \times \frac{d\mathbf{m}_j}{dt} \right)$$

Where,  $M_{si}$  and  $V$  are the saturation magnetization and volume of magnetic layer,  $i$ ,  $\alpha_0$  is the intrinsic damping. Due to the spin pumping, the intrinsic damping,  $\alpha_0$  will be modified. Specifically, the effective damping of layer,  $i$  will be increased to  $\alpha_i = \alpha_0 + \alpha'_i$ , where the damping enhancement can be expressed as the following:

$$\alpha'_i = \frac{\gamma}{M_{si} V} \frac{\hbar}{4\pi} \frac{g_i^{\uparrow\downarrow} g_j^{\uparrow\downarrow}}{g_i^{\uparrow\downarrow} + g_j^{\uparrow\downarrow}}$$

We see that the damping enhancement of a layer is inversely proportional to its saturation magnetization and volume.

To investigate the extent of dynamic coupling of the FI/FM heterostructure shown in Fig. 8-1a, we performed a broadband ferromagnetic resonance (FMR) study. The magnetization is forced to resonate by

applying microwave power. The FMR is measured by measuring the absorption of microwave power as a function of the applied field,  $H$  by applying a small microwave field of frequency,  $f$  perpendicular to  $H$ . To account for the different anisotropies of the FI and FM layer, FMR measurements are also carried out by changing the applied field angle,  $\theta$  with respect to the sample plane as indicated in the Fig. 8-1b. The absorption due to the imaginary part of the susceptibility  $\chi''$  of the rf magnetization component is fitted with Lorentzian to calculate the inflection point which is resonant frequency and the full width half maximum which is linewidth. Fig. 8-1c shows experimentally measured absorption and corresponding fitting for different field angles for FI/FM samples with  $x= 1$  nm CoFeB. In the FMR measurements, we observed two peaks due to the TmIG and CoFeB layers magnetizations which are identified and fitted with two Lorentzian, after background subtraction from the raw FMR data. Initially at small field angle,  $\theta$ , the CoFeB mode are observed to the left of the TmIG mode. As we increase the angle, the two modes almost merges at  $\theta=60^\circ$ . With further increase in the field angle the CoFeB modes shifts to the right of TmIG mode.

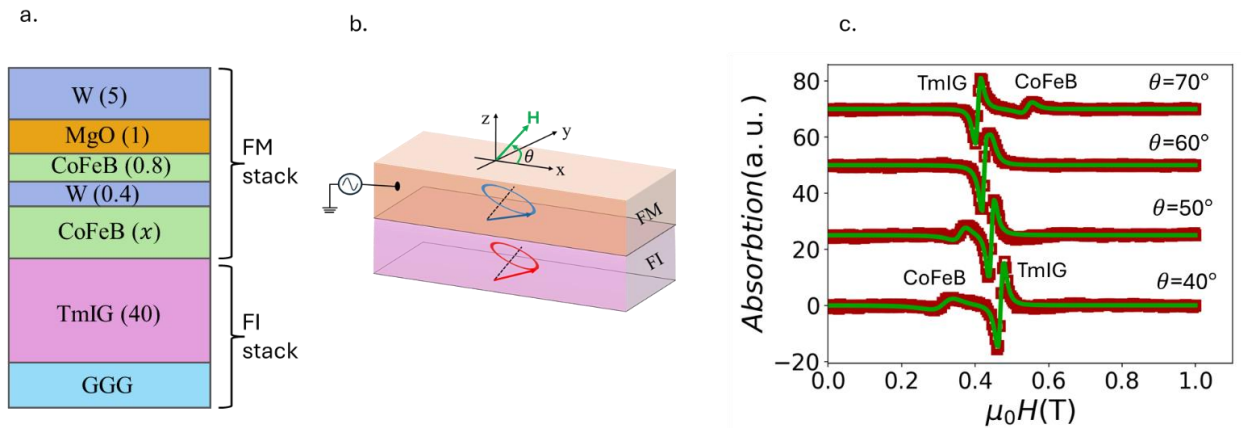


Figure 8-1 a. FI/FM heterostructure, GGG/TmIG(40)/CoFeB(x)/W(0.4)/CoFeB(0.8)/MgO(1)/W(5). All the numbers in parenthesis are in nanometers. b. The schematic of the FI/FM stack under the application of microwave power, the applied field,  $H$  and the angle of the field,  $\theta$  are indicated in the figure. c. The FMR absorption data and the corresponding fitting with two Lorentzian for TmIG and CoFeB modes for FI/FM heterostructure with  $x= 1$  nm CoFeB. The applied field angle for the FMR measurements is shown.

