ENERGY AND AREA EFFICIENT MACHINE LEARNING ARCHITECTURES USING SPIN-BASED NEURONS

by

HOSSEIN POURMEIDANI
M.S. University of Mississippi, 2018
M.S. Islamic Azad University, 2012
B.S. Islamic Azad University, 2010

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Electrical and Computer Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term

2021

Major Professor: Ronald F. DeMara

© 2021 Hossein Pourmeidani

ABSTRACT

Recently, spintronic devices with low energy barrier nanomagnets such as spin orbit torque-Magnetic Tunnel Junctions (SOT-MTJs) and embedded magnetoresistive random access memory (MRAM) devices are being leveraged as a natural building block to provide probabilistic sigmoidal activation functions for RBMs. In this dissertation research, we use the Probabilistic Inference Network Simulator (PIN-Sim) to realize a circuit-level implementation of deep belief networks (DBNs) using memristive crossbars as weighted connections and embedded MRAM-based neurons as activation functions. Herein, a probabilistic interpolation recoder (PIR) circuit is developed for DBNs with probabilistic spin logic (p-bit)-based neurons to interpolate the probabilistic output of the neurons in the last hidden layer which are representing different output classes. Moreover, the impact of reducing the Magnetic Tunnel Junction's (MTJ's) energy barrier is assessed and optimized for the resulting stochasticity present in the learning system. In p-bit based DBNs, different defects such as variation of the nanomagnet thickness can undermine functionality by decreasing the fluctuation speed of the p-bit realized using a nanomagnet. A method is developed and refined to control the fluctuation frequency of the output of a p-bit device by employing a feedback mechanism. The feedback can alleviate this process variation sensitivity of p-bit based DBNs. This compact and low complexity method which is presented by introducing the self-compensating circuit can alleviate the influences of process variation in fabrication and practical implementation.

Furthermore, this research presents an innovative image recognition technique for MNIST dataset on the basis of p-bit-based DBNs and TSK rule-based fuzzy systems. The proposed DBN-fuzzy system is introduced to benefit from low energy and area consumption of p-bit-based

DBNs and high accuracy of TSK rule-based fuzzy systems. This system initially recognizes the top results through the p-bit-based DBN and then, the fuzzy system is employed to attain the top-1 recognition results from the obtained top outputs. Simulation results exhibit that a DBN-Fuzzy neural network not only has lower energy and area consumption than bigger DBN topologies while also achieving higher accuracy.

To Sara and my dear parents.

ACKNOWLEDGEMENTS

This work was supported in part by the Center for Probabilistic Spin Logic for Low-Energy Boolean and Non-Boolean Computing (CAPSL), one of the Nanoelectronic Computing Research (nCORE) Centers as task 2759.006, a Semiconductor Research Corporation (SRC) program sponsored by the NSF through CCF 1739635.

I would like to express my sincere gratitude to Dr. DeMara who provided an opportunity for me to join the CAL research team, and conduct my research under his supervision. He has kindly supported me during this dissertation research, and his insightful comments helped me to proceed my research along the right direction leading to several publications. I would also like to thank my committee members Dr. Vik Kapoor, Dr. Mingjie Lin, Dr. Rikard Ewetz, and Dr. Fan Yao for supporting me and my research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION AND MOTIVATION

1.1 INTRODUCTION AND RELATED WORKS

The Restricted Boltzmann machine (RBM) is one of the well-known classes of unsupervised learning approach [1]. A set of RBMs connected hierarchically can be utilized to create deep belief networks (DBNs) with outstanding learning abilities such as natural language understanding for various applications [2]. Most of the research on RBM and DBN has focused on software implementations. Albeit the software implementation of DBNs on current von-Neumann-based platforms (e.g. CPU, GPU, FPGA) provides flexibility, it incurs significant power dissipation and high latency due to inherent data communication costs, a.k.a. the "memory wall" issue. There are various hardware implementations for RBMs such as FPGAs [3], [4] and CMOS multi-core processors [5] aiming to tackle existing software limitations.

Recently, processing-in-memory based solutions using emerging non-volatile memories (NVMs) such as resistive RAM (RRAM) [6], [7] and phase change memory (PCM) [8] are set forth to be used within the DBN architecture. NVMs provide the capability of performing logic beyond data storage by bringing an intrinsic computation parallelism alleviating the data transfer bottleneck. NVMs are typically used as weighted connections interconnecting building blocks in RBMs.

The existing FPGA-based acceleration solutions show 25- 145× speedup compared to software implementations [3], [4]. However, these designs have noticeable limitations such as constrained clock frequencies, routing congestion, and resource deficiencies due to the significant embedded memory utilization for weighted connections and activation functions. In [9] optimization methods to reduce memory requirements for weights and biases are proposed. However, in order to implement each of the activation functions, a random number generator (RNG), dedicated

piecewise linear approximator (PLA), and comparators are still required which increases area and energy consumption per neuron. As an alternative method, the stochastic CMOS-based RBM implementation have been set forth [10] that takes full advantage of low-complexity of the stochastic CMOS designs to improve area- and energy-efficiency. On the other hand, such implementation seeks extremely-long bit-stream that could lead to more energy consumption and longer latencies. Besides, it requires a significant amount of Linear Feedback Shift Registers (LFSRs) to generate the uncorrelated input and weight bit-streams. Both the FPGA and stochastic CMOS implementations leverage parallel Boolean circuits such as pseudo-random number generators, adder, and multipliers to improve the performance. Such designs impose significant area and energy overheads compared with leveraging the physical behaviors of emerging devices to perform the computation intrinsically.

Within the NVM domain, Bojnordi et al. [6] proposed to leverage resistive RAM (RRAM) devices to implement vectormatrix multiplication with up to 100× speedup and 10× energy savings over single-threaded cores. In the same way, Eryilmaz et al. [8] and Sheri et al. [7] have used resistive memories with CMOS activation function that ultimately imposes excessive area and power consumption overheads. Recently, spintronic devices with low energy barrier nanomagnets such as spin orbit torque-Magnetic Tunnel Junctions (SOT-MTJs) and embedded magnetoresistive random access memory (MRAM) devices are leveraged as a natural building block to provide probabilistic sigmoidal activation functions for RBMs, as studied in [11] and [12], respectively. These devices have realized significant energy and area improvements compared to previous RBM hardware implementations. Thus, we will investigate various circuit implementations to interpolate the stochastic output of the probabilistic spin logic devices (p-bit) proposed in [13]. In particular, inspired by a technique that is used to create an analog-to-digital

converter [14], we will develop two CMOS-based probabilistic interpolation recoder (PIR) circuits, which leverage a sampling methodology to provide a digital output corresponding to the probabilistic output of the p-bit based neurons. The proposed circuits achieve significant improvements in terms of resource utilization and energy consumption compared to conventional integration followed by analog-to-digital conversion methods.

1.2 NEED FOR IN-SITU ADAPTATION FOR PROCESS VARIATION IMMUNITY

Stochastic circuits play a significant role in the implementation of networks with probabilistic nodes. For instance, learning networks employing p-bits are worthwhile in realizing DBNs in a way that weights are trained offline by a learning algorithm in software and the hardware is utilized to repeatedly perform inference tasks effectively. Unstable low barrier nanomagnets present a direct mechanism to realize stochastic sigmoidal neurons in DBNs through leveraging the randomly fluctuating magnetization to produce a stochastic time varying output voltage. If these nanomagnets are designed to have as low energy barriers that are feasible, then many random outputs are produced in a short period of time. Under this strategy, a near-zero energy barrier nanomagnet has the capability of free magnetization layer flipping back and forth which can be tuned by modulated the voltage on the gate of p-bit's NMOS transistor.

The p-bit device is not entirely tolerant of defects and device-to-device variations even though is more error resilient than strictly digital computing devices [215]. The statistical distribution of the magnetization fluctuations, such as the power spectral density become affected by the presence of both localized and delocalized structural defects and moderate variations for the barrier height of the nanomagnet which is caused by small size variations [216]. It is investigated that the power spectral density is relatively insensitive to the presence of small localized defects

and moderate barrier height change. Nevertheless, the power spectral density is substantially affected by delocalized defects such as thickness variations over a significant fraction of the nanomagnet [217][218][219]. Delocalized defects can considerably change the fluctuation rate of the magnetization in low barrier nanomagnets. This will affect applications in p-bit-based neurons for neuromorphic architectures because the fluctuation rate is essential for stochastic computing applications. Thus, the defects caused by the fabrication imperfections are required to be addressed for neuromorphic applications using p-bit based neurons such as DBNs due to their significant impact on their performance and accuracy.

1.3 CONTRIBUTIONS OF THE DISSERTATION

1.3.1 AN EFFICIENT CONVERTER TO INTERPOLATE THE PROBABILISTIC OUTPUT OF THE NEURONS IN DBNS

The concept of using sampling and count operations to interpret the probabilistic output of a p-bit based neuron offers an intriguing approach to realize a CMOS-based probabilistic interpolation recoder (PIR) for a spin-based stochastic binary neuron. Herein, we proposed a PIR circuit as a replacement for an analog-based approach to interpolate the output of the p-bit based activation functions in the last layer of a DBN circuit. The conventional method involved: first, using an RC circuit to continuously integrate the analog output of the p-bit, next an op-amp based sample and holder is used to sample the output of the RC circuit, finally the analog sampled output is converted to a digital value through an op-AMP based ADC circuit and a priority encoder. Our proposed CMOS-based PIR circuit removes the need for all of area- and energy-consuming analog components existing in conventional circuits such as resistors, capacitors, and opamps, and performs the interpolation operation only by using MOS-transistor

based Boolean gates and flip-flops. In addition, the PIR circuits have an inherent single stuck-at fault tolerant features to tolerate either transient or permanent faults at the circuit's output without redundancy or active refurbishment overhead.

1.3.2 MITIGATING THE EFFECTS OF PROCESS VARIATION ON THE PERFORMANCE AND ACCURACY OF DBNS

We investigated two approaches to mitigate the effects of process variation on the energy barrier of the p-bit based neurons, and their consequent impact on the performance and accuracy of DBNs using p-bit devices as probabilistic sigmoidal neurons. In the first approach, it was shown that an increase in the energy barrier leads to decreased fluctuation speed in the magnetization direction of the p-bit' nanomagnet. It means that in order to observe the desired probabilistic sigmoidal behavior in the p-bit based neuron a temporal redundancy is required to be added to the sampling time of the p-bits output to give it enough time to have sufficient probabilistic fluctuations. While the temporal redundancy has shown to be an efficient mechanism, it was examined that it can lead to approximately 10-fold higher energy consumption in a 784×200×10 DBN which can tolerate maximum 2 kT of energy barrier variations compared to a variation-less DBN with similar topology. The second variation tolerance mechanism proposed herein involved implementing p-bit with a negative self-feedback, which could significantly increase the probabilistic fluctuation speed of the free layer. In this case, the drain of the NMOS transistor in the p-bit device tracks the magnetization direction of the free layer of the MTJ, and the inverter at the output of the device naturally generates the inverse voltage, hence realizing a negative feedback effect which successfully compensate the variation impacts with only ~10% energy consumption overheads.

1.3.3 HIGH ACCURACY DBN-FUZZY NEURAL NETWORKS USING MRAM-BASED STOCHASTIC NEURONS

In order to benefit from high-speed DBN hardware implementations, Probabilistic Inference Network-Simulator (PIN-Sim) has been developed and used to achieve up to 3× energy reduction and 20× area reduction in comparison with the prior DBN hardware implementations, however, this framework suffers from low accuracy in image classification. In order to improve the accuracy of MRAM-based DBN, we can increase the size of network but the accuracy will be improved only by 2.5% at the cost of ∼10× higher energy consumption and significantly larger area overheads. We addressed this problem by utilizing a fast fuzzy algorithm in the interest of improving the accuracy of MRAM-based neural networks while we still can benefit from a high-speed hardware implementation. In this system, the MRAM-based DBN and the fuzzy system are working sequentially, which can be suboptimal in terms of energy as well as accuracy. In the first phase, the MRAM-based DBN is employed to identify the top recognition results with the highest probability. In the second phase, a fuzzy system is utilized to obtain the top-1 recognition results. Simulation results exhibit that a DBN-Fuzzy neural network not only has lower energy and area consumption than bigger DBN topologies but also has higher accuracy. Neuro-fuzzy systems based on spintronic devices may offer a compact and computationally-efficient architectural approach to machine-based image recognition tasks.

CHAPTER 2: MAGNETIC TUNNEL JUNCTIONS (MTJ) CHARACTERISTICS AND OPERATION

The development of spintronics [15] was a direct result of the discovery of Giant Magnetoresistance (GMR) [16],[17], which in turn resulted in numerous significant advances. The demand for feasible alternatives to minimize power leakage rises with CMOS downscaling, and one of the most promising solutions is appeared to be spin-based devices [18].

In the spintronics development, Magnetic Tunnel Junction (MTJ) has an important role [19]. An insulator layer exists between ferromagnetic layers in the structure of an MTJ. In comparison to other technologies, MTJ has low power consumption, excellent scalability, and potentially infinite endurance. Moreover, MTJ devices can be entirely turned off without data loss resulting in saving energy which cause the MTJ appropriate for various applications [19] such as analog to digital converter and memory devices. MTJ-based non-volatile memories (NVMs) have shown excellent performance by considering its advantages such as endurance and energy efficiency [20]. At highly low energy levels, MTJ-based memories work with a ten-year retention time making them appropriate for low powered applications as internet of things (IoT) applications or batteries [21]. Other applications of MTJ devices are in storage and data processing [22],[25]. MTJs can be applied to mixed signals, such as analog to digital converters and comparators [24],[25] which can be useful in radio frequency. As an illustration, spectrum-optimizing applications based on compressive sensing [26] desire to lower power consumption and reduce the area of their circuits.

Figure 1 shows the fundamental structure of MTJ devices. As shown, an insulator layer MgO divides two ferromagnetic layers. In the free layer, the magnetization direction can be switched

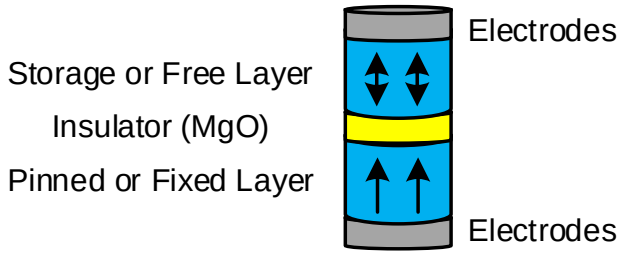


Figure 1: Magnetic tunnel junction (MTJ).

but the magnetization direction in the reference layer is unchangeable. Thus, electrical properties of the MTJ device is identified by the magnetic field. The two ferromagnetic layers' magnetization orientations (m_z) cause the two levels of the MTJ resistance to be in the high-resistance R_{AP} at an anti-parallel state or low-resistance R_P at a parallel state. The binary logic can be easily implemented by employing these two stable states of the MTJ [27]-[30]. Several methods for switching between the stable states of an MTJ have been presented. Below, three of the most prominent mechanisms for magnetization-switching are explained.

2.1 SPIN TRANSFER TORQUE (STT) SWITCHING

In order to enhance the density of the basic proposed MTJ circuits, spin transfer torque was presented in [31]. The bidirectional current I in STT device allows switch the MTJ state when I is bigger than a critical current I_{c0}. While STT enhances the scalability of the circuit, which provides a denser and simpler design, this method utilizes the same line to write and read the MTJ state which in turn leads to the issue of encountering an unexpected writing while a reading is occurring. Another drawback of the STT is that going from AP to P needs a smaller current than going from P to AP. Additionally, the application density is restricted in STT devices since a larger access transistor size is needed and increase in retention failures leads to unreliable operations [32]. The applications with high write speed have several problems since the

switching current of STT is proportional to the write pulse width in reverse [32]. The MTJ behavior model can be written as [33],[34]:

$$I_{CO} = \alpha \frac{\gamma^e}{\mu_B g} (\mu_0 M_s) H_k V \quad (1)$$

$$E = \frac{\mu_0 M_s H_k V}{2} \quad (2)$$

where the gyromagnetic ratio is γ, the magnetic damping constantis α, the elementary charge is *e*, the spin polarization efficiency factor is *g*, the Bohr magneton is μ_B, the permeability of free space is μ_0, the effective anisotropy field is H_k, the saturation magnetization is M_s, and the volume of the free layer is *V*. The average MTJ state switching delay time (*t*) can be attained as follows [33],[34]:

$$\tau = \tau_0 \exp\left(\frac{E}{K_B T}\left(1 - \frac{I}{I_{c0}}\right)\right), when\ I < I_{c0} \quad (3)$$

$$\frac{1}{\tau} = \left[\frac{2}{C + \ln\left(\frac{\pi^2 \epsilon}{4}\right)}\right] \frac{\mu_B P_{ref}}{e m_m (1 + P_{ref} P_{free})} (I - I_{c0}), when\ I > I_{c0} \quad (4)$$

where the Boltzmann constant is K_B, the attempt period is τ_0, the temperature is *T*, the thermal stability factor is ϵ, Euler's constant is *C*, the tunneling spin polarizations are P_{ref} and P_{free}, and the magnetization moment is m_m.

2.2 VOLTAGE-CONTROLLED MAGNETIC ANISOTROPY (VCMA) SWITCHING

Magnetoelectric effects have been employed in the interest of lowering the required energy consumption for switching the MTJ state [35]. By utilizing a voltage-controlled MTJ with an electric field, less area and energy consumption is achievable [36],[37]. An electric field with the

VCMA effect is utilized with the purpose of switching the MTJ state, which is when the occupation of atomic orbitals at the interface is altered through an accumulation of electron charges induced by the electric field. A change of magnetic anisotropy is obtained with this and the spin-orbit interaction [35],[38],[39].

Figure 2 exhibits the operational characterization of VCMA-MTJ. The barrier thickness increase would result in lower parasitic conductance and the effect of current-induced torques when the switching is carried out through voltage [32]. VCMA can decrease the energy barrier between the AP and P states which makes the switch of states easier. The energy barrier (E_b) between two stable magnetization states can be removed once the switching voltage V_b is more than MTJ critical voltage V_c. The minimum V_c for successful VCMA-MTJ switching is given by [40]:

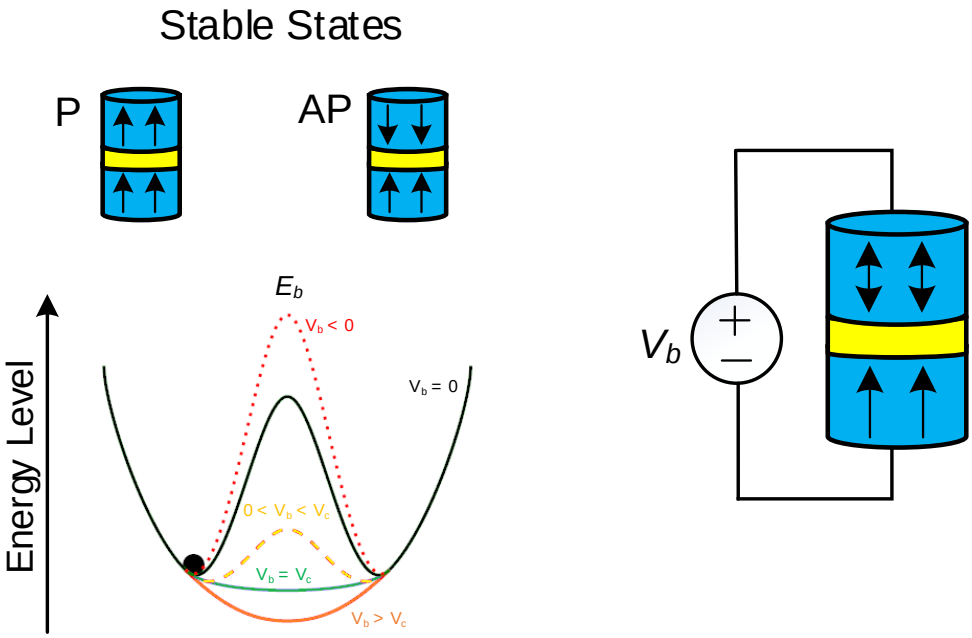


Figure 2: Structure and stable states of VCMA-MTJ device.

$$V_c = \Delta(0) k_B T t_{ox} / \xi A \quad (5)$$

where the thermal stability under zero voltage is $\Delta(0)$, k_B is the Boltzmann constant, the VCMA coefficient to weigh the perpendicular magnetic anisotropy (PMA) change under V_b is ξ, the temperature is T, the MTJ oxide layer thickness is t_{ox}, and the sectional area of the MTJ is A.

The VCMA-MTJ unstable states make its dynamics are changed consistently until V_c is obtained [41]. As soon as the excitation of MTJ terminals has finished, the energy barrier of the intermediate states returns to an amount more than the stable states resulting to stabilize the MTJ in its AP or P state. VCMA does not need large currents that lead to better scalability in its applications and less power consumption relating to STT. On the other hand, practical VCMA devices suffers from reliability issues which need to be examined more [42].

2.3 SPIN-ORBIT TORQUE (SOT) SWITCHING

A balanced switching current between the two MTJ states is allowed in the SOT devices by utilizing three terminals with the purpose of separating the read and write paths. In these devices, the read stability is increased since during the read operation, the possibility of a bit flip is decreased [32]. Spins are gathered once a current passes the non-magnetic layer and over the magnetization of the ferromagnetic layer, a torque switching is produced. Moreover, by utilizing SOT with the elimination of the time-demanding precessional motion, a faster switching can be happened [43]. On the other hand, SOT devices cannot be compatible with high-density applications since its three terminal structure causes a bigger cell size than STT-based applications.

2.4 POST-CMOS ROLES OF SPIN-BASED DIGITAL CIRCUITS

This subsection discusses some digital applications of MTJ devices.

2.4.1 MTJ-BASED MRAM

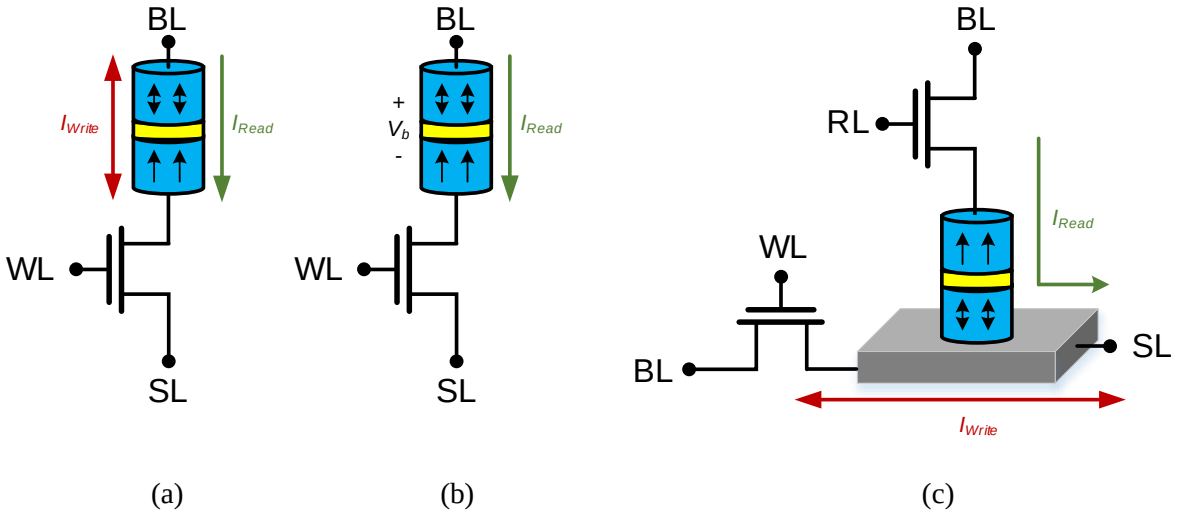
MTJ-based memories are the most famous applications of MTJ. The best characteristics of static random-access memory (SRAM), dynamic random-access memory (DRAM), and flash memory can be found in Magnetic Random-Access Memory (MRAM) [44],[45]. MTJs are considered as the primary elements in information storage by employing the difference of the MTJ resistance in its antiparallel and parallel states to represent the “1” and “0” in the binary system and the intrinsic spin of electrons as a storage unit.

According to the needed MOS transistors’ numbers and approach of write operation to build a memory cell, the MTJ-based MRAM’s structures can be different [46],[49]. Figure 3 (a) shows the bit-cell structure of spin transfer torque MTJ-based MRAM (STT-MRAM). As shown, the STT-MTJ has higher density memories since each bit-cell has only one transistor with the STT-MTJ. The MTJ state switching during the write operation is performed by the bi-directional current I_{Write} and the comparison of read current I_{Read} with a reference current defines the MTJ state [30]. It should be noted that an asymmetric write operation characterizes STT-MRAM since the current needed to switch from the P to AP state is bigger than that of switching from the AP to P state [48]. As a result, the access transistor has to be large in the interest of obtaining the requirement of the write operations’ worse case [47],[48],[50].

Relative to other techniques, STT-MRAM has intrinsic problems such as high write power and long latency [41]. However, magnetization flipping upon a voltage pulse is provided by MTJ with VCMA [36],[51]. A lower energy dissipation can be obtained by utilizing voltage instead of a charge current for MTJ write operations [41]. Furthermore, access transistor size reduction can be attained by the needed driving current reduction for the write operation [41]. By considering the switching energy and density of devices, VCMA-MTJ-based memory has better performance than STT-MRAM [35],[41],[52],[53].

Figure 3 (b) illustrates the structure of VCMA-MeRAM. This device has 1 access transistor and 1 MTJ in series like STT-MRAM. By switching the MTJ state or maintaining it, write operation can be performed. For this purpose, an extra circuit is required in order to check the state of MTJ and decide to maintain or switch the state of MTJ.

Herein, we discussed the structure of SOT-MRAM as its bit-cell design is shown in Figure 3 (c). As shown, the two access transistors in each bit-cell decreases the capacity of integration density [47],[49]. In this device, the read path and the write path are separated. The voltage applied



(a) (b) (c)

Figure 3: Memories devices: (a) STT-MRAM (b) VCMA-MeRAM (c) SOT-MRAM.

between the bit line (BL) and the source line (SL) generates the write current I_{Write} that is polarized and as a result, switches the MTJ free layer's magnetization direction. However, based on the magnitude of I_{Read}, the MTJ state is read during the read operation [54].

Recent researches claim that SOT-MRAM needs a lower write energy and a lower write time than STT-MRAM [55]-[58]. However, it is shown in [41] that VCMA-MeRAMs has better energy consumption, speed and area than STT-MRAM. Table 1 exhibits a comprehensive comparison between these devices provided in [59]. On the other hand, we need to consider that intensive development and research are being done on SOT and VCMA devices despite commercialized products are based in STT MTJ.

Recently, combination of read and write MTJ device's mechanisms have been employed for implementing memories. As an illustration, NAND-SPIN is an MTJ-based memory [47] which utilizes couple of the aforementioned switching mechanisms. As shown in Figure 4, the advantages of both SOT and STT mechanisms are taken in the interest of gaining better performance. The integration density of NAND-SPIN memory is better relative to SOT-MRAM since the transistors are shared by several MTJs. However, NAND-SPIN has better energy performance than STT-MRAM [47].

Table 1: Comparison between some MTJ-based memories.

	STT-MRAM	VCMA-MeRAM	SOT-MRAM
Read Time (ns)	1–5	1–5	1–5
Write Time (ns)	5–10	<1	<1
Cell Size (area in F^2)	40–50	20–30	50–70
Bit Density (Gb/cm^2)	1	2	0.75
Read Energy/Bit (fJ)	10–20	1–5	10–20
Write Energy/Bit (fJ)	100–200	<5	<10

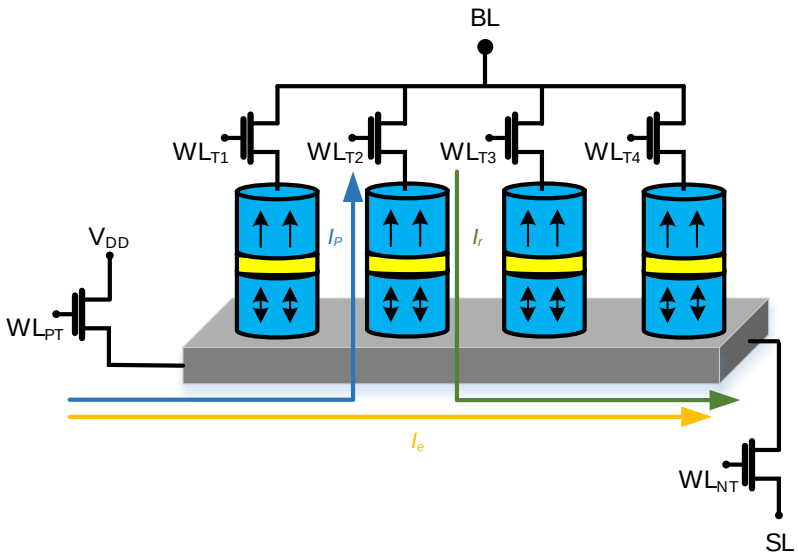


Figure 4: The structure of 4-bit NAND-SPIN.

2.4.2 NON-VOLATILE LOGIC GATES

Other applications of MTJ devices are non-volatile logic gates which provide less power consumption and area. Figure 5 shows the structure of NV-AND / NV-NAND. The NAND and AND operations are represented by $\overline{Q}$ and Q, respectively. Table 2 exhibits the truth table and Equations (6) and (7) illustrates the logic functions. For any MTJ's resistive level, this structure functions correctly.

$$Q = AB \tag{6}$$

$$\overline{Q} = \overline{AB} = \bar{A} + \bar{B} = \bar{A}B + \overline{AB} + A\bar{B} \tag{7}$$

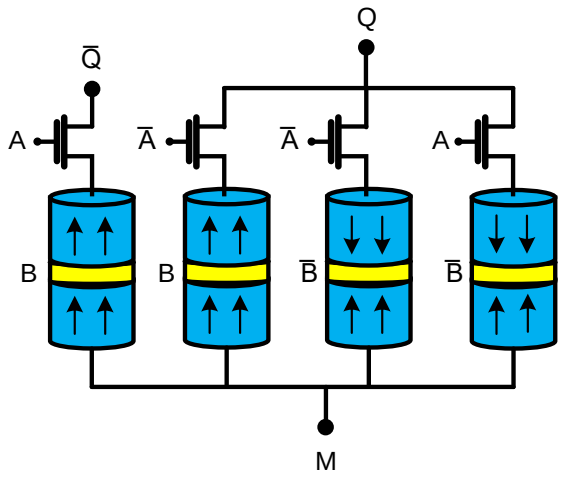


Figure 5: The structure of NV-AND / NV-NAND.

Table 2: Truth table of AND/NAND.

A	B	Q (AND)	$\overline{Q}$ (NAND)
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

Figure 6 illustrates the structure of NV-OR / NV-NOR. The NOR and OR operations are represented by $\overline{Q}$ and Q, respectively. Table 3 exhibits the truth table and Equations (8) and (9) illustrates the logic functions.

$$Q = \bar{A}B + A\overline{B} + AB \tag{8}$$

$$\overline{Q} = \bar{A}\overline{B} \tag{9}$$

Table 3: Truth table of OR/NOR.

A	B	Q (OR)	$\overline{Q}$ (NOR)
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

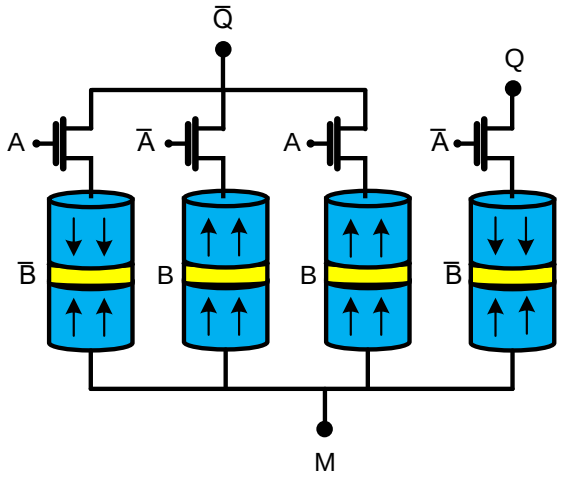


Figure 6: the structure of NV-OR / NV-NOR.

Figure 7 illustrates the structure of NV-XOR / NV-XNOR. The XNOR and XOR operations are represented by $\bar{Q}$ and Q, respectively. Table 4 exhibits the truth table and Equations (10) and (11) illustrates the logic functions.

$$Q = \bar{A}B + A\bar{B} \tag{10}$$

$$\bar{Q} = \bar{A}\bar{B} + AB \tag{11}$$

Table 4: Truth table of XOR/XNOR.

A	B	Q (XOR)	$\bar{Q}$ (XNOR)
0	0	0	1
0	1	1	0
1	0	1	0
1	1	0	1

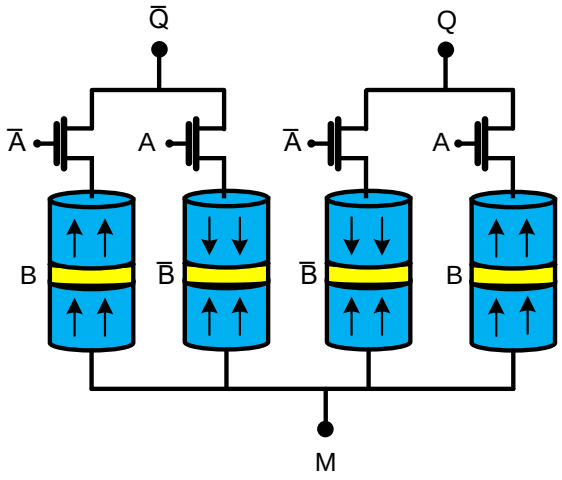


Figure 7: The structure of NV-XOR / NV-XNOR.

In [60], Deng presents some structures for logic gates that decrease the number of MTJ and NMOS transistors relative to the presented logic gate structures. On the other hand, some MTJ and NMOS settings such as their resistance configurations must be followed to be operated properly.

2.4.3 NON-VOLATILE CLOCKLESS LOOK-UP TABLE (C-LUT)

Look-Up Tables (LUTs) are one of the major FPGAs' components which are usually implemented by SRAM cells [61]. However, SRAM-based LUTs have several limitations such as high static power, volatility, and low logic density [62]. On the contrary, MTJ-based LUTs have lower mutual disturbance and power consumption since all parts are powered off except the data processing portion which is active.

Table 5 exhibits a comprehensive comparison of some MTJ-based LUTs provided in [63]. In [62], Salehi et al. employs spin Hall effect (SHE)-based MTJ for implementing a 6-input fracturable non-volatile Clockless LUT (C-LUT) for combinational logic operations without

Table 5: Characteristics of LUT designs.

Design	Write/Read Operation	Features and Challenges
FIMS-LUT [68]	Magnetic Field/TMR	High Speed High Power Consumption High Area Overhead
TAS-LUT [64]	Magnetic Field/TMR	Relatively High Speed High Power Consumption Medium Area Overhead
STT-LUT [63]	STT/TMR	High Speed Low Power Consumption Low Area Overhead
A-LUT [63]	STT/TMR	High Speed Scalable Power Consumption Low Area Overhead

requiring a clock while the proposed spin-based LUTs in [63]-[67] need a clock. This C-LUT reduces the area in comparison with the STT-MTJ-based C-LUT by removing the sense amplifier.

2.4.4 SPIN-MTJ BASED NON-VOLATILE FLIP-FLOP

Flip-flop based on non-volatile memory prevents the data loss due to system crashes and power failures. In [69], Zhao et al. proposes one of the first MTJ-based non-volatile flip-flop for System On Chip (SoC) and FPGA circuits. These circuits are fully non-volatile since all the processed data is stored in the cells of Spin-MTJ memory permanently. Figure 8 shows the full schematic of this circuit.

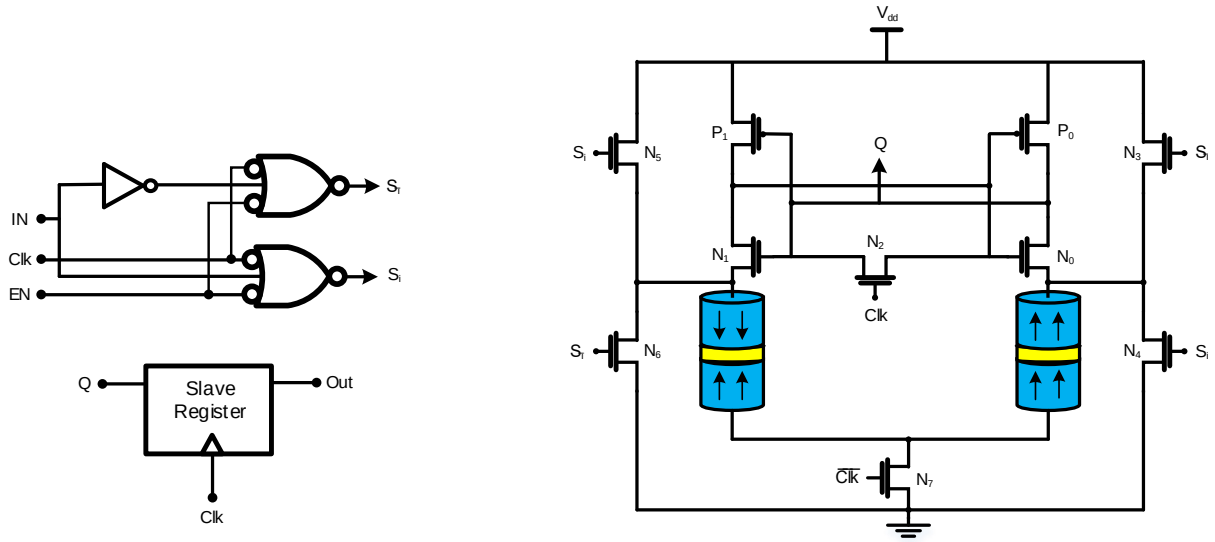


Figure 8: Spin-MTJ based Non-Volatile Flip-Flop [69].

The transistors N3, N4, N5, N6 are controlled by the NOR gates that two of them are active each time. While the circuit is in static mode, the power dissipation is reduced since the signal EN enables the current source. The pair of MTJs is written by the signal IN and then, the current direction is given. N7 switches between the reading and writing mode. The slave register keeps the prior data and the input data is stored when *Clk* = 1. However, the sense amplifier reads the stored data and the slave register updates with Q when *Clk* = 0. Several implementations of non-volatile flip-flop are available in [34],[70]-[72].

2.4.4 SPIN-MTJ BASED NON-VOLATILE FULL ADDER

Figure 9 shows the structure of a single-bit full adder including two outputs (S and C_o) and three inputs (A, B and C_i) given by Equations (12) and (13). In a CPU, high-density and low-power FA are desirable since FA is a fundamental unit to an arithmetic operation.

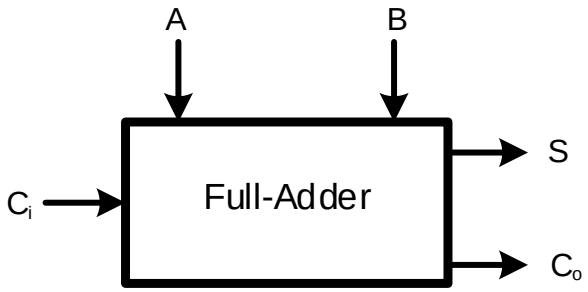
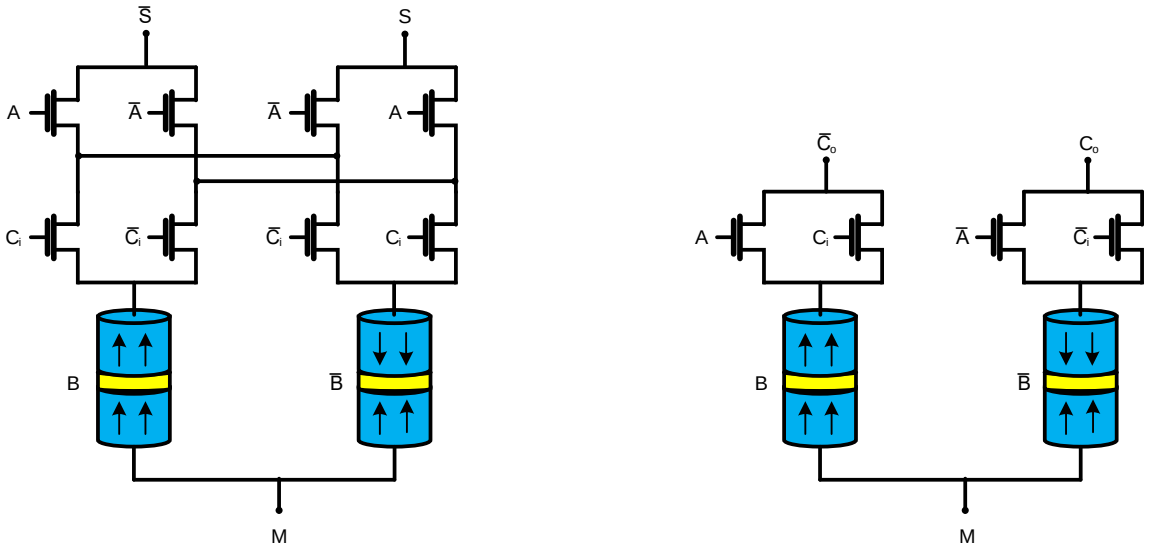


Figure 9: The schematic of Single-bit full adder (FA).

$$S = A \oplus B \oplus C_i = ABC_i + A\bar{B}\bar{C}_i + \bar{A}B\bar{C}_i + \bar{A}\bar{B}C_i \quad (12)$$

$$C_o = AB + AC_i + BC_i \quad (13)$$

The MTJ-based SUM and CARRY sub-circuits are shown in Figure 10 [60]. In [23],[73]-[75], many more non-volatile full adders were presented.