To investigate damping enhancement, we performed FMR measurements on pristine FI sample, GGG/TmIG (40 nm) and FM/FI heterostructure samples with variable CoFeB layer thickness of  $x= 0.84$  nm, 1 nm and 3 nm. The external field,  $H$  is applied perpendicular to the sample plane by keeping the field angle at constant  $\theta=90^\circ$ . The resonance frequency,  $H_{res}$  is fitted by the Kittel equation for perpendicular geometry to extract the effective magnetization,  $M_{eff}$  and Lande  $g$ -factor [5]:



$$H_{res} = \frac{2\pi f}{|\gamma|\mu_0} + M_{eff}$$

Where,  $f$  is the excitation frequency and  $\gamma = \frac{g\mu_B}{\hbar}$ ,  $\mu_B$  is the Bohr magnetron and  $\hbar$  is the reduced Planck constant.  $M_{eff}$  can be further expressed as [6],

$$M_{eff} = M_s - \frac{2(K_u - \frac{1}{2}\mu_0 M_s^2)}{\mu_0 M_s}$$

$M_s$  is the saturation magnetization and  $K_u$  is the perpendicular anisotropy of the sample.

The damping,  $\alpha$  and inhomogeneous full width half maximum linewidth,  $\Delta H_0$  values are obtained by fitting the linewidths,  $\Delta H$  using the following equation [7,8]:

$$\Delta H = \frac{4\pi\alpha f}{|\gamma|\mu_0} + \Delta H_0$$

Where  $\Delta H$  is the experimentally measured linewidths determined by fitting the absorption due to the complex susceptibility as explained above.

Fig. 8-2 shows the experimentally extracted linewidths,  $\Delta H$  of the TmIG mode observed in the pristine TmIG and the heterostructure samples. The corresponding line fittings are shown where the slopes of the lines give the damping coefficients. The intercepts of the fitted line in the linewidth axis represents inhomogeneous linewidth broadening,  $\Delta H_0$ . The damping coefficient of pristine TmIG sample is estimated to be  $\alpha=0.0147$ . After inserting CoFeB layers, the damping of the TmIG mode enhances. The enhancement is significant for heterostructure samples with thick CoFeB,  $x=3$  nm, where the damping is increased to  $\alpha=0.0226$ . The heterostructure with thin CoFeB such as  $x=0.84$  nm and  $x=1$  nm also show damping enhancement; however, the increase is smaller ( $\alpha=0.0173$ ). We explain this in the following. When the TmIG magnetizations are resonantly excited, spins are pumped into the adjacent CoFeB layer. The spin coherence length,  $\lambda_{sc}$  in transition metal (such as Fe, Co, Ni) is a few Angstroms. Thus, most of the pumped spins out of the TmIG are absorbed in thicker CoFeB (such as  $x=3$  nm) which acts as spin sink. However, in thin CoFeB sample ( $x \leq 1$  nm), the pumped spins from TmIG are not entirely absorbed in the CoFeB layer. Thus, spin accumulation may happen in the TmIG/CoFeB interface which results in backflow of spin currents. As a result, the net spin current,  $I_s$  reduces which decreases the enhancement of damping. This is analogous to the case found in ferromagnetic and non-magnetic (NM) multilayers (FM/NM), where the spin accumulation can be quantified as the backflow factor [9],

$$\beta = \frac{1}{g^{\uparrow\downarrow} \sqrt{\frac{4\epsilon}{3} \tanh\left(\frac{t}{\lambda_s}\right)}}$$

Where,  $\epsilon$  and  $\lambda_s$  represents spin-flip scattering probability and spin diffusion length [10] in NM and  $t$  is the thickness of NM. Considering backflow, the effective spin mixing conductance is as follows [9,11]:

$$\frac{1}{g_{eff}^{\uparrow\downarrow}} = \frac{1}{g^{\uparrow\downarrow}} + \beta$$

As the thickness,  $t$  decreases the backflow factor  $\beta$  increases, which decreases the effective spin mixing conductance. Similar to the FM/NM multilayer, in FI/FM multilayer as the FM layer thickness decreases, the spin backflow increases which decreases the net spin current and subsequently the effective spin mixing conductance, thus resulting in a smaller enhancement in the damping coefficients.