(a) (b)

Figure 10: The structure of full adder: (a) SUM sub-circuit (b) CARRY sub-circuit.

2.5 FABRICATION OF MAGNETIC TUNNEL JUNCTIONS

Researchers utilize several materials such as $Mg_3B_2O_6$ [89], ZnO [88], NaCl [87], and Al_2O_3 [86] as barrier layer material but the most frequently used material is MgO [85]. Other barrier layer material such as titanates [90],[91] and ferrites [92] have been utilized too. Several groups investigate barriers with graphene as insulating layer [93],[94]. The most popular ferromagnetic layers are CoFeB, Co, and Fe [95]. In the past few years, several studies broadly have been done in MTJ with rare earth metals [98] and heusler alloys [96],[97]. Additionally, ferromagnetic electrodes in these devices' fabrication is being implemented by MTJ with magnetic oxides [99] [100][101].

Lower resistance (R_P) will be achieved in the barrier's tunneled electrons once spins have the same orientation in lower and upper ferromagnets. The opposite spin-orientation in two ferromagnets results in higher electrical resistance (R_{AP}). Therefore, such combination of an insulator and ferromagnets demonstrate spin-orientation dependent electrical behavior, which is not allowed in regular magnetic materials or conducting [102][103].

Magneto-resistance (MR) effect defines as a change in the device's electric resistance with an applied magnetic field's effect once spin-orientation of such a combination of insulator and ferromagnets is controlled by applying magnetic field [84][86]. This effect is recognized as TMR effect since these devices' MR effect is relevant to tunneling phenomena [84]. As a result, these devices demonstrate tunneling magnetoresistance (TMR) that relies on the barrier layer's thickness and the ferromagnetic layer's spin polarization [104]. Several successful effect of such combinations have been experimentally demonstrated in various groups[84][98]. Moreover, this

effect's technological development can be observed by many memory devices' successful implementation [76][83].

These devices are on the basis of the spin dependent tunneling principle. Huge value of TMR is achieved from insulating barrier's spin dependent tunneling. Theoretically, a TMR value of ~1000% was achieved [106][107] but experimentally, a value of 604% could be attained so far [108]. These devices' deposition is a difficult task in spite of interesting phenomena in these devices. In this subsection, we will explain these devices' fabrication, but before, we will focus on the junction area's importance.

2.5.1 JUNCTION SIZE

In future storage technology, the possibility to form micrometer-scale junctions is one essential condition for the MTJ's application as read head sensors or MRAM elements. Working for longer duration and obtaining low dimensions is another challenge for these devices' fabrication. It is shown that for data retention over 10 years, a high enough thermal stability can be achieved for magnetic cells with size of nanometer [93][98][108][109]. By utilizing lithography process, MTJs are fabricated down to the nanometer levels' dimensions [110][111].

For a specified set of barrier parameters, a memory application should have a definite value for the junction area's resistance. Therefore, the product of barrier's resistance and junction area, resistance area (RA) product, is defined to understand the MTJ's characteristics[112][113]. On the basis of the operational requirements on access time and noise, an upper limit of around 20 $k\Omega$ μm^2 to the RA product of MRAM cells is set [114][116]. Thus, researchers assess TMR of these devices along with RA product [117][121]. It is reported RA = 1 $k\Omega\ \mu m^2$ and TMR = 12% in RF plasma oxidized Al barrier [117], RS = 2 $k\Omega\ \mu m^2$ and TMR = 12% in natural in situ

oxidation [118], and RS = 960 $\Omega\ \mu m^2$ and TMR = 6% in multiple oxidation of successive layers of Al in O_2 [119]. In order to minimize the junction area, lithography can be considered as an alternate solution by considering these several issues. Therefore, the MTJ's fabrication is a two-step process: (1) growth of multilayer structures by utilizing appropriate deposition technique, and (2) grown multilayers are fabricated into devices. We have explained these steps in the next subsections [101].

2.5.2 GROWTH OF MULTILAYER STRUCTURE

One of the major steps in the MTJ's fabrication is growth of a multilayer structure. For deposition of multilayer structure, we need dedicated deposition chamber with ultra-high vacuum (10^{-10}–10^{-11} Torr) [96][100].

2.5.3 MOLECULAR BEAM EPITAXY (MBE)

The most popular tool for the MTJ's deposition is MBE. This tool is efficient in maintaining different layers' orientation and stoichiometry. For this reason, researchers mostly utilize this technique to grow MTJ. In this technique, $CaTiO_3$ as insulating layer and trilayer heterostructure $La_{1-x}Sr_xMnO_3$ as the ferromagnet was utilized. This structure demonstrates magnetoresistance ΔR/R(H)ΔR/R(H) of as much as 450% in 200 *Oe* applied field at 14 *K* which persists up to ∼250 K [122]. In [123], RA product of the order of $10^6\ \Omega$–μm^2 and TMR value ∼120% are reported by depositing Fe/MgO/Fe structure with MBE method. In [124], RA product of few $k\Omega$–μm^2 and MR ratio of 88% (T = 293 *K*) are reported for the structure of fully epitaxial Fe/MgO/Fe MTJ with this technique. For Fe(001)/MgO(001)/Fe(001) junctions, RA product of 25 $k\Omega$–μm^2 and TMR value 180% are reported by this group [105].

Likewise, several research groups utilize this technique to fabricate Fe/MgO/Fe MTJ with considerable values of TMR [125][128]. For MgO(100)/Fe/MgO/Fe/Co/Pd MTJ with thickness of 0.8 nm for insulating layer, these junctions show a very small interlayer magnetic coupling, TMR up to 17%, and a low resistance around 4 $k\Omega$–μm^2 [129]. Several other groups also employ this technique for MTJ structure growth such as Co/MgO/Co tunnel junctions [133], MgO–EuO composite tunnel barriers [132], heusler based MTJ [131], and Fe/MgO/Gd [130].

2.5.3.1 E-BEAM EVAPORATION

E-beam evaporation method can be utilized for growing Fe/MgO/Fe structures [84][134][135]. This set-up permits online monitoring of thickness by utilizing quartz crystal monitor and substrate heating up to 500 $°C$ [136]. In [134][135], the details of these structures' growth procedure are reported. These structures do not show the formation of perfect interfaces but instead of that show the presence of Fe-oxides at interfaces [137]. Interface oxidation can be resulted from vacuum level of the order of 10^{-8} Torr by utilizing this technique and growing MgO/Fe/MgO structures [140] and Fe/MgO/Fe/Co MTJ structures [138][139]. We should avoid this condition for the good quality MTJ's fabrication.

2.5.3.2 SPUTTERING DEPOSITION

Another choice for growing these structures is a combination of RF and DC sputtering since a typical MTJ multilayer structure consists of insulating and ferromagnetic layers. For the deposition of CoFeB/MgO/CoFeB structure, a Six-Gun RF sputtering set-up is utilized in [141][142].

This technique recently is efficiently utilized for the multilayer structure's deposition of this CoFeB based MTJ. Sputtering method is utilized for growing MgO tunnel barriers with CoFe

electrodes which demonstrate TMR values of up to about 300% at low temperatures and almost 220% at room temperature [143]. In [101] is shown that these structures demonstrate crystallinity from well distinguishable lattice for each layer.

Several researchers used this technique to develop MTJ with varying compositions of ferromagnetic CoFeB electrodes while these MTJs demonstrate highest value of TMR [144][148]. A value of TMR around 604% at room temperature employing this technique is reported in [108] which is the TMR maximum value has been reported so far. Moreover, this group has achieved a low switching current of 49 μA in Ta/CoFeB/MgO/CoFeB/Ta, high thermal stability at dimension as low as 40 *nm* diameter, and TMR value of the order of 120% by employing this technique. Therefore, these parameters try to use these MTJs in spintronic devices [149].

2.5.3.3 ION BEAM SPUTTERING DEPOSITION

To deposit these structures, Ion beam sputtering has been utilized by several research groups [150][155]. TMR values up to 110%, with RA products of 100–400 $\Omega\ \mu m^2$ for CoFeB/MgO/CoFeB MTJ have been reported by these authors. NiFe/Mg/MgO/CoFe MTJ is grown by utilizing this technique by Singh and Chaudhary [152][154]. These authors report a TMR value of 1% for ion beam sputtered MTJ [153].

To deposit multilayer structure, several famous methods has been discussed in this subsection which further fabricate these devices in well-defined junctions by using lithography process. For growing oxide heterostructures in context of MTJ, several other deposition techniques such as atomic layer deposition [158][161] and pulsed laser deposition [156][157] are also employed by researchers.

2.5.4 LITHOGRAPHY

For the semiconductor devices' fabrication, one of the famous phenomena in electronic industry is Lithography [162][163]. E-beam lithography is employs even though in semiconductor industry, optical lithography meets the device size requirement [164]. In [165][166], e-beam lithography is preferred through advanced etching procedure and in [164][166], e-beam lithography is preferred to design devices free from mechanical damage and chemical impurity. Two lithography types are defined based on the fabricated device's size: (1) Nanolithography features smaller than 100 *nm* and (2) Microlithography for growing features smaller than 10 μm. MTJs with junction size scaling down to few *nm* are under fabrication through the development of device requirements and technological advances. Therefore, the MTJ's fabrication is considered as a subcategory of nanolithography.

2.5.4.1 PHOTOLITHOGRAPHY

One of the approaches that often used for microchips' semiconductor manufacturing is Photolithography. For fabricating micro-electro-mechanical-systems (MEMS) devices, photolithography is usually utilized too. Several steps are required to fabricate device from layer grown on substrate (wafer) in typical lithography process. These steps can be outlined as follows [101]:

I. As explained in Radio Corporation of America [167], surface layer cleaning.
II. Heating the wafer surface to drive off any moisture that may be exist.
III. Application of photoresist by spin coating.
IV. Exposure of photoresist by pattern of intense light.
V. Etching.

VI. Photoresist removal.

Therefore, the two most critical steps in lithography process are: (1) exposure and (2) etching. The photoresist removal from layer is Etching. Chemical etching is a cost-effective and simple technique of etching by utilizing chemicals [168][169]. Several different techniques are used for etching procedure these days. These techniques are ion beam milling [172], reactive ion etching [171], and plasma etching [170]. While wet etching is considered as chemical etching procedure, these approaches are usually considered as dry etching.

Other principal step is exposure while the lithography's category is defined by the radiation's nature utilized for exposure. Photolithography is a process that ultra-violet radiation is employed for exposure. This technique is used for both MTJ fabrication and semiconductor devices' fabrication. Photolithography has been used in several researches in micron-sized Fe/MgO/Fe [174], $Co_{75}Fe_{25}/Al_2O_3/Co_{75}Fe_{25}$ [173], and Ni–Fe/Al_2O_3/Co [172] junction. Chen et al. employed this technique to grow micron-sized junction on flexible substrate [175][176]. The MTJ fabrication is a regular process while the layers' number for pattering is more than three and each layer requires pattering. Patterning of Co, Al_2O_3, and NiFe layers is needed for NiFe/Al_2O_3/Co fabrication. The MTJ device patterning is a difficult job and requires a lot of expertise since each layer's patterning goes through several steps.

2.5.4.2 E-BEAM LITHOGRAPHY

Electron beam lithography has the ability of much greater patterning resolution. In the manufacture of photomasks, electron beam lithography is essential too. Electron beam lithography is kind of maskless lithography which a mask is not needed to produce the ultimate pattern. Instead, through controlling an electron beam while scans across a resist-coated

substrate, the ultimate pattern is made directly from a digital representation on a computer. The drawback of electron beam lithography is that it is much slower than photolithography. As a result, e-beam lithography is used in most of the MTJ uses [140][155]. This process is used to fabricate $Ni_{80}Fe_{20}/Co_{75}Fe_{25}$/Al-O/$Co_{75}Fe_{25}$/Ta MTJ. A part of the MTJ structure is removed by utilizing Ar ion milling and e-beam lithography in this procedure. Then, Pt layer of thickness 10 *nm* was vacuum-evaporated obliquely on both sides of Al_2O_3/Cu films, substrate was covered with a thick Al_2O_3/Cu film and liftoff in organic solvent. At the end, Ar ion milling defines junction area, in which the Pt films were utilized as etching masks with 100 $\mu m \times 10$ *nm*.

2.5.5 PATTERNING OF FE/MGO/FE SYSTEM

Recently, ion milling is preferred over chemical etching in case of etching in spite of significant results in chemical etching. We first discuss the Fe/MgO/Fe/Au MTJ procedure utilizing e-beam lithography and chemical etching in the interest of understanding the MTJ fabrication phenomena. The detailed information of this lithography procedure is given by Fe/MgO/Fe MTJ. As depicted below, the process containing almost 21 steps which are demonstrated in terms of different steps for better understanding [101]:

1. Multilayer structure growth.
2. Deposition of photoresist (PR) on the structure.
3. PR Masking.
4. Radiation exposure on PR.
5. Desired structures are grown in PR.
6. Au layer etching.
7. Fe layer etching using appropriate etching agent.

8. MgO barrier layer etching.
9. Photoresist removal.
10. PR Deposition.
11. Utilization of another mask to define exposure to radiation and lower layer.
12. After exposure structures are formed through PR.
13. Lower electrode etching.
14. PR removal.
15. Silica deposition for insulation among several devices.
16. PR deposition.
17. Mask to exposure through radiation and grow contact pads.
18. Structure formation for lower contact pads.
19. Formation of structure and silica etching for contact pads.
20. Contact layer deposition and PR removal.
21. PR deposition and exposure through mask to deposit contact pads of around 1 *mm*.

2.5.6 FABRICATION OF DEVICE USING PSEUDO/METAL MASKING PROCEDURE

To fabricate these devices, we need to pass through a complicated lithography process. These devices' properties can be affected through a number of such treatments. Thus, researcher employs the technologies that are free from complicated lithography process. Pseudo-masking is a technology that is utilized by researchers at the beginning of this technology as these junctions' fabrication. In [177][178], this methodology is employed to develop the procedure for growing CoFe/Al_2O_3/Co and CoFe/Al_2O_3/NiFe junctions. At room temperature, these junctions could achieve approximately 11% of TMR.

To minimizing magnetic coupling, the Jullieres's observation of TMR in Fe/Ge/Co junction has been employed in this procedure through oxidizing Ge in dry oxygen [179]. To deposit Gd/Gd_2O_3/NiFe junctions, metal masking is used by this group [180]. As a replacement for metal mask that allows in situ deposition of MTJ, Ootuka et al. reported fabrication method on the basis of Si_3N_4 membrane [181]. As reported by Julliers [179] and Moodera group [177][178], this process also prevents the situation where vacuum breaking is needed for designing junctions. Although mask formation requires e-beam lithography, the pseudo-masking fabrication methods to grow MTJ without lithography process. In [67], pseudo masking is employed by Perkin et al. employed in the interest of fabricating these devices, which results in approximately 220% TMR value at room temperature. Thereafter, Barraud et al. used this technique for the purpose of growing $Co/Al_2O_3/Co$ junction on organic substrate and Si [182] and on Kapton substrate by another group[183].

As described, MTJ fabrication is a multistep and complicated procedure. The most significant steps are choice of radiation, etching, and masking. In the lithography process a substantial improvement has been achieved through the technological advancement that makes good quality junction with smaller junction size and perfect shape. Researchers utilize various lithography techniques such as free electron lithography [186], X-ray lithography [185], and ion beam lithography [184]. Because of typical instrumentation associated with these techniques, the utilization of these techniques is restricted [187][188].

CHAPTER 3: BACKGROUND

3.1 DEEP BELIEF NETWORK (DBN)

DBN can be easily realized by stacking Restricted Boltzmann machines (RBMs), which are classes of recurrent stochastic neural networks, in which state of the network, *k*, has an energy expressed by (1), determined by the connection weights between nodes and the node bias, where s_i^k denotes the state of node *i* in *k*, b_i represents the bias, or intrinsic excitability of node *i*, and w_{ij} is the weight of connection between nodes *i* and *j* [189].

$$E(k) = -\sum_i s_i^k b_i - \sum_{i<k} s_i^k s_j^k w_{ij} \tag{14}$$

The probability of each node in a RBM to be in state one is determined based on (2), where σ denotes the sigmoid function. RBMs can reach a Boltzmann distribution in which the system probability to be in state *v* is represented by (3), and *u* could be any possible system state. Therefore, given sufficient time, the system moves towards the states with the lowest associated energy.

$$P(s_i = 1) = \sigma\left(b_i + \sum_j w_{ij} s_j\right) \tag{15}$$

$$P(v) = \frac{e^{-E(v)}}{\sum_u e^{-E(u)}} \tag{16}$$

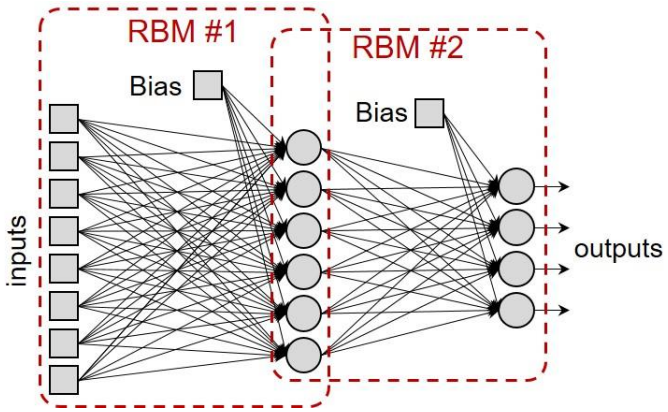


Figure 11: An example of DBN structure including a visible layer

RBM consists of two fully-connected layers called the *visible layer* and the *hidden layer*, as shown in Figure 11. Crossbar architecture is a widely-explored method to implement such networks. The most known method for training RBMs is contrastive divergence (CD), which is based on approximate gradient descent procedure using Gibbs sampling [190]. CD operates in four steps as described below:

I. Feed-Forward I: The hidden layer, *h*, is sampled based on the applied training input vector, *v*, to the visible layer.

II. Feed-back: The generated input (v') is sampled based on the sampled hidden layer output which is fed-back to the network.

III. Feed-Forward II: The reconstructed hidden layer, h', is sampled by applying v' to the visible layer.

IV. Update: The weights are updated according to Equation (17), where *W* is the weight matrix and η is the learning rate.

$$\Delta W = \eta(\mathrm{v}h^T - v'h'^T) \tag{17}$$

Algorithm 1: Contrastive Divergence Unsupervised Learning Algorithm

Input: train dataset (D_{train}), # of training samples (S), # of RBMs (M)
Output: weight (n).mat, bias(n).mat, where n is the RBM number
Require: Maximum iteration (MaxIter), Learning Rate (η)
for i= 1:S **do**
  v = D_{train}(i);
  for j= 1:M **do**
    for k= 1:MaxIter **do**
      Feed-Forward 1: h = $\sigma\ (b + \sum w.v)$;
      Feed-Back: v′ = $\sigma\ (c + \sum w.h)$;
      Feed-Forward 2: h′ = $\sigma\ (b + \sum w.v')$;
      Update:
      $\Delta W(j) = \eta\ (vh^T - v'h'^T) \Rightarrow W(j) = W(j) + \Delta W(j)$
      $\Delta B(j) = \eta\ (h - h') \Rightarrow B(j) = B(j) + \Delta B(j)$
      $\Delta C(j) = \eta\ (v - v') \Rightarrow C(j) = C(j) + \Delta C(j)$
    end
  end
end
for j= 1:M **do**
weight(j).mat ⇐W(j) ;
bias(j).mat ⇐ B(j) ;
end

A DBN can be formed by stacking RBMs and trained similarly to RBMs. The visible layer and the first of the hidden layers within the network are trained first with CD. Then, the CD is repeated as much as needed, which will adjust the weights in a hierarchical flow as described in Algorithm 1.

3.2 EMBEDDED MRAM-BASED NEURON

In this subsection, we show how a recently-proposed building block based on embedded MRAM technology can realize a neuron with probabilistic sigmoidal activation function [13]. The MRAM-based stochastic device (p-bit) structure is shown in Figure 12. It consists of a magnetic tunnel junction (MTJ), which is a 2-terminal device with two possible resistive levels based on the orientation of its ferromagnetic (FM) layers, i.e. *fixed layer* and *free layer*. The fixed layer has a fixed magnetic orientation, while the free layer's magnetization orientation can be

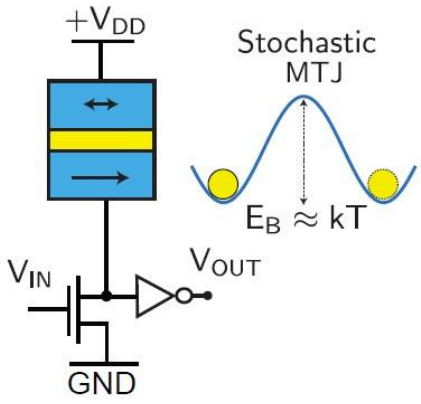


Figure 12: The diagram of the embedded MRAM-based neuron

switched. In conventional MRAM cells, free layer of the MTJ is manufactured with a thermally-stable nanomagnet with a large energy barrier with respect to the thermal energy (kT). Accordingly, the fixed layer works as a non-volatile storage. Recently, in search of functional spintronic paradigms, thermally-unstable MTJs based on superparamagnetic materials have been theoretically and experimentally explored [191], [192], [193], [194], [11], [195], [196], [197], [198].

In this work, we use a thermally-unstable MRAM device with a low energy-barrier nanomagnet (E_B << 40 kT) [13]. The MTJ resistance of this device randomly fluctuates between the two possible resistive states. This leads to a fluctuating output voltage at the drain of the NMOS transistor connected to a CMOS inverter. The inverter amplify such voltage deviation from the threshold voltage and generate a stochastic output modulated by the input voltage. Particularly, by reducing the drain-source resistance (r_{ds}) through increasing the input voltage (V_{IN}), the voltage at the drain of the NMOS transistor is shorted to the ground. Alternatively, it can get to V_{DD} by increasing the r_{ds} through decreasing V_{IN}. Such device operation is formulated considering the MTJ conductance [13]:

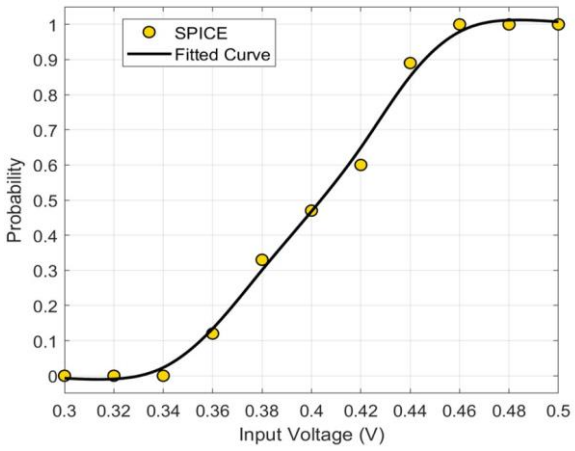


Figure 13: Output probability of MRAM-based neuron vs. its input voltage.

$$G_{MTJ} = G_0\left[1 + m_z \frac{TMR}{(2+TMR)}\right] \tag{18}$$

where m_z is the free layer magnetization, G_0 denotes the average MTJ conductance, (G_P+G_{AP})/2, and TMR represents the tunneling magnetoresistance ratio. The drain voltage can be written as:

$$V_{DRAIN}/\,V_{DD} = \frac{(2+TMR) + TMR\, m_z}{(2+TMR)(1+\alpha) + TMR\, m_z} \tag{19}$$

where α is the ratio of the transistor conductance (G_T) to the average MTJ conductance (G_0).

The p-bit device uses a circular nanomagnet with near-zero energy barrier without shape anisotropy. The free layer magnetization for the MTJ conductance discussed in Equation (18) is given by the stochastic Landau-Lifshitz-Gilbert (LLG) equation:

$$(1 + \alpha^2)d\hat{m}/dt = -|\gamma|\hat{m} \times \vec{H} - \alpha|\gamma|\left(\hat{m} \times \hat{m} \times \vec{H}\right) + 1/qN\left(\hat{m} \times \overrightarrow{I_S} \times \hat{m}\right) + (\alpha/qN(\hat{m} \times \overrightarrow{I_S})) \tag{20}$$

where α is the damping coefficient of the nanomagnet, γ is the electron gyromagnetic ratio, q denotes the electron charge, and $\overrightarrow{I_S}$ is the spin current incident to the free layer. Figure 13 shows

Table 6: Various hardware implementations for DBN architecture.

Design	Weighted Connection	Activation Function	Energy per Neuron	Normalized Area per Neuron
[4]	Embedded multipliers	▪ 2-kB BRAM ▪ PLA ▪ RNG	~10-100 nJ	~3000×
[10]	▪ LFSR ▪ AND/OR gates	▪ LFSR ▪ Bit-wise AND ▪ Tree adder ▪ FSM-based *tanh*	~10-100 nJ	~90×
[6]	RRAM	▪ 64×16 LUTs ▪ Pseudo Random Number Generator ▪ Comparator	~1-10 nJ	~1250×
[8]	PCM	Off-chip	N/A	N/A
[11]	SOT-DWM	near-zero energy barrier SOT-MTJ	~1-10 fJ	~1.25×
[12]	Memristive Devices	Embedded MRAM-based Stochastic Neuron	~10-30 fJ	1×

the correlation between the probability of output being in state "1" and V_{IN}. A close observation shows that $V_{IN} = V_{DD}/2 = 400$ mV produces an output probability of 50%.

Some of the most recent hardware implementations of DBNs are listed in Table 6. In [4], FPGAs are utilized to achieve speedups of 25-145 in comparison with software implementations, however they still suffer from constrained clock frequencies and routing congestion along with substantial resource deficiencies because of the significant embedded memory utilization for both weighted connections and activation functions. The design presented in [10], benefits the low-complexity characteristics of stochastic CMOS-based arithmetic for implementing RBMs with reduced area and power consumption but the increased latencies in this design significantly restrains the energy savings due to the enormous number of linear feedback shift registers (LFSRs) that are required to generate the long input and weight bit-streams. In [6] and [8], the crossbar arrays have been employed with emerging technologies such as resistive RAM (RRAM) and phase change memory (PCM) to implement matrix multiplication within RBMs. In [6], Bojnordi et al. have employed RRAM devices as weighted connections to achieve 100-fold and

10-fold improvement with respect to operation speed and energy consumption, respectively, relative to single-threaded cores. The CMOS-based circuits such as multipliers and RNGs are employed in all the aforementioned designs to realize the probabilistic behavior of activation functions, which results in significant area and energy overheads. In [11], Zand et al. have achieved substantial area and energy reductions by employing low energy barrier spin-orbit torque (SOT) MTJs to implement the probabilistic sigmoidal activation function. Nevertheless, this design requires weighted connections with very large resistance values which results in considerable area overhead and fabrication complexity. Moreover, the current-mode behavior of the SOT-MTJ devices imposes considerable power consumption to the activation functions [199], [200]. Voltage-driven embedded MRAM-based neuron with low energy barrier (p-bit) has been proposed to take advantage of intrinsic thermal noise to generate sigmoidal probabilistic activation functions required for RBMs [12]. As listed in Table 6, the p-bit based RBM implementation can attain approximately three orders of magnitude energy reduction relative to the previous energy-efficient CMOS-based implementations, as well as at least 90- fold decrease in the CMOS device count.

3.3 PROBABILISTIC INFERENCE NETWORK-SIMULATOR (PIN-SIM)

Herein, we use the Probabilistic Inference Network-Simulator (PIN-Sim) proposed in [12] to realize a circuit-level implementation of DBNs using memristive crossbars as weighted connections and embedded MRAM-based neurons as activation functions. As shown in Figure 14, PIN-Sim is a hierarchical simulation framework that consists of five main modules: (1) trainDBN: a MATLAB-based module used for training various DBN topologies [201] (2) mapWeight: a module developed in MATLAB that converts the trained weights and biases to their corresponding resistance values, (3) mapDBN: a python-based module which provides a

circuitlevel implementation of the restricted Boltzmann machine using the obtained weight and bias resistances, (4) neuron: a SPICE model of the MRAM-based stochastic neuron [13], (5) testDBN: the main module developed in Python that executes test evaluations to assess the error rate and power consumption using the other modules in PIN-Sim.

As described in Algorithm 2, a MATLAB implementation of DBN is modified to train the network and obtain the bias (B) and trained weight (W) matrices. Then, the extracted bias and weight matrices are applied to *mapWEIGHT* as a MATLAB module to convert both the positive and negative elements in bias and weight matrices to two separate matrices with only positive elements as described in Algorithm 3 and below:

$$w^{+}_{(i,j)} = \begin{cases} w_{(i,j)}, & if\ w_{(i,j)} \geq 0 \\ 0, & if\ w_{(i,j)} < 0 \end{cases}, \qquad w^{-}_{(i,j)} = \begin{cases} 0, & if\ w_{(i,j)} \geq 0 \\ -w_{(i,j)}, & if\ w_{(i,j)} < 0 \end{cases} \tag{21}$$

$$b^{+}_{j} = \begin{cases} b_j, & if\ b_j \geq 0 \\ 0, & if\ b_j < 0 \end{cases}, \qquad b^{-}_{j} = \begin{cases} 0, & if\ b_j \geq 0 \\ -b_j, & if\ b_j < 0 \end{cases} \tag{22}$$

Algorithm 2: PIN-Sim Methodology

Input: test dataset (D_{test}) with the target labels (*Label*), # of test samples(*S*), #of RBMs(*M*), #of nodes in hidden layer x (N_x)
Output: Error Rate
Initialize: *Err* = 0
weight.mat, *bias.mat* ⇐**Contrastive_Divergence** Algorithm
posWeight.txt, *negWeight .txt*, *posBias.txt*, *negBias.txt* ⇐ **mapWeight** (*Weight.mat*, *Bias.mat*)
for *i= 1:S* **do**
    input_data = D_{test} (i);
    for *j= 1:M* **do**
        RBM(j).sp ⇐ **mapRBM**(input_data, N_{j+1}, *posWeight.txt*, *negWeight.txt*, *posBias.txt*, *negBias.txt*);
        Run *RBM(j).sp* in HSPICE and store the obtained output voltages in array outRBM;
        for *k= 1:N_j* **do**
            Run *neuron.sp* model with *outRBM(k)* as the input of the k_{th} Neuron;
        end
        Store the output of the neurons in array *OUTPUT*;
        if *(j = M)* **then**
            if (*OUTPUT*, *Label(i)*) **then**
                Err+ = 1;
            end
        else
            input_data = OUTPUT;
        end
    end
end
ErrorRate = Err/S;

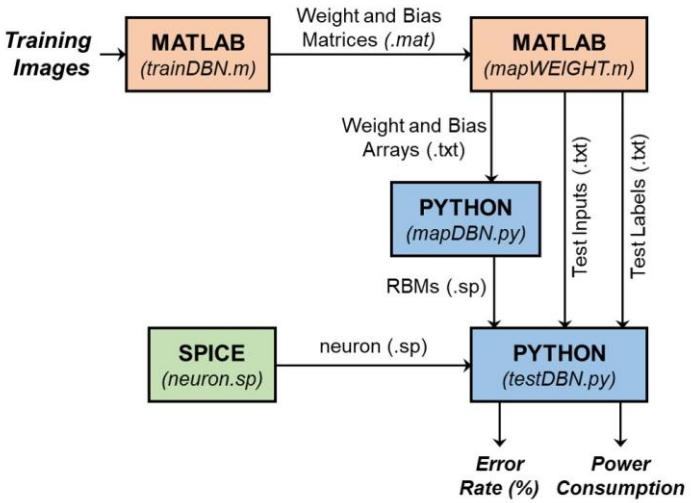


Figure 14: The block diagram of PIN-Sim framework including five main modules [12].

Next, the *mapWEIGHT* module obtains the corresponding conductance values of the elements in W^+, W^-, B^+, and B^- matrices by using the below equations:

$$\forall\, w_{(i,j)} \in (W^+, W^-): gw_{(i,j)} = \frac{(g_{max} - g_{min}) \times (w_{(i,j)} - w_{min})}{w_{max} - w_{min}} + g_{min} \tag{23}$$

$$\forall\, b_{(i,j)} \in (B^+, B^-): gb_{(i,j)} = \frac{(g_{max} - g_{min}) \times (b_{(i,j)} - b_{min})}{b_{max} - b_{min}} + g_{min} \tag{24}$$

where $\forall\, g_{(i,j)} \in G : g_{min} \leq g_{(i,j)} \leq g_{max}$, in which $g_{max} = 1/r_{min}$ and $g_{min} = 1/r_{max}$ are maximum and minimum conductances of all weighted connections in the crossbar weighted array. Moreover, w_{min}, w_{max}, b_{min}, and b_{max} are the minimum and maximum values in all of the weight and bias matrices, respectively. For implementing the required resistive crossbar array, all of the obtained conductance values are converted and quantized to their corresponding resistance values by utilizing Equation (25):

$$\forall\, g_{(i,j)} \in (GW^+, GW^-, GB^+, GB^-): r_{(i,j)} = \frac{round(Q \times 1/g_{(i,j)})}{Q} \tag{25}$$

Where GB^-, GB^+, GW^-, and GW^+ are negative bias, positive bias, negative weight, and positive weight conductance matrices, respectively, and Q is the quantization factor.

Algorithm 3: *mapWeight* Methodology

Input: *weight.mat*, *bias.mat*, #of RBMs (*M*)
Output: *posWeight(n).txt*, *negWeight(n).txt*, *posBias(n).txt*, *negBias(n).txt*, where *n* is the RBM number
Require: r_{min}, r_{max}, Quantization Factor (*Q*)
$g_{max} = 1/r_{min}$;
$g_{min} = 1/r_{max}$;
$Q = Q/(r_{max} - r_{min})$;
for *i= 1:M* **do**

W^+, W^- ⇐ *weight(i) Matrix*;
B^+, B^- ⇐ *bias(i) Matrix*;
w_{min} = smallest weight value in W_{pos}, W_{neg};
w_{max} = largest weight value in W_{pos}, W_{neg};
b_{min} = smallest weight value in B_{pos}, B_{neg};
b_{max} = largest weight value in B_{pos}, B_{neg};

$$GW^+ = \frac{(g_{max} - g_{min}) \times (W^+ - w_{min})}{w_{max} - w_{min}} + g_{min},\ RW^+ = \frac{round(Q \times 1/GW^+)}{Q};$$

$$GW^- = \frac{(g_{max} - g_{min}) \times (W^- - w_{min})}{w_{max} - w_{min}} + g_{min},\ RW^- = \frac{round(Q \times 1/GW^-)}{Q};$$

$$GB^+ = \frac{(g_{max} - g_{min}) \times (B^+ - b_{min})}{b_{max} - b_{min}} + g_{min},\ RB^+ = \frac{round(Q \times 1/GB^+)}{Q};$$

$$GB^- = \frac{(g_{max} - g_{min}) \times (B^- - b_{min})}{b_{max} - b_{min}} + g_{min},\ RB^- = \frac{round(Q \times 1/GB^-)}{Q};$$

posWeight(i).txt ⇐ RW^+;
negWeight(i).txt ⇐ RW^-;
posBias(i).txt ⇐ RB^+;
negBias(i).txt ⇐ RB^-;
end

Then, the negative and positive bias and weight resistance matrices will be converted to text files. As shown in Figure 14, a Python module called *mapRBM.py* receives the obtained matrices and based on the defined network topology, automatically produces plural crossbar weighted array circuits in SPICE. At the end, another Python module named *testDBN.py* uses the model of the probabilistic neuron and generated circuit of the DBN to perform a SPICE circuit simulation and calculate the error rate by utilizing the text files of test inputs and test labels.

A possible hardware implementation of an RBM is shown in

Figure 15 whereby the needed probabilistic sigmoidal activation function neurons is generated by the concise embedded MRAM-based design described in the prior subsection. For realizing the matrix multiplication elaborated in Equation (15), the resistive crossbar arrays are employed in this implementation. The resistive weighted connections will be programmed on the basis of the off-chip trained weights. As illustrated in

Figure 15, two resistive weighted arrays with the same dimensions are needed to implement the positive and negative weights in the *w* matrix. The differential amplifiers which are implemented by op-amps link the outputs of the positive and negative weighted connections. The input signal of the MRAM-based neuron is the output voltage of the op-amp. The embedded MRAM-based neuron will generate an output voltage signal with a probability, which is modulated based on the applied input voltage and fluctuates between V_{DD} and GND. At the end, the probabilistic output of the neuron is converted to an analog voltage level by employing a resistor-capacitor (RC) integrator circuit, which can be later converted to a digital output within an analog to digital converter. Based on the application requirements, several parameters can be tuned in the PIN-Sim framework as listed in Table 7.

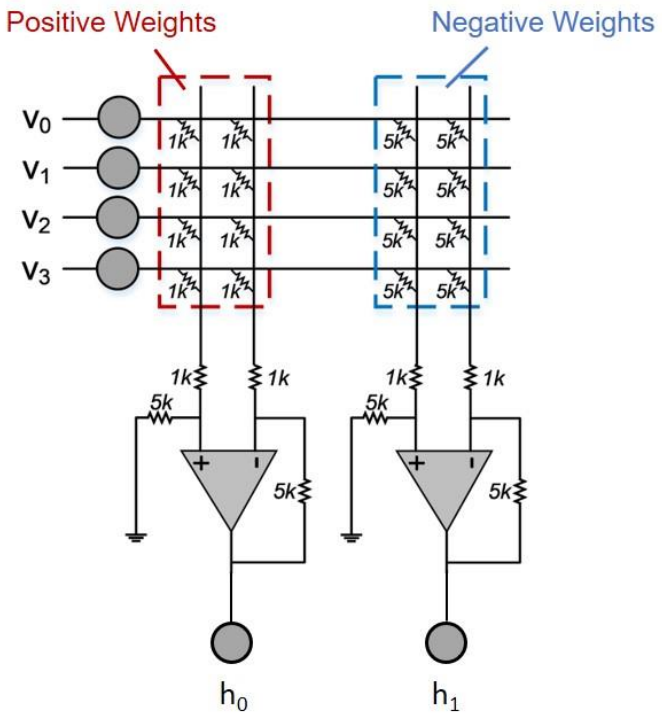


Figure 15: An RBM hardware implementation. Two resistive arrays are leveraged along with differential amplifiers to implement both positive and negative weights. The embedded MRAM-based neurons are used to evaluate the activation functions. The fluctuating output voltage of the neurons are integrated through an RC circuit to generate the output of the proposed RBM structure.

Table 7: PIN-Sim Tunable Parameters and their default values

Parameters	Description	Default Value
Topology	Defines the number of layers and nodes	784×200×10
TrainNum	# of training images	3,000
R_{min}	Minimum resistance of the weighted connections	1 kΩ
ΔRW	Difference between min and max resistances of weighted connections	400%
Q	Quantization factor	8
R_0, R_1	Resistances of the resistors in the differential amplifiers	1 kΩ, 5 kΩ
R_i, C_i	Resistance and capacitance of the RC integrator circuits	100 kΩ, 20 fF

CHAPTER 4: PROBABILISTIC INTERPOLATION RECODER

In this dissertation research, we use a 784×200×10 DBN for MNIST pattern recognition tasks. Figure 16 indicates the output voltages of the neurons for a sample digit of "4" in the last hidden layer whereas each neuron represents an output class. Figure 16 (a) shows the probabilistic outputs of the p-bit devices while the outputs of their corresponding integrator circuits is demonstrated in Figure 16 (b). The outputs of the integrators are connected to the proposed PIR circuits described in this section to interpolate the probabilistic outputs of the neurons representing each class in the MNIST dataset.