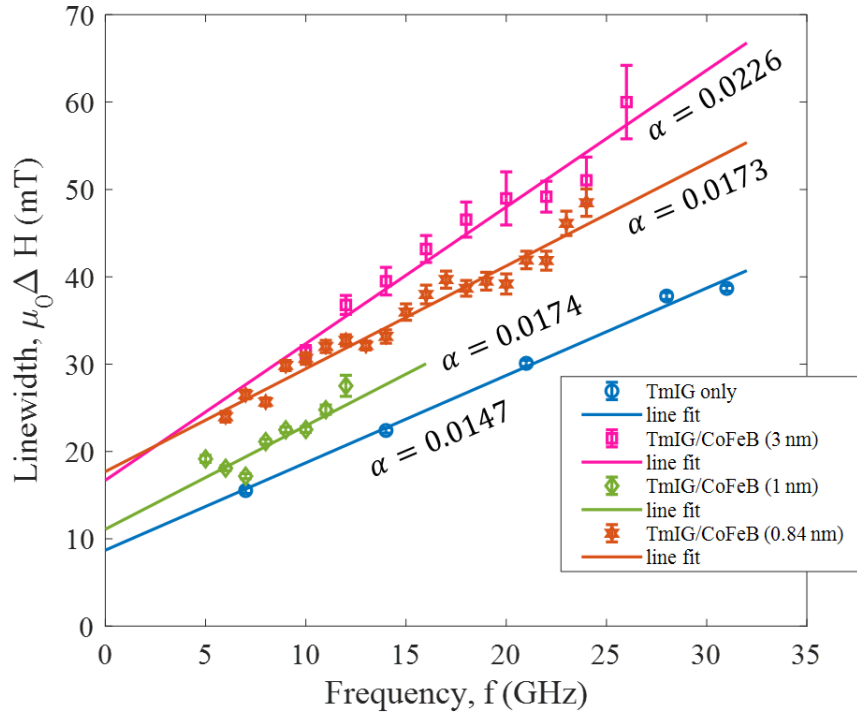


Figure 8-2 The FMR linewidths of the TmIG modes for pristine TmIG sample and the heterostructure samples with different thickness CoFeB are plotted as a function of the rf field frequencies. The corresponding line fits are shown where the slopes indicate the values of the damping coefficients. Damping enhancement of the TmIG modes in heterostructures compared to the pristine TmIG sample are evident. The enhancement is highest for samples with x= 3 nm CoFeB.

Interesting phenomena due to this dynamic coupling are observed when both of the FI and FM layers in the heterostructure are resonantly excited by tuning the angle of the external fields. Due to different anisotropy of the FI and FM, when the field angle is swept, the resonant fields match at some specific field angle. Interestingly, the damping enhancement vanishes when equal resonance conditions are satisfied. Referring to the previous LLG equation, when one of the magnetic layers become resonant while the other is not, the change in magnetization in the non-resonating layer can be extremely small,  $\frac{dm_j}{dt} \approx 0$ . This case is shown in Fig. 8-3a-b. In this scenario, the damping enhancement of the resonating layer is the highest. However, when both the magnetic layers are resonating spins are pumped into each other as shown in Fig. 8-3c. In this scenario, the damping of both resonating layers reaches close to their bulk damping values.

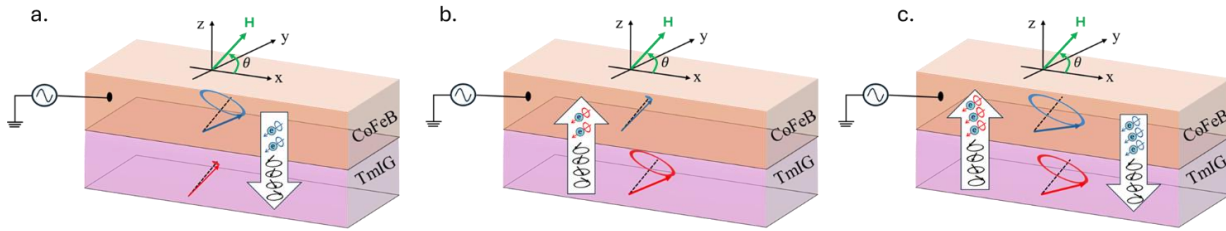


Figure 8-3 Schematics of conditions where a. FM layer magnetization is resonantly excited b. FI layer magnetization is resonantly excited and c. both of the FM and FI layers' magnetizations are resonantly excited.

This scenario is tested by sweeping the field angles from  $\theta=0^\circ$  to  $\theta=90^\circ$  and measuring the FMR spectra for FI/FM heterostructure sample with  $x=3$  nm and 1 nm CoFeB. For  $x=3$  nm sample, the FMR measurements are carried out with a rf field frequency,  $f=14$  GHz. The individual TmIG and CoFeB modes are identified from the FMR spectra for the field angle sweep. Fig. 8-4a and 8-4b show the resonant fields and linewidths for different field angles. From Fig. 7-4a we see that when the field is applied parallel to the sample plane,  $\theta=0^\circ$ , the resonant field of the TmIG is highest which decreases in a sinusoidal manner and reaches to a minimum at  $\theta=90^\circ$ . This confirms that the TmIG layer has perpendicular magnetic anisotropy (PMA). The CoFeB mode shows opposite trend which confirms the 3 nm thick CoFeB has predominant in-plane anisotropy. By sweeping the field angle, the resonance fields of the TmIG and CoFeB magnetization are tuned to match, and the resonance field becomes almost identical at  $\theta=72^\circ$ . We see a simultaneous decrease in the linewidths of both the CoFeB and TmIG mode at this crossover field angle. Specifically, the linewidth of the TmIG mode reaches 22.15 mT which is close to the linewidth of the pristine sample ( $\sim 22.40$  mT at 14 GHz). The overall trend is explained as follows. Initially the linewidths of both of the CoFeB and TmIG are seen to decrease with field angle ( $\theta < 70^\circ$ ). This is due to the fact that, before crossover, only one of the magnetic layers was resonant. Thus, the pumped spins from the resonating layers are