4.1 SAMPLE AND COUNT BASED PIR (SC-PIR)

Conventional methods for designing an interpolation circuit for probabilistic neurons involve using an integrator circuit, e.g. resistor-capacitor (RC) circuit, along with an analog-digital-

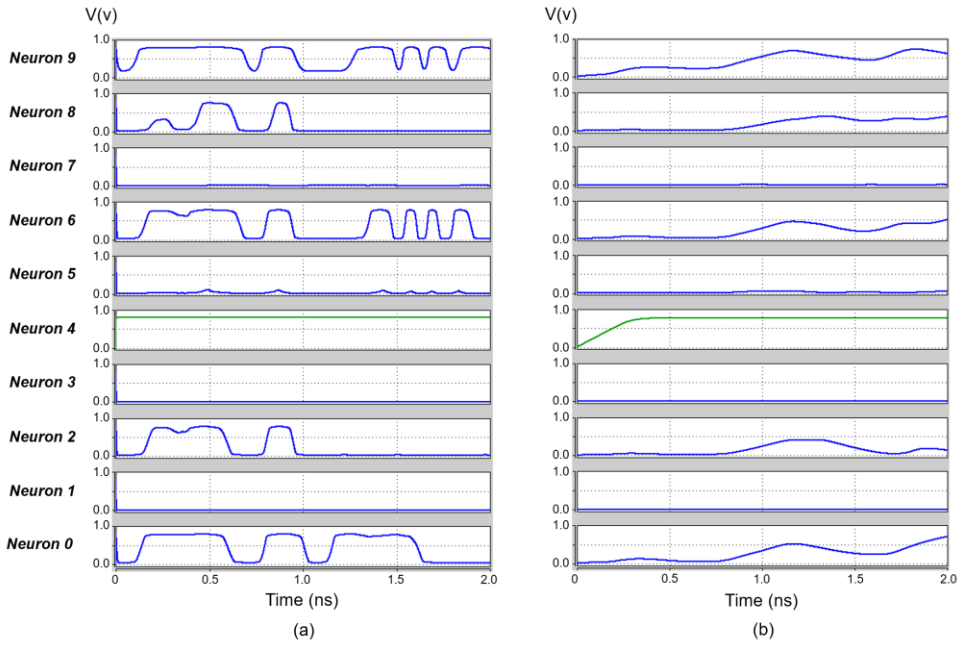


Figure 16: Output voltages of a 784 × 200 × 10 DBN for a sample digit of "4": (a) Probabilistic output of the p-bit devices, (b) Output of the integrator circuit [12].

converter (ADC) to convert the probabilistic outputs of the neurons to a digital output, as shown in Figure 17 (a). Interpolation circuits such as ADCs, which are required for a completely operational network, are being investigated as an emerging topic in computing. These are identified as useful targets to further reduce energy and area demands [193], [194], [195], [196], [197]. For instance in [194], significant reduction of the ADC energy and area overhead is achieved by using bit-slice sparsity since the power-hungry ADCs prevent the practical deployment of Resistive Random-Access Memory (ReRAM)-based DNN accelerators on end devices with limited chip area and power budget. In [195], they painstakingly attempted to reduce the overhead of ADCs by partitioning the input into several segments which are fed sequentially into the crossbar. An alternate technique is presented in [196] to reduce the overhead of ADCs in ReRAM neuromorphic computing systems by normalizing and quantizing data. In [197], it is an explicit focus to considerably decrease the overhead of the peripheral circuit to reduce the total design area and power consumption by quantizing the weights to fewer bits. Herein, we propose a CMOS-based probabilistic interpolation recoder (PIR), which is directly connected to the p-bits to generate a discrete n-bit output for each of the neurons in the last layer of the network. Figure 17 (b) shows the circuit structure of 3-bit SC-PIRs.

In the proposed SC-PIR circuits, the probabilistic output of the embedded MRAM-based neuron ($Neuron_{OUT}$) is sampled at the positive edge of each clock (*clk*), and the sampled outputs are accumulated through a counter. A ctrl signal is utilized to reset the counter and control the PIR circuit's sampling time window. An *n*-bit PIR circuit counts the sampled outputs for *2n-1* clocks and then returns the accumulated value in the form of an n-bit output (OUT_{n-1}-OUT_0). Figure 18 (a) exhibits the transient response of the proposed 3-bit SCPIR circuits, while the input of the p-bit based neuron is set to $V_{IN} = V_{DD}/2 = 400mV$. When the ctrl signal is "0", the counter is reset

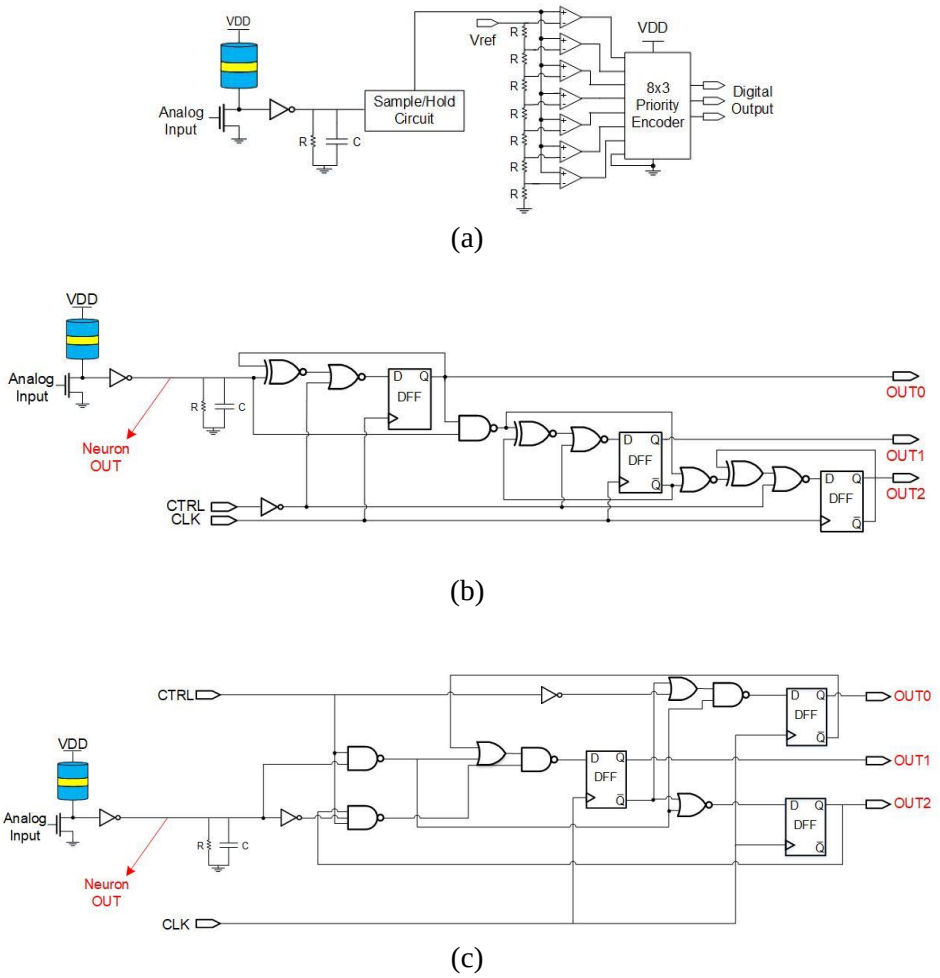
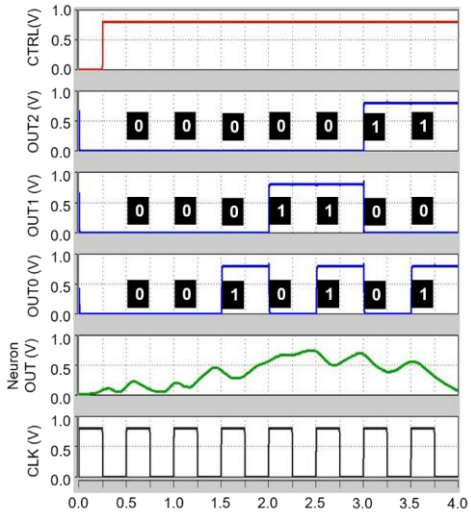
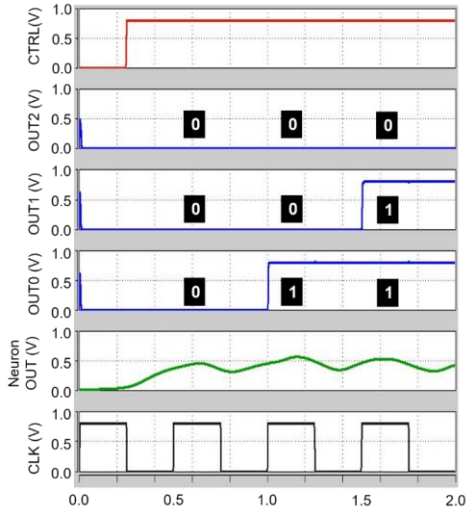


Figure 17: (a) 3-bit ADC circuit, (b) 3-bit SC-PIR circuit, and (c) 3-bit SS-PIR circuit.

and the output of the PIR circuit will be connected to GND, i.e. (OUT_2 -OUT_0 = 000). When the *ctrl* = 1, the counter is activated, and the output of the neuron is sampled at every positive edge of the clock signal. If the output of the RC integrator circuit connected to the neuron is greater than $V_{DD}/2$ during the sampling time, the PIR circuit will increment the counter, else the counter remains unchanged. For instance, in Figure 18 (a), the counter is incremented from 000 to 001 at the fourth positive edge of the clock since ctrl signal is equal to ”1” and the voltage of the $Neuron_{OUT}$ is greater than $V_{DD}/2$ = 400 mV. An n-bit SC-PIR circuit continues this process for



(a)



(b)

Figure 18: Timing waveforms of (a) 3-bit SC-PIR circuit and (b) 3-bit SS-PIR circuit.

2n clock periods and after the *2n*-th period, the output of the counter is used as the interpolated output of the probabilistic neuron.

4.2 SAMPLE AND SHIFT BASED PIR (SS-PIR)

In this dissertation research, we develop another alternative implementation of PIR circuits that is called sample and shift based PIR (SS-PIR) in the interest of improving energy consumption

while obtaining a comparable error rate. In the proposed SS-PIR circuit, the sampled outputs are interpolated through a bidirectional shift register at the positive edge of clock (*clk*). The SS-PIR circuit shifts by one position the bit array stored in it, shifting in $Neuron_{OUT}$ and shifting out the last bit in the array at each transition of the clock input. The shift register in the SS-PIR circuit must be shifted right or left if the sampled output voltage of the neurons integrator ($Neuron_{OUT}$) is less than or greater than $V_{DD}/2$, respectively. In other words, the bit array that is stored in shift register multiplies or divides by 2 if $Neuron_{OUT}$ is less than or greater than $V_{DD}/2$, respectively. A ctrl signal is utilized to reset the shift register and control the SS-PIR circuit's sampling time window. An n-bit SS-PIR circuit counts the sampled outputs for n clock periods and then returns the shifted value in the form of an n-bit output (OUT_{n-1}-OUT_0). Figure 18 (b) exhibits the transient response of the proposed 3-bit SS-PIR circuits while the input of the p-bit based neuron is set to $V_{IN} = V_{DD}/2 = 400$ mV. For instance, as shown in the figure, when the ctrl signal is "1", the value stored in the shift register changes from 000 to 001 at the third positive edge, and from 001 to 011 at the fourth positive edge of the clock since $Neuron_{OUT} > (V_{DD}/2 = 0.4V)$.

4.3 PIR FOR SPIKING NEURAL NETWORKS

With some minor changes in the PIR circuit design, they can be utilized in the Spiking Neural Network (SNN) architectures as well. There are various implementations of spiking neurons, whereby some require a compatible counting and sampling while others do not utilize such techniques. In Seo et al. [202], each SNNs neuron circuit has its own 16-bit adder, Op-amp comparator, and a 4:1 mux to integrate all presynaptic weights and determine firing activity, which essentially imposes a large power overhead to the design. In [203], the neurons readout block includes a column ADC, which contains a summing amplifier, a sample-and-hold circuit

and a high-resolution ADC, which again shows a large area-overhead. Wang et al. [204] exploit a capacitive accumulator and then a comparator as well as a flip-flop to readout the data from a SNN-based RRAM crossbar. On the other hand, in a recent work [205], the authors present an efficient three-step memristive-based SNNs neuron; or reference [206] presents an all-spin SNNs by using a domain wall-based neuron, where neither of these designs need adder/comparator-based techniques.

The proposed sequential PIR circuit can be modified to a combinational circuit which instead of sampling the output of the neuron at the positive edges of the clock, would increment the counter or shifts the shift register in SC-PIR and SS-PIR circuits, respectively. This occurs when the input voltage of the circuit (i.e. output of the neuron in SNN) is greater than a specific voltage threshold. However, the proposal of our PIR circuit is particularly important for DBNs whereas listed in Table 6, the p-bit based neurons achieve orders of magnitude energy and area reduction compared to their CMOS-based counterpart, but they require an efficient interpolation circuit to fully-leverage their advantages. Thus, in this dissertation research, we have focused on developing energy and area efficient interpolation circuits for DBN architectures.

4.4 SIMULATION RESULTS

In order to assess the performance of the proposed PIR circuits, we have utilized them within the structure of a 784×200×10 DBN circuit implemented by the PIN-Sim framework for MNIST digit recognition application. As shown in Figure 19, the PIR circuits are connected to the output layer to interpolate the probabilistic output of the neurons which represent the 0-9 digit classes of the MNIST dataset. Moreover, we employ a circular disk magnet that have been fabricated and

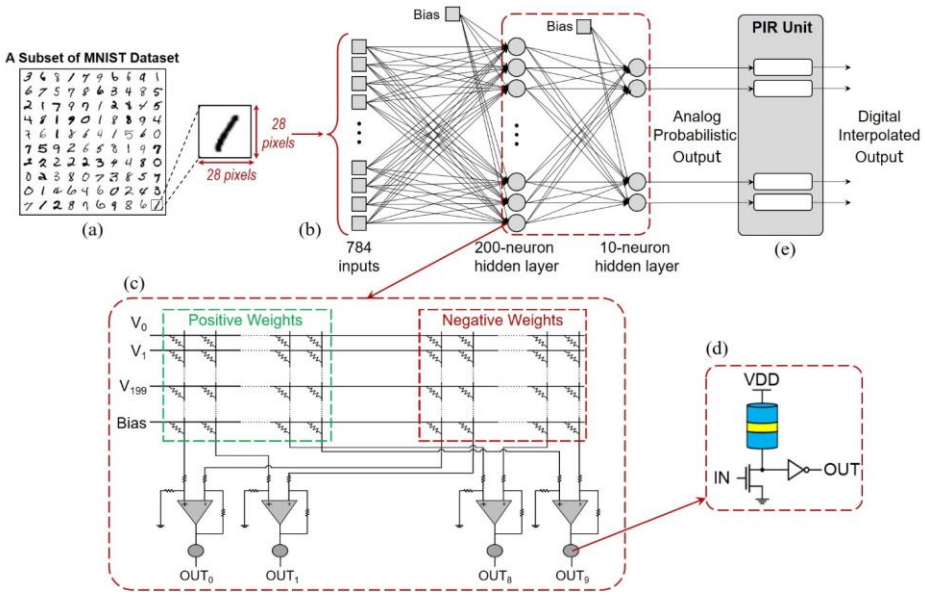


Figure 19: Simulation framework utilized for application-level simulations. (a) subset of MNIST dataset with 100 test images, (b) a 784×200×10 DBN developed for MNIST pattern recognition application, (c) hardware implementation of the 784×200×10 DBN using PIN-Sim tool, (d) stochastic MRAM-based neuron (p-bit), and (e) PIR unit used to interpolate the probabilistic output of the p-bit based output neurons to digital output.

characterized in [207], [208], and [198] with near-zero energy barrier without any shape anisotropy.

Table 8 shows the device parameters that are used in the simulations in this dissertation research [13]. It should be emphasized that the results are not considerably influenced by the current that is flowing at the midpoint ($V_{IN} = V_{DD}/2$) for the selected parameters with a circular free layer with an in-plane anisotropy, and any pinning at higher input voltages takes advantage of switching operation of the device. By verifying the functionality and efficiency of the PIR circuits for MNIST dataset, their efficiency for larger datasets will be validated as well. This is because the PIR circuits are only used to interpolate the probabilistic output of the last layer in the network, while the accuracy of the network for various datasets rely on other factors such as

Table 9: The binary outputs generated by ADC-based and PIR-based interpolation circuits for an input digit "2" from the MNIST dataset of handwritten digits.

Output Class	3-bit ADC	3-bit SC-PIR	3-bit SS-PIR	4-bit SC-PIR	4-bit SS-PIR	5-bit SC-PIR	5-bit SS-PIR
Digit-0	001	100	001	0011	0000	10001	00000
Digit-1	001	000	000	0001	0000	00111	00000
Digit-2	110	111	111	1110	1111	11110	11111
Digit-3	010	111	011	0110	0011	11111	00000
Digit-4	001	001	000	0001	0000	01000	00000
Digit-5	000	000	000	0010	0000	00010	00000
Digit-6	000	001	000	0011	0000	01011	00000
Digit-7	000	000	000	0000	0000	00000	00000
Digit-8	001	100	000	0111	0000	10101	00000
Digit-9	000	000	000	0010	0000	00010	00000
1st Selected	2	2,3	2	2	2	3	2
2nd Selected	3	0,8	3	8	3	2	-
3rd Selected	0,1,4,8	4,6	0	3	-	8	-

Temperature	26.85 °C

number of hidden layers and number of nodes in each hidden layer which is not the focus of this work. Thus, once it is shown that PIR circuits can properly interpolate the output of the network for MNIST dataset, it is also verified that they can interpolate the outputs of different DBN topologies for different datasets.

4.4.1 ACCURACY ANALYSES

Herein, 100 images from MNIST dataset are selected, which induced the most discrepancy in recognition accuracy when classified using ADC and PIR circuits. Output classes are selected according to the binary values given to them by the ADC-based or PIR-based interpolation circuits. For instance, Table 9 exhibits the binary values generated for each output classes in the 784×200×10 DBN for a sample digit "2" from the selected images of the MNIST dataset. The output class(es) with the largest binary value represents the first class(es) selected by the interpolation circuit. As listed in Table 9, the 3-bit and 5-bit SC-PIR circuits produced similar output binary values for digit classes 2,3 and 3 respectively as its top selections, which is an incorrect recognition, while other circuits successfully selected the correct output class.

Table 10 provides a recognition accuracy comparison between DBN circuits with 3-bit ADC and DBNs with 3-bit, 4-bit and 5-bit PIR in their structure. As listed, the 3-bit PIR circuits could obtain a comparable error rate with 3-bit ADC circuit, which led to a top-2 error rate of 0.23 and 0.24 for SC-PIR and SS-PIR respectively. This is mainly due to the low number of samples in the sampling time window for the 3-bit PIR circuits, i.e. only 7 and 3 samples for SC-PIR and SS-PIR respectively, which results in giving the same value to different classes. On the other hand, 4-bit SC-PIR and 5-bit SS-PIR circuits could achieve better error rate than 3-bit ADC circuit as shown in Figure 20. It is worth emphasizing that the network topology, weights, and neurons in each of these DBN implementations are similar, even a similar random seed is utilized in the SPICE simulations to generate the probabilistic behavior of the p-bit based neurons, thus the discrepancy in the recognition accuracy is only induced by the difference in the interpolation circuits and no other factors are involved.

Table 10: Various DBN hardware implementations with a focus on activation function structure.

Design		3-bit ADC	3-bit SC-PIR	3-bit SS-PIR	4-bit SC-PIR	4-bit SS-PIR	5-bit SC-PIR	5-bit SS-PIR
Resource Utilization	OP-AMP	9	-	-	-	-	-	-
	Capacitor	2	1	1	1	1	1	1
	Resistor	22	1	1	1	1	1	1
	Transistor	94	114	90	156	128	208	152
Required Number of clocks		-	8	4	16	5	32	6
Error Rate		0.2	0.23	0.24	0.17	0.27	0.18	0.18
Power Consumption (μW)		70.3	39.2	32	38.4	43.3	42.6	39.5
Energy Consumption (fJ)		351.5	156.8	64	307.2	108.25	681.6	118.5
Energy-Error -Product		702.6	360.6	153.6	522.2	292.2	1226.8	213.3

Figure 21: EEP for 3-bit, 4-bit and 5-bit SC-PIR and SS-PIR.

4.4.2 PERFORMANCE ANALYSES

In this work, the authors expected an increase in the error rate by replacing the ADCs with PIR circuits since a continuous integration operation followed by a sample-and-hold operation, and analog-to-digital conversion is replaced by a simple sample and accumulation method that is implemented only by CMOS transistors. Thus, to better comprehend the advantages of our proposed circuits, we have defined a metric called energy-error-product (EEP) as follows, which incorporates the energy costs to achieve a particular accuracy:

$$EEP = N \times E \times err \tag{26}$$

where N is the number of output neurons, E is the energy consumption of the PIR circuit, and err is the error rate of the network.

Table 10 provides a comparison between the 3-bit ADC and 3-bit, 4-bit and 5-bit PIR circuits in terms of resource utilization, power/energy consumption, and EEP values. A comparison between 3-bit ADC and PIR circuits shows a significant improvement in the effectiveness of resource utilization. In the PIR circuits, all of the area-consuming elements in the conventional circuits such as operational amplifiers (op-amps), resistors, and capacitors are removed and for

Table 11: Power and energy consumption of weighted array, activation function and interpolation circuits for several DBN topologies.

Topology	Power Consumption (mW)				Energy Consumption (pJ)			
	Weighted Array	Activation Function	Interpolation Circuits		Weighted Array	Activation Function	Interpolation Circuits	
			3-bit ADC	3-bit SS-PIR			3-bit ADC	3-bit SS-PIR
784×10	4.146	0.194	0.703	0.32	8.292	0.388	3.515	0.64
784×200×10	80.4	5.6	0.703	0.32	321.6	22.4	3.515	0.64
784×200×200×10	117.57	10.5	0.703	0.32	705.42	63	3.515	0.64

example, only 58 MOS transistors are increased for 5-bit SS-PIR compared to the 3-bit ADC circuit. Moreover, more than 54% and 81% reductions in power and energy are achieved, respectively, whereas EEP reduction is 78% for 3-bit SS-PIR circuit compared to 3-bit ADC as shown in Figure 21. The results obtained verify the advantage of our proposed circuit in terms of the individual and combined metrics of accuracy and energy consumption.

Table 11 lists the power and energy consumption of the weighted array, activation function, and interpolation circuits for several DBN topologies. In smaller networks, such as the 784×10 DBN, energy consumption of the ADC-based interpolation circuit is approximately 9-fold greater than the energy that is consumed in the activation functions, while it constitutes almost 28% of the total energy consumption of the entire network. On the other hand, the proposed SS-PIR circuit achieves more than 5-fold energy consumption reduction compared to ADC-based circuit, which significantly reduces the contribution of the interpolation circuit to the total energy consumption of the network from 28% to only 6%. By enlarging the size of the network, the activation function and interpolation circuit will be minority sources of energy consumption, which is partially realized by the considerable energy reductions achieved by utilizing the p-bit devices as neurons and proposed PIRs as interpolation circuits.

4.4.3 AREA ANALYSIS

One of the major challenges of ADC circuits are their significant area consumption, which is mainly induced by the large analog components existing in their structure such as Op-Amps. Herein, we have used a Flash ADC, which uses a linear voltage ladder with Op-Amp based comparators and an encoder circuit to interpolate the probabilistic output of the circuit and compared its energy and area consumption with our proposed PIR circuits. For the Op-Amp circuits we have used the CMOS-based design proposed in [211], which reports an area consumption of approximately 250 μm^2 for 130nm CMOS technology, scaling it down to 14nm nodes using the scaling method proposed in [212] results in an approximate area consumption of 2.9 μm^2 for each Op-Amps utilized in the ADC circuits. On the other hand, the layout design results of MRAM-based neuron demonstrate that the area consumption of the MRAM-based neuron is approximately equal to 32 $\lambda \times 32 \lambda$, where λ = 14nm/2 = 7nm for 14nm FinFET technology, thus leading to the approximate area consumption of 0.05 μm^2 per neuron [12].

Herein we have used the area consumption of the p-bit neuron as the baseline and all the other

Table 12: Area of weighted array, activation function and interpolation circuits for several DBN topologies relative to the area occupied by a single p-bit-based neuron.

Topology	Normalized Area			
	Weighted Array	Activation Function	Interpolation Circuits	
			3-bit ADC	3-bit SS-PIR
784×10	2600×	10×	4400×	330×
784×200×10	52000×	2000×	4400×	330×
784×200×200×10	66000×	400000×	4400×	330×

estimated area values are normalized according to the p-bit area consumption. For instance, the area required to implement the RC circuit with 100 K resistor and 20fF capacitor is almost three times larger than that of the p-bit [12], i.e. RC_{Area}=3X, i.e. 3×(p-bit neuron area). On the other hand, we have used the well-known 1T-1R structure for each weight in the weighted array, which allocates one transistor to each weight and the resistive devices are fabricated on top of the MOS transistors thus incurring no area overhead. Therefore, the estimated area consumption for each weight is approximately 0.02 μm^2 =0.4X. Table 12 provides the normalized area consumptions for weighted arrays, activation functions, and interpolation circuits for various network topologies. As it is listed in table, the area consumption of the activation function and interpolation circuits constitute a significantly smaller portion of the entire networks area, when the DBNs become larger which is in part realized by significant area reductions achieved by p-bit devices and PIR circuits.

Table 13: Stuck-at fault table for 4-bit SC-PIR.

Bit	Stuck-at	Output = $O_3O_2O_2O_0$															
		1111	1110	1101	1100	1011	1010	1001	1000	0111	0110	0101	0100	0011	0010	0001	0000
O_3	0	X	X	X	X	X	X	X	X								
	1									X	X	X	X	X	X	X	X
O_2	0	X	X	X	X					X	X	X	X				
	1					X	X	X	X					X	X	X	X
O_1	0	X	X			X	X			X	X			X	X		
	1			X	X			X	X			X	X			X	X
O_0	0	X		X		X		X		X		X		X		X	
	1		X		X		X		X		X		X		X		X

4.4.4 FAULT ANALYSIS

High performance integrated circuits must be protected against either transient or permanent faults. The most commonly fault model is the single stuck-at fault, in which faults are modeled in a way that only one circuit node is permanently connected to either 0 (stuck-at 0) or 1 (stuck-at 1). When a node is stuck-at 0 or 1, the value is still readable, but cannot be altered. In a write operation, the stuck-at node is faulty if the desired value is not equal to the stuck-at value but if the two values are equal, the node is not faulty. In order to do a fault simulation, it is necessary to execute two simulations: one for the fault-free circuit and another for the faulty circuit with some faults. In this way, when using the single stuck-at model, the fault injection includes a node that is permanently set either to 0 or 1. By comparing the output of the two simulations, if the simulation results are different for the same input, it is concluded that the circuit is faulty [213]. In this dissertation research, we evaluate ADC and PIR circuits in terms of reliability to achieve more efficient DBNs. For a circuit with n outputs, 2n single stuck-at faults would be possible since each output can set to 0 or 1. In 4-bit and 5-bit circuits, 8 and 10 single stuck-at faults can transpire respectively as shown in Table 13 and 14. The ’X’ shows the states that a faulty bit causes faulty output and the blank state illustrates that output is still correct despite a faulty bit. For example, the output of SS-PIR will be faulty when the desired output must be 31 just in a case that most significant bit becomes stuck at 0 ($O_4/0$). We calculate the faulty rate of each circuit by dividing the number of states that cause faulty outputs by all possible stuck-at fault states for each circuit. The faulty rate for 4-bit ADC and SC-PIR circuits is 50% because a bit flip for each output causes faulty output. In SS-PIR, the fault rate is 33% which is achieved by providing some dropouts between all possible outputs. To better comprehend the reliability

advantages of SS-PIR circuits, we have defined a metric called energy-error-faulty-product (EEFP) as follows, where *F* is the fault rate:

$$EEFP = N \times E \times err \times F \tag{27}$$

This incorporates the energy and reliability costs to achieve a particular accuracy. As shown in Figure 22, all PIR-based circuits have better EEFP than 3-bit ADC up to 84% reduction except 5-bit SC-PIR. The SS-PIR can offer better performance also in the matter of reliability in comparison to ADC and SCPIR.

Table 14: Stuck-at fault table for 5-bit SS-PIR.

Bit	Stuck-at	Output = $O_3O_2O_2O_0$					
		11111	01111	00111	00011	00001	00000
O_4	0	X					
	1		X	X	X	X	X
O_3	0		X				
	1			X	X	X	X
O_2	0			X			
	1				X	X	X
O_1	0				X		
	1					X	X
O_0	0					X	
	1						X

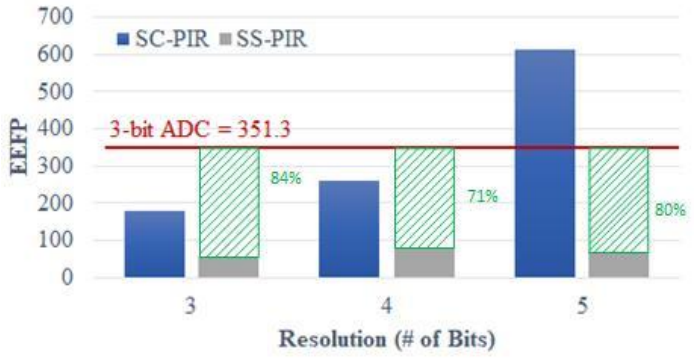


Figure 22: EEFP for 3-bit, 4-bit and 5-bit SC-PIR and SS-PIR.

4.5 RECODER BASED CONVERSION CIRCUIT

To obtain better accuracy, we develop a CMOS-based recoder which produces a digital *n*-bit output from each stochastic neuron in the last layer of the neural network. In the proposed recoder circuit, the sampled outputs are counted through a bidirectional shift register at the positive edge of each clock (*clk*). The circuit is updated based on the paradigm of two last $Neuron_{OUT}$. On the positive edge of each clock, if the most recent two $Neuron_{OUT}$ values are "01" or "10", then these states are considered as transitionary states and thus the reading remains unchanged. On the other hand, if the two last $Neuron_{OUT}$ values are "11" or "00", then these states are considered as stable states and recoder's output will be updated. In the update process, the recoder circuit shifts by one position the bit array stored in it to right or left if the $Neuron_{OUT}$ is "0" or "1", respectively. Indeed, the recoder circuit is shifting in the $Neuron_{OUT}$ and shifting out the last bit in the array at each transition of the clock input. Thus, the bit array that is stored in shift register multiplies or divides the value by 2 if the $Neuron_{OUT}$ is "1" or "0", respectively. Figure 23 shows the circuit structure of 3-bit recoder-based interpolation circuit. A *ctrl* signal is used to reset the counter and control the recoder circuit's sampling time window. An *n*-bit

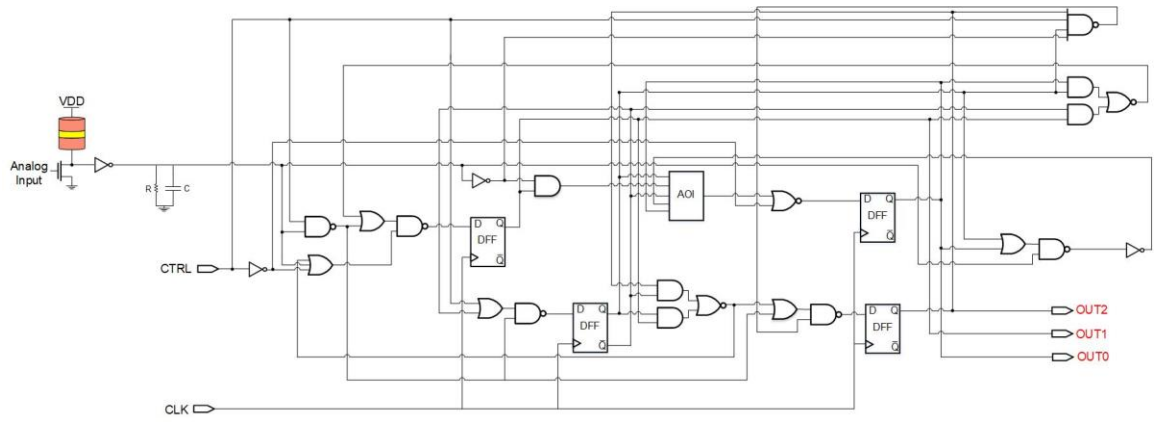


Figure 23: 3-bit recoder circuit.

Table 15: Performance comparison between 3-bit, 4-bit and 5-bit recoder circuits.

Design		3-bit ADC	3-bit Recoder	4-bit Recoder	5-bit Recoder
Resource Utilization	OP-AMP	9	-	-	-
	Capacitor	2	1	1	1
	Resistor	22	1	1	1
	Transistor	94	202	232	278
Required Number of clocks		-	5	6	7
Error Rate		0.20	0.20	0.21	0.14
Power Consumption (μW)		70.3	36.3	41.4	50.3
Energy Consumption (fJ)		351.5	90.75	124.2	176.0
Energy-Error -Product		702.6	181.5	260.8	246.4

recoder circuit counts the sampled outputs for $n - 1$ clocks and then outputs the shifted value in form of an n-bit output ($OUT_{n-1} - OUT_0$). We utilize the p-bit device model developed in [13] along with 14nm HP-FinFET PTM library to perform SPICE circuit simulations using the nominal voltage of V_{DD} = 0.8 for the purpose of verifying the functionality of the proposed recoder circuits.

Herein, the recognition accuracy comparison between DBN circuits with 3-bit ADC and DBNs with 3-bit, 4-bit, and 5- bit recoders is presented. As listed in Table 15, the 3-bit and 4-bit recoder circuits could obtain a comparable error rate with 3-bit ADC circuit, which led to a top-2 error rate of 0.20 and 0.21 respectively. Furthermore, 5-bit recoder circuit could gain better error rate than 3-bit ADC circuit. It should be noted that differences in the recognition accuracy are created only by the interpolation circuits. Other factors such as the network topology, weights, and neurons in each of these DBN implementations including random seed are identical in the SPICE simulations to produce the probabilistic behavior of the p-bit based neurons. Table 15 provides a comparison between the 3-bit ADC and 3-bit, 4-bit and 5-bit recoder circuits with

regards to resource utilization, power/energy consumption, and EEP values. A comparison between 3-bit ADC and recoder circuits shows a substantial enhancement in the efficiency of resource utilization. In the recoder circuits, all of the area-consuming elements in the conventional circuits like operational amplifiers (op-amps), resistors, and capacitors are omitted and for instance, only 108 MOS transistors are increased for 3-bit recoder in relation to the 3-bit ADC circuit. As shown in Table 15, more than 48% and 74% reductions in power and energy are achieved, respectively, while EEP reduction is 74% for 3-bit recoder circuit in relation to 3-bit ADC.

4.6 PYTHON-DRIVEN SIMULATION FRAMEWORK

The goal with this script is to gather data on how process variation will affect the energy barrier of the MTJ, thus changing the realization of the sigmoid function and potentially adverse effects to energy consumption. The Python script invokes SPICE to gather outputs under consecutive voltage data points applied to the p-bit device as shown in Figure 24. Given a SPICE neuron file with the small magnetic anisotropy field, H_K, defined in the parameter file, the script changes H_K values based on the propagated energy barrier value. It then runs the simulation while piping the bash output to a text file in case of any SPICE errors. Finally, it extracts the neuron output

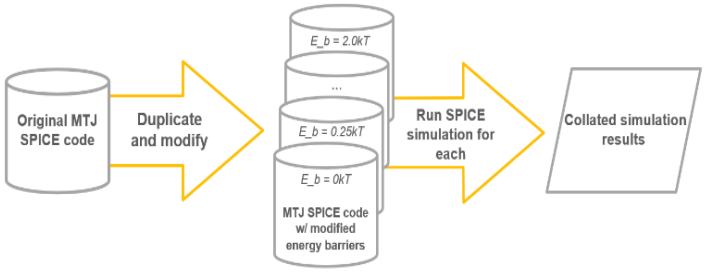


Figure 24: MTJ Energy barrier simulation using Python scripting.

voltage data points and collates them within the results text file. Multiple energy barrier values can be run sequentially if a text file containing a list of energy barriers with each entry is passed as an argument to the script.

Data points can then be plotted in MATLAB to view the effect that the energy barrier has on the realization of the sigmoid activation function. The formula used by the Python script for calculating the energy barrier, E_B, is:

$$E_B = \frac{1}{2} H_K M_s V \tag{28}$$

where V is the volume and M_S is the magnetization saturation of the MTJ. The pseudocode for the script is identified in Algorithm 4.

Algorithm 4: Effect of Process Variation on the MTJ Sigmoid Activation Function Realization

Input: energy barriers, neuron source file
Output: neuron output voltage
for each energy barrier, E:
  calculate $H_K = \frac{2E_b}{M_s V}$
  search for "H_K=" in SPICE code
  replace anisotropy field value with calculated value
  run SPICE simulation piping bash output to text file
  search for voltage output and **write** to results file

Another tool developed was a Python script to aid testing of the DBN's accuracy with the MNIST dataset. The development of these scripts extends PIN-sim and will aid future development and testing of p-bit based, DBN networks with a PIR digitization output stage.

4.7 MNIST DATASET EVALUATION

To analyze the performance of the PIR circuit, a Python script was developed to compare the large amounts of data commonly found in machine learning datasets. This accuracy analysis

script operated as follows: first, it reads one line of the MNIST dataset file to find the expected output for that testcase and the testcase label. Secondly, it locates each instance of the testcase label in the PIR output file. Third, it reads the neuron data into a list until it encounters the subsequent case. Upon finding the next testcase, all the neurons for the current testcase have been read and now can be processed. The list is sorted by constituent probabilities from high to low. Next, the first two neurons are examined to see if either of them is the neuron indicating the expected output from the MNIST dataset. If any output digit neuron subsequent to the top two neurons have the same probability, then the output of the PIR circuit counts as a *fail* even if the expected output was within the top two confidence selections, whereas the circuit was not able to tell a clear difference between which neuron was correct. If the expected output was a neuron in the top two likelihood categories, then the testcase is regarded as a *pass*.

The process is repeated until either file reaches its end. Upon finishing all testcases, then the total number of testcase passes and failures are tabulated so that the overall error rate is determined. The corresponding pseudocode is listed in Algorithm 5. The script was tested by comparing the error rates generated against previous works [214]. Namely, the script was fed outputs of a PIR circuit with 100 testcases and the MNIST dataset. The results obtained are listed in Table 15.

Algorithm 5: MNIST DBN Performance Analysis

Input: MNIST dataset, PIR output
Output: number of testcases that passed/failed
for each testcase
  for each neuron
    append neuron data to list
  sort list by neuron probability high to low
  if the expected output was in the top two neurons
    AND its probability doesn't match any neurons
    beyond the top two neurons
  then
    testcase **passes**
  else
    testcase **fails**

CHAPTER 5: ELECTRICALLY-TUNABLE STOCHASTICITY FOR SPIN-BASED NEUROMORPHIC CIRCUITS

Stochastic circuits play a significant role in the implementation of networks with probabilistic nodes. For instance, learning networks employing p-bits are worthwhile in realizing DBNs in a way that weights are trained offline by a learning algorithm in software and the hardware is utilized to repeatedly perform inference tasks effectively. Unstable low barrier nanomagnets present a direct mechanism to realize stochastic sigmoidal neurons in DBNs through leveraging the randomly fluctuating magnetization to produce a stochastic time varying output voltage. If these nanomagnets are designed to have as low energy barriers that are feasible, then many random outputs are produced in a short period of time. Under this strategy, a near-zero energy barrier ($E_b \ll k_B T$) nanomagnet has the capability of free magnetization layer flipping back and forth which can be tuned by modulated the voltage on the gate of p-bit's NMOS transistor.

5.1 EFFECTS OF PROCESS VARIATION ON THE PROBABILISTIC BEHAVIOR OF P-BIT

The p-bit device is not entirely tolerant of defects and device-to-device variations even though is more error resilient than strictly digital computing devices [215]. The statistical distribution of the magnetization fluctuations, such as the power spectral density become affected by the presence of both localized and delocalized structural defects and moderate variations for the barrier height of the nanomagnet which is caused by small size variations [216]. It is investigated that the power spectral density is relatively insensitive to the presence of small localized defects and moderate barrier height change. Nevertheless, the power spectral density is substantially affected by delocalized defects such as thickness variations over a significant fraction of the nanomagnet [217]-[219]. Delocalized defects can considerably change the fluctuation rate of the

magnetization in low barrier nanomagnets. This will affect applications in p-bit-based neurons for neuromorphic architectures because the fluctuation rate is essential for stochastic computing applications.