absorbed in the non-resonating which increases the damping and the linewidths. As the two resonant fields become identical, non-equilibrium spin currents pumped from the resonating layers almost cancel each other. Thus, the net spin current becomes nearly zero which suppresses the damping enhancements, and the linewidth of the modes almost reaches the bulk values of pristine samples. As we increase the field angle,  $\theta < 74^\circ$ , the separation in resonance fields become higher which further enhances the resonance linewidths for both of the TmIG and CoFeB modes. Similar trends are also observed in heterostructures involving FM/NM/FM in previous studies [12].

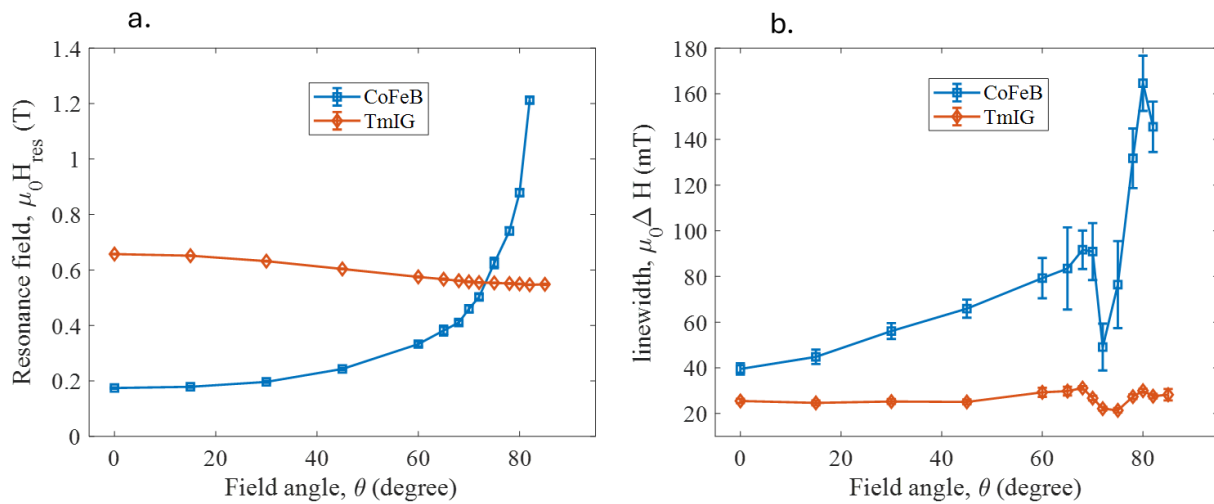


Figure 8-4 a. Resonant fields vs the field angle for the CoFeB and the TmIG modes in FM/FI heterostructure with  $x=3$  nm. b. linewidths vs the field angles of the two modes for the same heterostructures. The FMR measurements are carried out for rf field frequency of 14 GHz.

Similar trends are observed for FI/FM sample with  $x=1$  nm CoFeB. Fig. 8-5a and 8-5b show the resonance fields and the linewidths of the individual TmIG and CoFeB modes. For this sample, the resonant field crossover occurs near  $\theta=60^\circ$ . Similar to the sample with  $x=3$  nm, the linewidths of the modes become minimum at the crossover resonant field. At this equal resonance condition, the linewidth of the TmIG mode is measured to be  $\sim 22.14$  mT which is close to the pristine TmIG sample ( $\sim 20$  mT at 11 GHz).