The near-zero energy barrier in p-bit devices is achievable by reducing the total magnetic moment through decreasing volume (V) and/or manage a small anisotropy field (H_K) [220], according to the below relation:

$$E_B = \frac{1}{2} H_K M_S V = \frac{1}{2} H_K M_s (\pi (d/2)^2 t_f) \tag{29}$$

where d and t_f are the diameter and thickness of the MTJ's free layer. Due to the variations in the fabrication process of low energy barrier nanomagnets, p-bits may exhibit different "as-built" energy barriers [221]. Based on Equation (29), variations in MTJ's anisoptropy field (σH_K) and nanomagnet diameter (σd) cause linear and quadratic variations in energy barrier (σE_B), respectively.

As described in previous section, the near-zero energy barrier free layer will fluctuate arbitrarily between the parallel and anti-parallel magnetic states. The magnetization dwell time in the parallel and anti-parallel states creates a distribution which confirms that the nanomagnet fluctuates stochastically. By switching the magnetization direction of the free layer between parallel and anti-parallel states, a sigmoidal distribution is observed over a sequence of samples. These state transitions are instigated by thermal energy which is adequate to randomly fluctuate when using a sufficiently small energy barrier. The fluctuation speed of a nanomagnet can be obtained from the average dwell time in parallel and anti-parallel states τ_P and τ_{AP} as follows [222]-[224]:

$$\tau^{-1} = \tau_P^{-1} + \tau_{AP}^{-1} \tag{30}$$

and, this time scale is related to the energy barrier (E_B) of the nanomagnet [225]

$$\tau = \tau_0 \times \exp(E_B/K_BT) \tag{31}$$

Thus, the fluctuation speed of nanomagnet can be increased or decreased by reducing or increasing the energy barrier, respectively, which will impact the probabilistic behavior of the p-bit devices.

The defects caused by the fabrication imperfections are required to be addressed for neuromorphic applications using p-bit based neurons such as DBNs due to their significant impact on their performance and accuracy. Generally, these challenges raised by variations can be addressed by two approaches. Firstly, a fabrication-oriented approach aims to refine materials and production processes. Alternatively, a post-fabrication mechanism is proposed herein which leverages temporal redundancy as well as a circuit-level mechanism to address the aforementioned PV-imposed challenges. Moreover, a sensitivity-analysis will be conducted to inform the production process with the acceptable range and tolerances for critical parameters impacting the energy-barrier and resulting stochasticity of the p-bit device.

5.2 VARIATION-LESS P-BIT BASED DBN AS THE BASELINE

In this dissertation research, we utilize a variation-less 784 × 200 × 10 DBN circuit as the baseline to analyze the reliability and energy consumption tradeoffs. The Probabilistic Inference Network-Simulator (PIN-Sim) is used to realize a circuit-level implementation of DBNs. In PIN-Sim, resistive crossbars and embedded MRAM-based p-bit neurons are employed as weighted connections and activation functions, respectively.

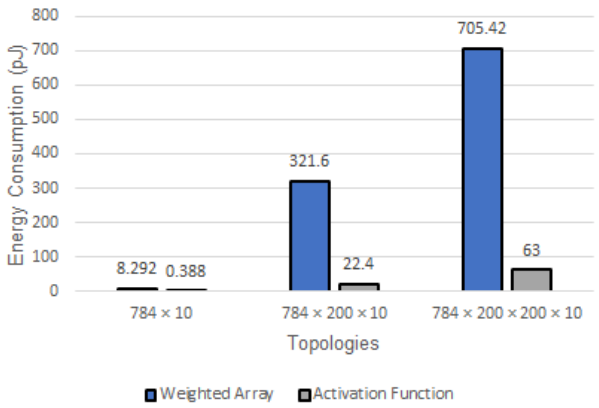


Figure 25: Energy consumption of weighted array and activation function for several DBN topologies.

The energy consumption of the weighted array and activation function for various DBN topologies is shown in Figure 25. As shown, a substantial amount of energy is consumed in the weighted connections, while less than 10% of the total energy is consumed in the neurons of an embedded MRAM-based p-bit approach. For example, the total energy consumption of a 784 × 200 × 10 DBN is almost equal to 344 pJ, only 22.4 pJ of which is dissipated in the activation functions. This is achieved by using the proposed energy-efficient embedded MRAM-based p-bit neurons to implement the activation functions, as opposed to more elaborate floating-point circuits and pseudo-random number generators.

5.3 PROPOSED VARIATION-IMMUNE P-BIT IMPLEMENTATION

Herein, the impact of energy barrier variation is assessed by using a random distribution of parameters for several ranges from near-zero kT to 2.0 kT. The higher energy barrier of 1.5 kT, 1.75 kT, and 2.0 kT are realized by increasing the small anisotropy field (H_K). As expressed in Equations (30) and (31), increasing the energy barrier decreases the probabilistic fluctuation speed of the nanomagnet in p-bit devices, which means if we do not change the sampling time of the p-bit's output the probabilistic sigmoidal activation function will be distorted as shown in Figure 27: (a) to (c). The results obtained by MATLAB simulation, depicted in Figure 26, show

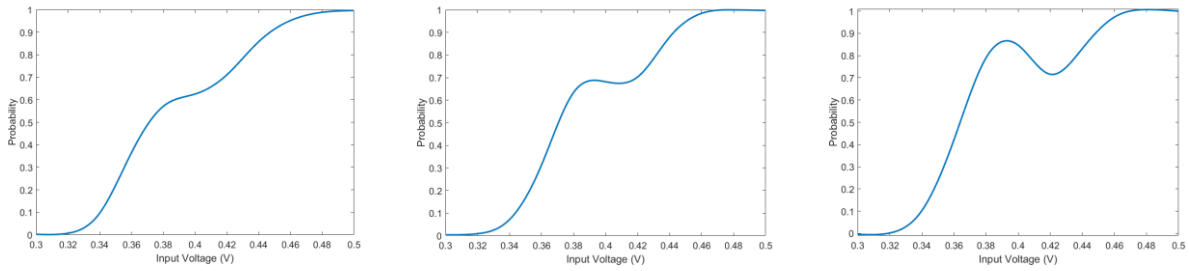


Figure 27: (a) Output probability of MRAM-based neuron for (a) E_B = 1.5 *kT*, (b) E_B = 1.75 *kT*, and (c) E_B = 2.0 *kT*.

that while the energy barriers less than or equal to 1.5 *kT* yield an recognition error of approximately 5% (i.e., accuracy rate ~95%) for MNIST hand-written digit recognition application, the error rate will be drastically increased to an unacceptable value of ~90% (i.e. accuracy rate ~10%) for the energy barriers more than 1.75 *kT* on a 784×200×10 DBN which is trained by 60,000 training images. Thus, the process variation sensitivity of DBNs utilizing low energy barrier MTJs are seen to encounter a sharp "knee effect" drop-off when energy barriers exceed 1.75 *kT* as illustrated in Figure 26.

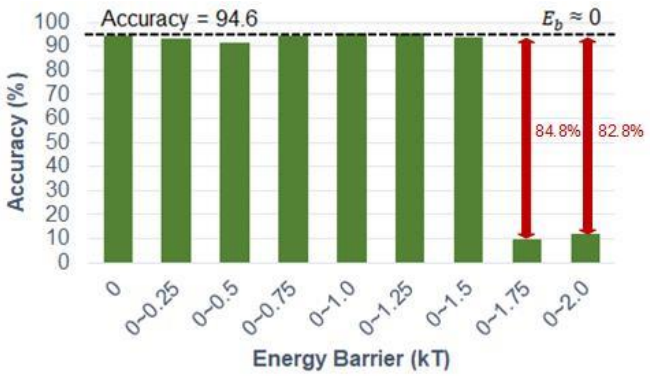


Figure 26: Effects of neuron's energy barriers on the DBN accuracy.

5.4 P-BIT WITH TEMPORAL REDUNDANCY

As mentioned, the increase in the energy barrier of p-bit results in a decrease in the probabilistic fluctuation speed of its nanomagnet. It means that given sufficiently long sampling time, the p-bit's output voltage can realize its probabilistic sigmoidal behavior without any distortions. To verify the effect of increasing the sampling window period (τ_S) of p-bit's output to address the energy-barrier variation issues, we have examined p-bits with four different energy barriers: 0.5 *kT*, 1 *kT*, 1.5 *kT*, and 2 *kT*. Figure 28: shows an experiment conducted in SPICE circuit simulator, in which the input voltage of the p-bit neurons with different energy barriers is incrementally increased from 0.3 *V* to 0.5 *V* (i.e. the active region of the p-bits probabilistic sigmoidal activation function) with 20 *mV* steps. In every step the input voltage remains fixed for τ_S period of time and the output voltage is monitored using CosmosScope. It is shown that in order to achieve the sigmoidal output required to be realized by p-bit based neurons with 0.5 *kT*, 1 *kT*, 1.5 *kT*, and 2 *kT*, the minimum τ_S should be tuned to 4 *ns*, 11 *ns*, 16 *ns*, and 19 *ns*, respectively, while the sampling window period for a p-bit with near-zero energy barrier is 2 *ns*. These results are obtained using the SWEEP function provided by HSPICE circuit simulator.

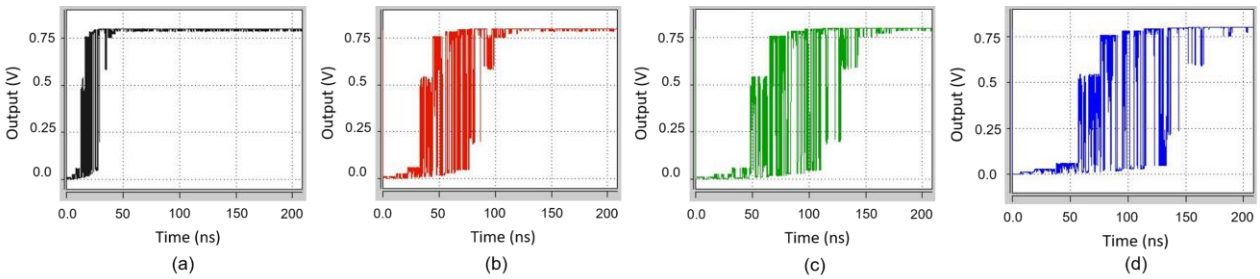


Figure 28: Output of MRAM-based neuron vs. time for different energy barriers (a) E_B = 0.5 *kT*, (b) E_B = 1.0 *kT*, (c) E_B = 1.5 *kT*, and (c) E_B = 2.0 *kT*.

In order to assess differences in energy consumption of DBNs under different energy barriers, we examined a 784×200×10 DBN circuit implemented by the PIN-Sim framework for MNIST digit recognition application using p-bits models with the maximum energy-barrier variations ranging from ~0 *kT* to 2.0 *kT* with 0.5 *kT* steps. Herein, we consider the energy consumption of the p-bit neuron with near-zero energy barrier as the baseline. Figure 29 illustrates the energy consumptions of a 784×200×10 DBN with various levels of maximum energy barrier variation tolerance using the proposed temporal redundancy mechanism. As depicted, the energy that is consumed in DBN with ~2 *kT* energy barrier variation tolerance is approximately 10-fold greater than variation-less DBNs utilizing p-bits with near-zero energy barrier. It is worth noting that variations are applied via PIN-Sim tool by using a randomly generated energy barrier value between 0 *kT* and a maximum energy barrier variation defined by user. Hereby, we seek to examine the "knee effect" point of energy barrier for p-bit devices with different energy barriers, whereby too high of a barrier increases the energy consumption of the neural network. Thus, in terms of energy consumption, a "knee effect" point for the energy barrier is seen to be around 0.5 *kT* for our DBN. This knee effect factor can be alleviated in practice by configuring a feedback mechanism to increase the fluctuation rate of the nanomagnet, as described in the following

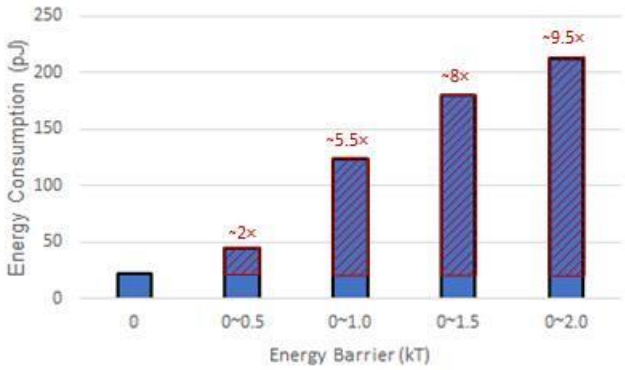


Figure 29: Influence of increasing energy barrier on energy consumption for 784×200×10 topology.

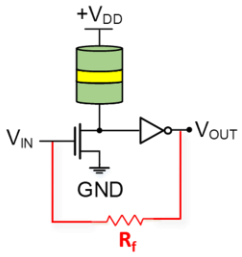


Figure 30: Device configuration with feedback for the embedded MRAM-based p-bit.

Section.

5.5 P-BIT WITH FEEDBACK

In this Section, we demonstrate a p-bit neuron circuit, in which the fluctuation rate of its nanomagnet can be tuned using an electrical feedback. In this neuron, the average fluctuation frequency (f_0) is determined by the energy barrier of the nanomagnet through the following equation [225]:

$$f_0 = (\tau_0 \times \exp(E_B/K_B T))^{-1} \quad (32)$$

Herein, the output of the p-bit device is amplified and fed back to the NMOS transistor, thus the magnetization fluctuation becomes faster, depending on the polarity and strength of the feedback, as a modulation method to compensate towards optimal levels of thermal noise. An implementation of the feedback configuration is illustrated in Figure 30. In this case, the drain of the NMOS transistor tracks the magnetization direction of the free layer of the MTJ. The inverter at the output of the device naturally generates the inverse voltage, hence realizing a feedback compensation mechanism. The feedback can be controlled by changing the value of the resistor R_f, which changes the feedback current flowing through the NMOS transistor. Figure 31 (a) shows the output of p-bit with an energy barrier of $E_B \approx 1.5\ kT$ and a feedback with $R_f = 100\ K\Omega$, while in the no feedback case (R_f = infinity), the nanomagnet of device fluctuates extremely

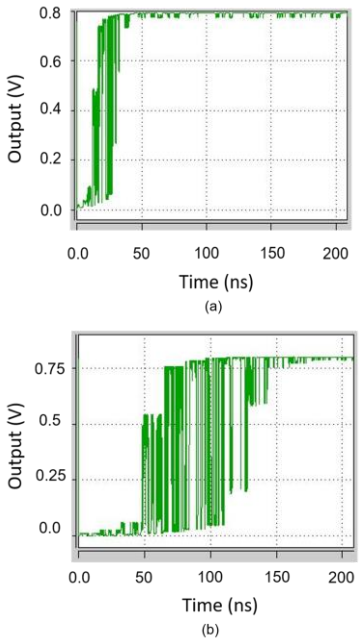


Figure 31: Tuning the effective energy barrier through electrical feedback. (a) Measurement of the output fluctuations of the device without feedback for $E_B = 1.5\ kT$. (b) Measurement of the output fluctuations of the device with the feedback implemented through a simple resistor of value 100 KΩ for $E_B = 1.5\ kT$.

slower as shown in Figure 31 (b). Employing the feedback resistor decreases the fluctuation time scale, τ, by approximately 5 times, which reduces the need for temporal redundancy, in consequent of which the energy consumption in the variation-tolerance p-bit neuron will be decreased.

A similar experiment involving feedback of the p-bit output to its input was performed in a current controlled device scheme [226]. In that case, the effect of feedback on frequency tunability can be understood by considering the change to the energy landscape of the nanomagnet. In the feedback configuration, when magnetization is in the "*P*" state, the device output feeds back a negative current to its input, thus tilting the energy barrier in favor of the "*AP*" state, i.e, the barrier that needs to be overcome to transition from the "*P*" to the "*AP*" state becomes smaller than the barrier for the reverse transition. Similarly, when the magnetization is

in the "AP" state, the barrier for transitioning from the "*AP*" to the "*P*" state is smaller than the barrier for the reverse transition. So, the energy landscape is dynamically modified in a way such that the energy barrier appears to be lower to transition from the occupied state to the other state. This effect increases the fluctuation frequency of the device output, expressed as:

$$f_0 = (\tau_0 \times \exp(E_{B,eff}/K_B T))^{-1} \tag{33}$$

where the effective energy barrier ($E_{B,eff}$) is given by:

$$E_{B,eff} = E_B(1 \pm I_{feedback}/I_C) \tag{34}$$

where E_B is the intrinsic energy barrier of the nanomagnet given in Equation (29), $I_{feedback}$ is the feedback current and I_C is the critical current for magnetization switching at zero temperature. $I_{feedback}$ can be replaced by $V_{DD}R_f$ from analyzing the circuit configuration, considering that the NMOS transistor resistance is much smaller than R_f (which can be realized by choosing a large enough R_f). Next, by defining $V_{DD}I_C$ as R_0, we get the following expression for the effective energy barrier of the magnet:

$$E_{B,eff} = E_B(1 \pm R_0/R_f) \tag{35}$$

Equations (33) and (35) elaborate that the fluctuation frequency of the p-bit can be controlled by changing the feedback resistor, as also demonstrated in the experiment [226]. The above analysis generally holds true for the device presented in this dissertation research. The circuit simulation results exhibit that maximum variations of 0.5 kT, 1 kT, 1.5 kT and 2 kT can be compensated using R_f with $30\ K\Omega$, $60\ K\Omega$, $100\ K\Omega$, $120\ K\Omega$ resistances, respectively. This is realized with only ~12% energy overhead, which is 25.1 pJ for p-bit with $120\ K\Omega$ feedback resistor compared to 22.4 pJ for p-bit without feedback.

5.6 PROCESS VARIATION ANALYSIS OF SOT PERPENDICULAR NANOMAGNETS IN DBNS

The probabilistic spin logic device (p-bit) with perpendicular magnetic anisotropy (PMA) is amongst the new building block which is completely tunable by spin orbit torque (SOT) [222],[224],[226]. The output of p-bit can be varied by adjusting a DC current through the giant spin Hall effect (GSHE) Ta Hall bar as illustrated in Figure 32. By adjusting the direction of the DC current (I_C), then the magnetization direction will probabilistically favor either the "UP" or "DOWN" state that produces a sigmoidal curve by taking the average of the states as the current is swept across a range of values. The charge current flowing through the layer with giant spin Hall effect (GSHE) modifies the dwell time in the two stable states and as a result, changes the output significantly for a LBNM with a thermal barrier close to zero *kT*. According to the thermal energy, the p-bit is implemented with a thermally-stable nanomagnet in regular MRAM cells with a high energy barrier. As a result, the p-bit offers a thresholding behavior appropriate

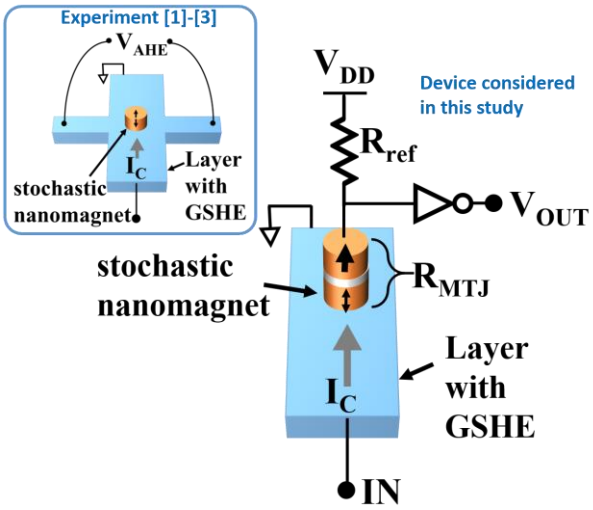


Figure 32: The diagram of the probabilistic device (p-bit) with perpendicular magnetic anisotropy (PMA) as a binary stochastic neuron for DBNs [222],[224],[226]. The experiments in [222],[224],[226] used AHE to read the magnetization state. This read scheme can be replaced by an MTJ. The magnetization state of the weak perpendicular anisotropy free layer can be read through the resistance change of an MTJ as proposed in [194].