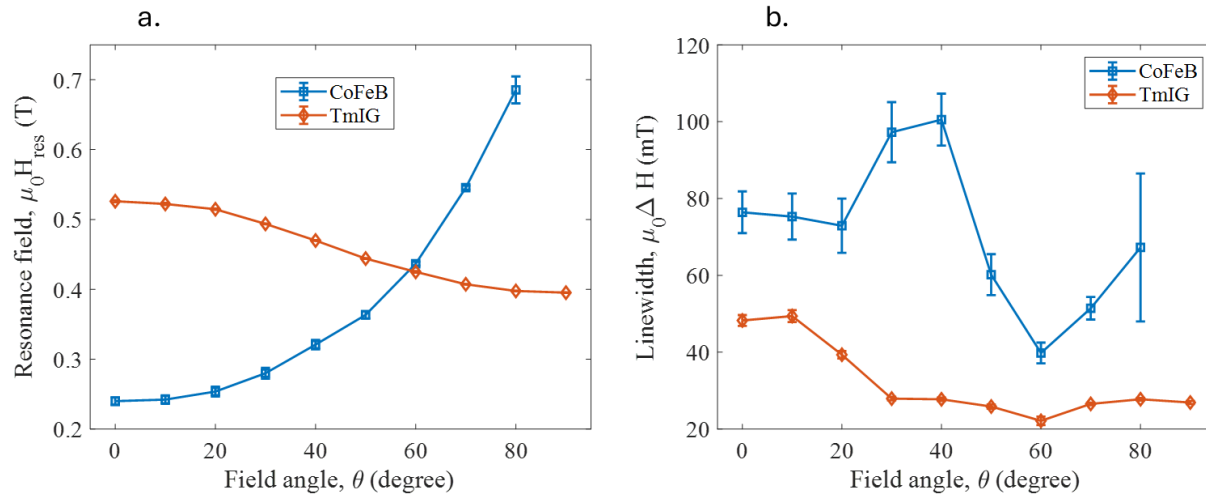


Figure 8-5 a. Resonant fields vs the filed angle for the CoFeB and the TmIG modes in FM/FI heterostructure with  $x = 1$  nm. b. linewidths vs the filed angles of the two modes for the same heterostructures. The FMR measurements are carried out for rf field frequency of 11 GHz.

### 8.3 Conclusion:

In summary, we have shown dynamic coupling in magnetic heterostructures consist of GGG/TmIG(40)/CoFeB( $x$ )/W(0.4)/CoFeB(0.0)/MgO(1)/W(5) due to the exchange of non-equilibrium spin currents. When one of the FM or FI layers becomes resonant, it emits spin currents which are absorbed by the non-resonating layer. Thus, the absorbing layers act as the nonlocal spin momentum brake by dissipating spin moments and applying a relaxation torque on the resonating layer. The damping of the resonating layer is observed to be enhanced. When both magnetic layers are tuned into equal resonance condition by adjusting the applied field angle, the damping enhancement is suppressed as confirmed from the linewidth minima. Our study shows modulation of magnetic damping is possible in heterostructures by nonlocal relaxation and dynamic coupling.

### References:

- [1] Y. Tserkovnyak, A. Brataas, and G.E.W. Bauer, Enhanced Gilbert Damping in Thin Ferromagnetic Films, Phys. Rev. Lett. 88, 117601 (2002)
- [2] A. Brataas, Y. V. Nazarov, and G. E. W. Bauer, Finite-Element Theory of Transport in Ferromagnet–Normal Metal Systems, Phys. Rev. Lett. 84, 2481 (2000)

- [3] X. Waintal, E. B. Myers, P. W. Brouwer, and D. C. Ralph, Role of spin-dependent interface scattering in generating current-induced torques in magnetic multilayers, *Phys. Rev. B* 62, 12317 (2000)
- [4] Yaroslav Tserkovnyak, Arne Brataas, Gerrit E. W. Bauer, Dynamic exchange coupling and Gilbert damping in magnetic multilayers, *J. Appl. Phys.* 93, 7534–7538 (2003)
- [5] N. Nakamura, H. Ogi, M. Hirao, T. Fukuhara, K. Shiroki, N. Imaizumi, Fast Recovery of Elastic Stiffness in Ag Thin Film Studied by Resonant-Ultrasound Spectroscopy, *Jpn. J. Appl. Phys.* 2008, 47, 3851.
- [6] R. O. Handley, *Modern Magnetic Materials: Principles and Applications*, John Wiley & Sons, New York, NY 1999.
- [7] B. Heinrich, J. F. Cochran, and R. Hasegawa, FMR line broadening in metals due to two-magnon scattering, *Journal of Applied Physics* 57, 3690 (1985)
- [8] D. B. Gopman, C. L. Dennis, R. D. McMichael, X. Hao, Z. Wang, X. Wang, H. Gan, Y. Zhou, J. Zhang, Y. Huai, Enhanced ferromagnetic resonance linewidth of the free layer in perpendicular magnetic tunnel junctions, *AIP Advances* 7, 055932 (2017)
- [9] S. Mukhopadhyay, P. K. Pal, S. Manna, C. Mitra and A. Barman, All-optical observation of giant spin transparency at the topological insulator BiSbTe<sub>1.5</sub>Se<sub>1.5</sub>/Co<sub>20</sub>Fe<sub>60</sub>B<sub>20</sub> interface, *NPG Asia Materials*, vol. 15,57 (2023)
- [10] H. Nakayama, K. Ando, K. Harii, T. Yoshino, R. Takahashi, Y. Kajiwara, K. Uchida, and Y. Fujikawa, Geometry dependence on inverse spin Hall effect induced by spin pumping in Ni<sub>81</sub>Fe<sub>19</sub>/Pt films. *Phys. Rev. B* 85, 144408 (2012).
- [11] Yaroslav Tserkovnyak, Arne Brataas, Gerrit E. W. Bauer, and Bertrand I. Halperin, Nonlocal magnetization dynamics in ferromagnetic heterostructures, *Rev. Mod. Phys.* 77, 1375 (2005)
- [12] Bret Heinrich, Yaroslav Tserkovnyak, Georg Woltersdorf, Arne Brataas, Radovan Urban, and Gerrit E. W. Bauer, Dynamic Exchange Coupling in Magnetic Bilayers, *Phys. Rev. Lett.* 90, 187601 (2003)

## Chapter 9: Conclusion and Future Works

In this thesis, we have investigated energy efficient spintronic devices for non-volatile memory and hardware AI applications. Our study on magnetization precession dynamics with surface acoustic wave (SAW) induced ferromagnetic resonance (FMR) in nanomagnets with inhomogeneity and thermal noise has shown the potential for scaling spin-transfer torque (STT) memory devices below 20 nm lateral dimensions while achieving significant reduction in write energy. We have also proposed an energy efficient strain-controlled synapse utilizing domain wall (DW) motion in chiral systems with Dzyaloshinskii-Moriya interaction (DMI), demonstrating the feasibility of multi-state synapses for edge computing applications.

Furthermore, we have developed deep neural networks (DNNs) with extremely low resolution and stochastic DW device-based synapses, achieving high classification accuracy when trained with appropriate in-situ and ex-situ learning algorithms. This technology is particularly attractive for low power intelligent edge devices in future Internet of Things (IoT) applications. We have also demonstrated long-term autonomous prediction using a skyrmion reservoir, which can perform predictions with significantly lower energy consumption compared to Long Short-Term Memory (LSTM) based approaches, making it suitable for hardware and memory constraint edge computing platforms.

Our investigation of interlayer exchange coupling between a ferrimagnetic insulator with perpendicular magnetic anisotropy (PMA) and a ferromagnetic compound has provided insights into the relative strength of competing interactions, which can be tuned by varying the thickness of the ferromagnetic layer. Strong coupling can ensure efficient readout of the ferrimagnetic insulator which has great potential in realizing high-density magnetic data storage with ultra-low energy dissipation. We have also shown dynamic coupling in ferrimagnetic insulator/ferromagnetic (FI/FM) layers due to the exchange of non-equilibrium spin currents, demonstrating the modulation of magnetic damping through nonlocal relaxation and coupling.

Lastly, we have demonstrated  $90^\circ$  switching of the magnetization easy axis in a multiferroic heterostructure using voltage-induced strain, highlighting the potential for electric field control of spintronic devices utilizing the magnetoelastic effect. This is significant for realizing energy efficient voltage-controlled spintronic devices for storage and neuromorphic implementations.

Through the comprehensive exploration of energy-efficient spintronic devices presented in this thesis, we hope to contribute significantly to the ongoing transformation of computing technology and pave the way for a more sustainable and capable technological future.

In future we are planning to pursue the followings:

### **9.1. DW-MTJ for complex neural network architectures and tasks:**

Our research has demonstrated the potential of energy-efficient DW-MTJ devices for implementing deep neural networks (DNNs). Moving forward, we plan to explore the application of our developed framework to more complex tasks using advanced neural network architectures. Recent studies from our group have shown successful classification of images from the CIFAR-10 dataset using a VGG-8 convolutional neural network with DW-MTJ [1]. By employing 5-state quantization, we achieved competitive accuracy compared to 32-bit precision weights (CMOS-based design), while realizing improvements of 9.6×, 3.5×, and 13.8× in overall energy efficiency, latency, and area, respectively. Furthermore, we have demonstrated the feasibility of using DW-MTJ for autoencoder-based unsupervised learning in anomaly detection tasks, utilizing quantized neural network learning [2]. Given the success and prospects of DW-MTJ based spintronic devices for efficient DNN implementation, our future work will focus on exploring more complex networks, such as transformers, to be implemented with DW-MTJ and other non-volatile memory technologies.

### **9.2. Demonstration of DW-MTJ based DNN:**

As a preliminary study, we patterned DW racetracks and associated electrode pads for injecting current pulses using photolithography and electron beam lithography at VCU. Our collaborators then used ion beam sputtering to deposit Pt(3nm)/Ta(3nm)/Co(0.3nm)/Ni(0.7nm)/Co(0.3nm)/Ta(0.3nm) and performed lift-off, while we also have the capability to perform the deposition at VCU. The prepared racetrack exhibits dominant perpendicular magnetic anisotropy (PMA). Figure 9-1 shows polar magneto-optical Kerr microscopy (MOKE) images of the racetrack stack when it is perpendicularly magnetized with magnetizations in the upward direction, DW nucleation and propagation under an external applied magnetic field, and when the racetrack magnetizations are pointing in the downward direction. The next step involves applying current pulses across the Pt(3nm)/Ta(3nm) layer of the electrodes, which will exert spin-orbit torque (SOT) on the racetrack magnetization to control the motion of the DW along the racetrack. We will first characterize the racetrack devices by evaluating the equilibrium DW position distribution within the racetracks in response to various amplitude and duration SOT current pulses. Based on the device characteristics, we will build a reduced-order model of our racetrack device. Secondly, using this model, we will train and optimize our neural networks. Finally, we will connect several racetracks to the contact



pads to mimic the functionality of a compact neural network and evaluate the implemented network's performance. Based on the device performance, we can further tune the neural network architecture for higher accuracy.

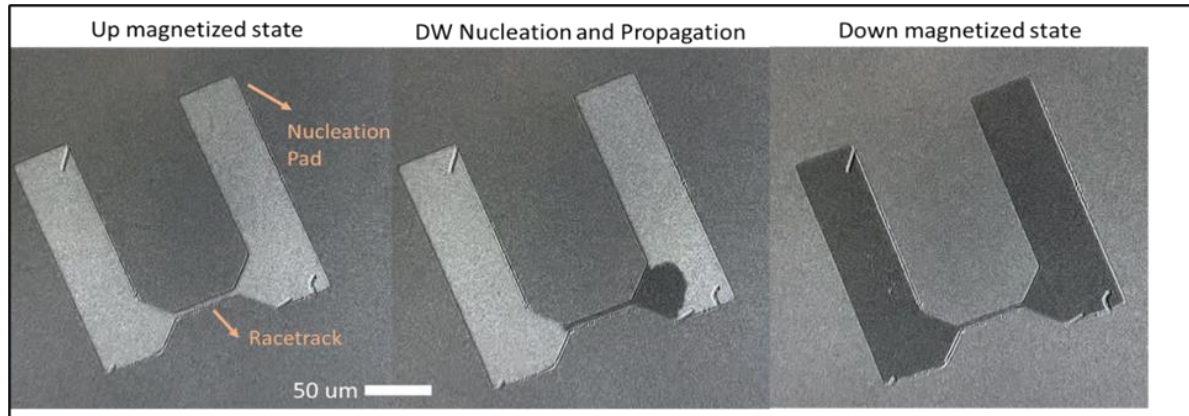


Figure 9-1 Polar MOKE images during the DW nucleation and propagation along the racetrack stacks of Pt (3 nm)/Ta(3nm)/Co (0.3 nm)/Ni (0.7 nm)/Co (0.3 nm)/Ta (0.3 nm) exhibiting perpendicular magnetic anisotropy (PMA).

## References:

- [1] S. Dhull, W. A. Misba, A. Nisar, J. Atulasimha and B. K. Kaushik, Quantized magnetic domain wall synapse for efficient deep neural networks, *IEEE Transactions on Neural Networks and Learning Systems*, (early access), 1-10 (2024)
- [2] M. S. Alam, W. A. Misba, and J. Atulasimha, Quantized non-volatile nanomagnetic synapse based autoencoder for efficient unsupervised network anomaly detection, *Neuromorphic Computing and Engineering*, 4, 024012 (2024)

## Article VI. A1: List of Journals

- [1] M. S. Alam, **W. A. Misba**, & J. Atulasimha, Quantized non-volatile nanomagnetic synapse based autoencoder for efficient unsupervised network anomaly detection, *Neuromorphic Computing and Engineering*, 4, 024012 (2024) [[link](#)]
- [2] S. Dhull, **W. A. Misba**, A. Nisar, J. Atulasimha & B. K. Kaushik, Quantized magnetic domain wall synapse for efficient deep neural networks, *IEEE Transactions on Neural Networks and Learning Systems*, (early access), 1-10 (2024) [[link](#)]
- [3] **W. A. Misba**, M. Lozano, D. Querlioz, & J. Atulasimha, Energy efficient learning with low resolution stochastic domain wall synapse for Deep Neural Networks, *IEEE Access*, 10, 84946 – 84959 (2022) [[link](#)]
- [4] M. J. Gross, **W. A. Misba**, K. Hayashi, D. Bhattacharya, D. B. Gopman, J. Atulasimha, & C. A. Ross, Voltage modulated magnetic anisotropy of rare earth iron garnet thin films on a piezoelectric substrate, *Applied Physics Letter*, 121, 252401 (2022) [[link](#)]
- [5] **W. A. Misba**, T. Kaisar, D. Bhattacharya, & J. Atulasimha, Voltage-controlled energy-efficient domain wall synapses with stochastic distribution of quantized weights in the presence of thermal Noise and edge roughness, *IEEE Transactions on Electron Devices*, 69(4), pp. 1658 – 1666. (2021) [[link](#)]
- [6] **W. A. Misba**, H. S. Mavikumbure, M. M. Rajib, D. L. Marino, V. Cobilean, M. Manic, & J. Atulasimha, Spintronic physical reservoir for autonomous prediction and long-term household energy load forecasting, *IEEE Access*, 11, 124725-124737 (2023) [[link](#)]
- [7] **W. A. Misba**, M. M. Rajib, D. Bhattacharya, & J. Atulasimha, Acoustic-wave-induced ferromagnetic-resonance-assisted spin-torque switching of perpendicular magnetic tunnel junctions with anisotropy variation, *Physical Review Applied*, 14, 014088, (2020) [[link](#)]
- [8] M. F. F. Chowdhury, **W. A. Misba**, M. M. Rajib, A. J. Edwards, D. Bhattacharya, J. Friedman, & J. Atulasimha, Focused surface acoustic wave induced nano-oscillator based reservoir computing, *Applied Physics Letter*, 121, 102402 (2022) [[link](#)]
- [9] M. M. Rajib, **W. A. Misba**, D. Bhattacharya, & J. Atulasimha, Robust skyrmion mediated reversal of ferromagnetic nanodots of 20 nm lateral dimension with high Ms and observable DMI, *Scientific reports*, 11, 20914 (2021) [[link](#)]
- [10] M. M. Rajib, **W. A. Misba**, D. Bhattacharya, F. Bhattacharya, & J. Atulasimha, Dynamic skyrmion-mediated switching of perpendicular MTJs: feasibility analysis of scaling to 20 nm with thermal noise, *IEEE Transactions on Electron Devices*, 67(9), 3883-3888. (2020) [[link](#)]

[11] M. Niknam, M. F. F. Chowdhury, M. M. Rajib, **W. A. Misba**, R. N. Schwartz, K. L. Wang, J. Atulasimha, & L. S. Bouchard, Quantum control of spin qubits using nanomagnets, *Communications Physics*, 5, 284 (2022) [[link](#)]

[12] M. M. Rajib, **W. A. Misba**, M. F. F. Chowdhury, M. S. Alam, & J. Atulasimha, Skyrmion based energy efficient straintronic physical reservoir computing. *Neuromorphic Computing and Engineering*, 2, 044011 (2022) [[link](#)]

[13] A. J. Edwards, D. Bhattacharya, P. Zhou, N. R. McDonald, **W. A. Misba**, ..., J. Atulasimha & J. S. Friedman, Passive frustrated nanomagnet reservoir computing, *Communications Physics*, 6, 215 (2023) [[link](#)]

#### **Article VII. A2: List of Conferences (Selected)**

[1] **W. A. Misba**, M. J. Gross, D. Bhattacharya, D. Gopman, C. Ross, & J. Atulasimha, Voltage Induced Strain Control of Perpendicular Magnetic Anisotropy in Yttrium Substituted Dysprosium Iron Garnets, *APS March Meeting Abstracts*, Y54. 003, (2022) [oral presentation]

[2] **W. A. Misba**, T. Kaisar, M. Lozano, D. Querlioz, C. Ross, & J. Atulasimha, Strain Controlled Domain Wall Synapse with Quantized Weights in the Presence of Thermal Noise and Edge Roughness, *APS March Meeting Abstracts*, P40. 008, (2021) [oral presentation]

[3] **W. A. Misba**, M. M. Rajib, D. Bhattacharya, & J. Atulasimha, Acoustic Ferromagnetic Resonance Assisted Spin-Torque Switching of Perpendicular MTJs, *Bulletin of the American Physical Society*, 65 (2020) [oral presentation]

[4] **W. A. Misba**, K. Hayashi, M. J. Gross, D. Gopman, C. Ross, & J. Atulasimha, Magnetic Anisotropy Modulation in Bismuth Substituted Yttrium Iron Garnet with Voltage Controlled Strain, *APS March Meeting Abstracts*, GG03. 008 (2023) [oral presentation]

[5] **W. A. Misba**, M. J. Gross, K. Hayashi, J. E. Shoup, D. B. Gopman, C. A. Ross, & J. Atulasimha, Interfacial Static and Dynamic Exchange Coupling in a Heterostructure of CoFeB and a Perpendicular Ferrimagnetic Insulator Thulium Iron Garnet, *68th Annual Conference on Magnetism and Magnetic Materials* (2023) [oral poster]

[6] **W. A. Misba**, M. S. Alam, M. M. Rajib, M. F. Chowdhury, & J. Atulasimha, Neuromorphic computing for secure IOT and medical devices, *Commonwealth Cyber Initiative Symposium*, Richmond, Virginia, April 4 (2022) [oral poster]

[7] **W. A. Misba**, M. M. Rajib, M. F. F. Chowdhury, M. S. Alam, A. Roe, D. Bhattacharya, & J. Atulasimha, Energy Efficient Nanomagnetic Devices for Non-volatile Memory and Hardware AI, *69th International Electron Device Meeting* (2023) [oral poster]