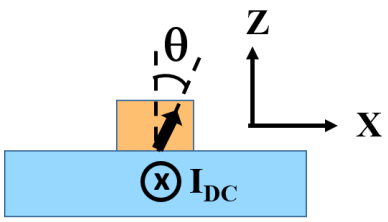


Figure 33: Tunability of the average magnetization component in the Z-direction while the magnetization lies in the ZX-plane.

for neural network applications and meanwhile works in non-volatile storages in the form of a single-bit element. The magnetization direction in the device is read through anomalous Hall effect (AHE), which requires large CMOS circuitry to amplify the weak output signal generated. However, if the single nanomagnet of the device can be replaced by an MTJ where the free layer is designed to have a similarly weak perpendicular anisotropy, then the magnetization fluctuations can be read through a much stronger tunneling magnetoresistance (TMR) effect. A similar device is proposed in [194] and shown in Figure 32. This device uses the same technology as the SOT-MRAM, with one modification, i.e. the MTJ free layer is made thermally unstable. Hence, the implementation of p-bits requires small changes to the MRAM fabrication flow. In a low perpendicular anisotropy p-bit device, any little in-plane anisotropy can result in a considerable tilt angle (θ) that may not be detectable in high perpendicular anisotropy magnets. The in-plane spins do not impact the desired direction of the average magnetization component in the Z-direction when there is no tilt in the magnet's anisotropy. As a result, the average magnetization component in the Z-direction remains around zero. Accordingly, the tilt direction is toward the X-axis in the ZX-plane as shown in Figure 33. It has been proved, in the Z-direction, perfect tunability of the average magnetization component can be attained for tilt angles around 25 degrees [222],[224],[226].

A three-terminal SOT-based neuron is much more suitable to be utilized into neural networks for several reasons. The shared read/write path across the whole device in STT-based neurons results in a read reliability issue while serious stress can be applied on the MTJ by the write current. Also, the shared read and write path causes the concatenation of these devices into a neural network to become complicated since the input and output signals are not isolated from each other. Moreover, due to considerable incubation delays of STT-MRAM devices, they are unable to work reliably at ns and sub-ns scales [227]-[230].

In this section, the impact of process variation on the SOT p-bit based DBN is evaluated. We have modeled a random variation distribution of three types of process variation which affects the fluctuation speed of nanomagnet; 1) (σH_K): variations in the anisotropy field (H_K), 2) (σd): variations in the diameter (d) of nanomagnet, and 3) (σt_f): variations in the thickness (t_f) of nanomagnet, for various temperatures and tilt angles (θ). The nanomagnet parameters used in our simulation for a variation-less p-bit based DBN as the baseline are: H_K = 400 mT, D = 36 nm, t_f = 1.3 nm, θ = 25 *degrees*, and Temperature = 300K. The lower and higher energy barriers are realized by decreasing and increasing these three parameters. As described in Equations (30) and (31), reducing and increasing the energy barrier increases and reduces the nanomagnet's probabilistic fluctuation speed in SOT-MRAM devices.

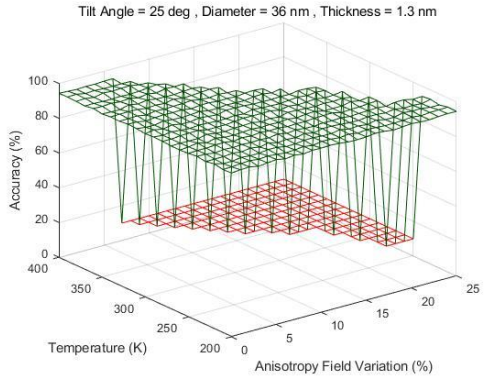
5.6.1 INDIVIDUAL VARIATION

Herein, we analyze the impact of individual parameter variation while temperature and tilt angle are varying for specified ranges. The results are achieved by MATLAB simulation for MNIST hand-written digit recognition application by utilizing 60,000 training images on a 784×200×10

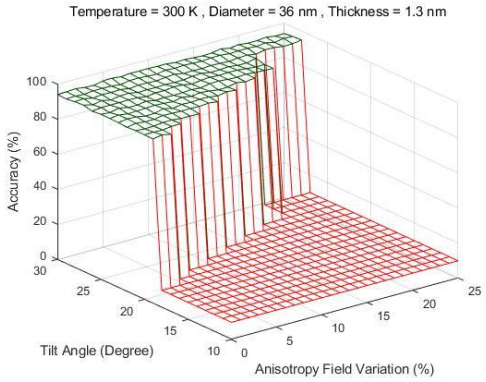
DBN. It is mentioned that variations in PIN-Sim framework are applied by utilizing a randomly generated parameters value between the baseline value of each parameter and a maximum parameter variation of 25%.

5.6.1.1 ANISOTROPY FIELD VARIATION

Figure 34 (a) and (b) show the accuracy of p-bit based DBN versus σH_{K} for various temperatures of 200K to 400K and tilt angles of 10 degrees to 30 degrees, respectively. Other nanomagnet



(a)



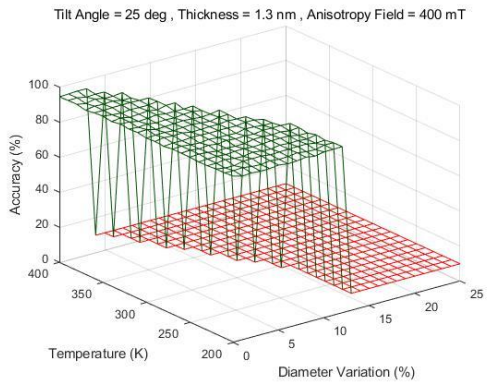
(b)

Figure 34: Accuracy of p-bit based DBN versus σH_M for: (a) Temperature of 200K to 400K, (b) Tilt angles of 10 degrees to 30 degrees.

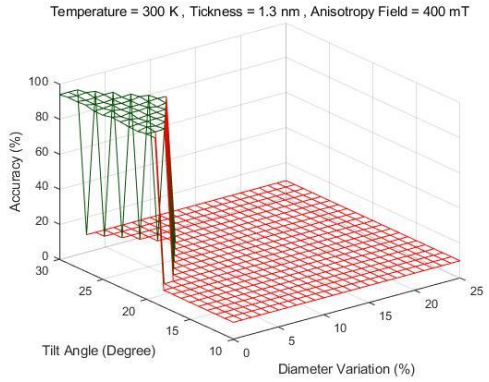
parameters are the same with the baseline and fixed. As shown in Figure 34 (a), anisotropy field variations do not affect the accuracy of p-bit based DBN while temperature values range from 200K to 250K. The worst-case scenario is when temperature = 400K in presence of around 7% process variation in anisotropy field which the accuracy will be reduced to an unsuitable value of around 10% (i.e. error rate around 90%). As it can be seen in the Figure 34 (b), tilt angle should be at least around 18 degrees while process variation in anisotropy field completely can be tolerated up to 25% (i.e. σH_K = 25%) for at least a tilt angle of 28 degrees. Thus, the process variation of anisotropy field in p-bit based DBNs are seen to be tolerated up to around 20% for the baseline values of temperature and tilt angle. In this case, for temperature of 300K and tilt angle 25 degrees baseline values are used throughout.

5.6.1.2 DIAMETER VARIATION

The accuracy of p-bit based DBN versus σd for various temperatures of 200K to 400K and tilt angles of 10 degrees to 30 degrees is demonstrated in Figure 35 (a) and (b), respectively, whereas other nanomagnet parameters are equivalent to the baseline values and fixed. By decreasing temperature, higher diameter variation can be tolerated as shown in Figure 35 (a). The best-case scenario is when temperature = 250K in presence of around 14% process variation in diameter which the accuracy of around 90% is still obtained. As can be observed in Figure 35 (b), diameter variation can be tolerated up to around 8%, i.e. σd = 8%, for a small range of tilt angles. Hence, the process variation of diameter in p-bit based DBNs is witnessed to be tolerated up to around 8% for the baseline values of temperature and tilt angle. In comparison to anisotropy field variation, p-bit based DBNs are more sensitive to diameter variation.



(a)

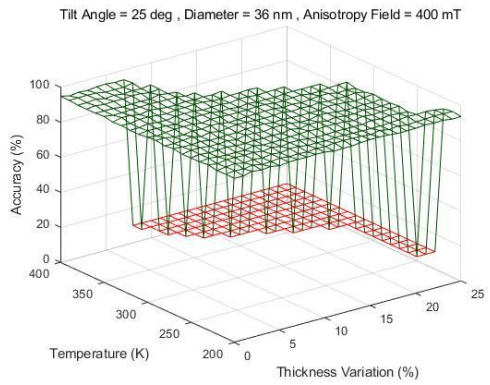


(b)

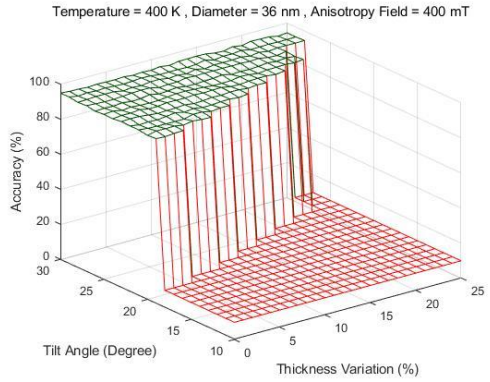
Figure 35: Accuracy of p-bit based DBN versus σd for: (a) Temperature of 200K to 400K, (b) Tilt angles of 10 degrees to 30 degrees.

5.6.1.3 THICKNESS VARIATION

Figure 36 (a) and (b) exhibit the accuracy of p-bit based DBN versus σt_f while temperature and tilt angle range from 200K to 400K and 10 degrees to 30 degrees, respectively. Other nanomagnet parameters are fixed values which are equivalent to the baseline values. Thickness variations do not impact the accuracy of p-bit based DBN while temperature values range from 200K to 220K as displayed in Figure 36 (a). The least thickness variation tolerance is achieved



(a)



(b)

Figure 36: Accuracy of p-bit based DBN versus σt_f for: (a) Temperature of 200*K* to 400*K*, (b) Tilt angles of 10 degrees to 30 degrees.

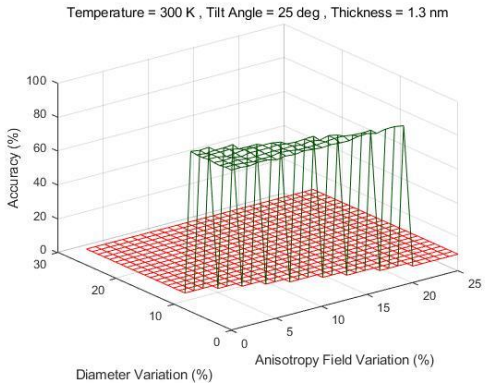
when temperature = 400K in presence of around 8% process variation in thickness which the accuracy will be reduced to an inappropriate value of around 10% (i.e. error rate around 90%). As illustrated in Figure 36 (b), tilt angle should be at least around 18 degrees while process variation in thickness completely can be tolerated up to 25% (i.e. σt_f = 25%) for at least a tilt angle of 28 degrees. Therefore, the process variation of thickness in p-bit based DBNs is seen to be tolerated up to around 23% percent for the baseline values of temperature and tilt angle. The

p-bit based DBNs are less sensitive to thickness variation in relation to anisotropy field and diameter variations.

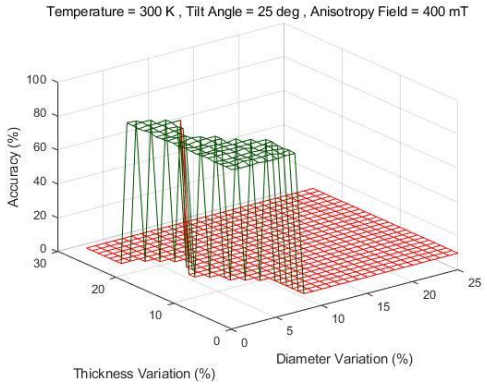
5.6.2 IMPACT OF MULTIPLE SOURCES OF VARIATION

Herein, we analyze the impact of multiple variation sources of σH_K, σd, and σt_f ranging from 0% to 25% on the p-bit based DBNs while temperature and tilt angle are fixed values which are equivalent to the baseline values of 300K and 25 degrees, respectively. The results are achieved by MATLAB simulation for MNIST hand-written digit recognition application by utilizing 60,000 training images on a 784×200×10 DBN. As previously mentioned, the maximum value of variation of all parameters is limited to 25%.

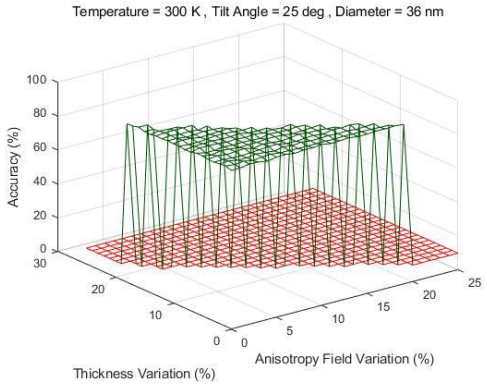
Figure 37 (a) to (c) show the accuracy of p-bit based DBN for three different combinations of σH_K, σd, and σt_f. As we can see, by increasing variation in a parameter, less variation in another parameter is tolerable. Thus, the highest variation tolerance for each parameter is obtained when variation in other parameters is 0% (i.e. other parameters are set to the baseline values). In other words, less variation in each parameter is tolerable in the presence of variation in more than one parameter compared to the scenarios that only one parameter has variation. As shown in Figure 37 (a), the highest multiple variation tolerance for a combination of σd and σH_K is obtained when σd = 5% and σH_K = 7% which results in an aggregated variation of 12%. Figure 37 (b) exhibits that the highest multiple variation tolerance for a combination of σt_f and σd is attained when σt_f = 15% and σd = 5% which leads to an aggregated variation of 20%. The highest multiple variation tolerance for a combination of σt_f and σH_K is achievable when σt_f = 11% and σH_K = 7% which results in an aggregated variation of 18%. The results show that the p-bit based DBNs are more sensitive to the multiple variations of (σd vs. σH_K), (σt_f vs. σH_K), (σt_f vs. σd),



(a)



(b)



(c)

Figure 37: Accuracy of p-bit based DBN for: (a) σd vs. σH_K, (b) σt_f vs. σd, (c) σt_f vs. σH_K.

practice by changing the nanomagnet's fluctuation rate, as explained in the next section.

5.6.3 SOT P-BIT WITH FEEDBACK

In this section, we present a method to control the fluctuation frequency of the output of a p-bit. In this device, the output voltage tracks the magnetization direction through the anomalous Hall effect and fluctuates randomly between two values, "UP" and "DOWN". The energy barrier of the nanomagnet defines the average fluctuation frequency (f_0) as described in Equation (33). A DC current through the layer with GSHE biases the stochastic output of the device towards one of the two states. However, instead of a DC current, when the device's output is amplified and fed back to the layer with GSHE, the fluctuation of the magnetization based on the strength and polarity of the feedback gets slower or faster, analogous to temperature annealing. Figure 38 (a)

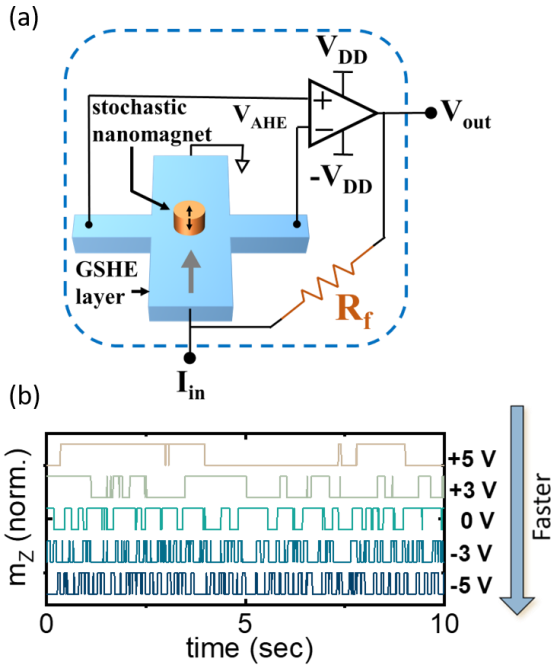


Figure 38: Tuning the effective energy barrier through electrical feedback. (a) Measurement configuration with the feedback implemented through a simple resistor of value 360 $K\Omega$. (b) Measurement of the output fluctuations of the device for various feedback configurations.

shows the device schematic with the feedback configuration. The value of the resistor R_f controls the feedback which alters the feedback current flowing through the layer with GSHE. However, for experimental simplicity, the resistor value is kept fixed and the V_{DD} of the amplifier is changed in both magnitude and sign. Figure 38 (b) shows the experimentally measured stochastic signal at the output of this device for different feedback configurations. A large positive V_{DD} corresponds to a strong positive feedback, while a large negative V_{DD} corresponds to a strong negative feedback. It can be seen that the fluctuations at the device output become progressively faster as the feedback changes from positive to negative (amplifier V_{DD} changing from +5V to -5V). The output signals for various feedback configurations are shifted artificially along the vertical axis for clarity. Please note that in this experiment, the average fluctuation time scale, t, is 144 *ms* at no feedback. This slow fluctuation speed is due to the fact that the nanomagnet used in this experiment had an energy barrier, E_B = 18 k_BT. In order to achieve fluctuations of 1 *ns*, this energy barrier can be reduced to be closer to 1 k_BT by two approaches:

1. By designing weaker perpendicular anisotropy stacks through the ferromagnetic layer thickness optimization. In a CoFeB/MgO perpendicular anisotropy magnetic stack, the effective anisotropy energy density is given by [231]:

$$\frac{1}{2}M_sH_K = K_{eff} = \frac{K_i}{t_F} - \frac{M_S^2}{2\mu_0} \tag{36}$$

where K_i is the interface anisotropy, t_F is the thickness of the ferromagnetic layer, M_S is the saturation magnetization and μ_0 is permeability of free space. It can be seen from the

above equation that by engineering the t_F to make the RHS close to zero, we can make the effective perpendicular anisotropy to vanish, hence resulting in very low E_B.

2. By fabricating magnets with smaller diameters through advanced lithography, the volume and hence the E_B of the nanomagnet can be made smaller. With currently available industrial lithographic technology, a diameter of 30 *nm* is possible [232].

In this device, a charge current input to the heavy metal electrode adjacent to the nanomagnet produces a torque on its magnetization via SOT. This was heuristically explained as a tilting of the energy landscape, where a positive current tilts the energy landscape towards the "UP" state and a negative current causes a tilt towards the "DOWN" state. When the device output is converted to a charge current and fed back to the device input, the tilt of the energy landscape dynamically depends on the instantaneous state of magnetization. This results in a dynamic modification of the effective energy barrier. The case for the negative feedback case is described as follows. When the magnetization is in the "UP" state, the negative feedback produces a negative charge current through the input that tilts energy landscape towards the "DOWN" state, thus reducing energy barrier to hop out of the "UP" state. likewise, once the magnetization is in the "DOWN" state, a positive feedback current is produced at the input that tilts the energy landscape towards the "UP" state, hence reducing energy barrier to hop out of the "DOWN" state. Hence, a negative feedback results is an overall reduction of the effective energy barrier to switch to the other state. Following the modified Neel-Brown model [233],[234] to include the effect of spin torque on the nanomagnet due to the current flowing through the layer with GSHE, Equations (33) and (34) are obtained for the fluctuation frequency. This results in a faster fluctuation rate, as seen by replacing E_B with a smaller $E_{B,eff}$ in Equation (29). By considering

GSHE layer resistance is much smaller than R_f, $I_{feedback}$ can be defined as V_{DD}/R_f. Then, by replacing V_{DD}/I_C with R_0, Equation (35) is obtained for the magnet's effective energy barrier. From Equation (33) and (35) it is clearly seen that the p-bit's fluctuation frequency can be managed by altering the feedback resistor, as is proved in the experiment.

CHAPTER 6: HIGH ACCURACY DBN-FUZZY NEURAL NETWORKS USING MRAM-BASED STOCHASTIC NEURONS

6.1 FUNDAMENTALS OF FUZZY SYSTEMS

Fuzzy system is a dynamic or static system utilizes fuzzy logic, fuzzy sets, and the corresponding mathematical framework [235],[272]. There are several ways that a system requires fuzzy sets, as explained below:

- *In the system's description*: As an illustration, a system can be defined as a fuzzy relation or as a collection of if-then rules with fuzzy predicates. For instance, the relationship between a room's temperature trend and a heating power would be described as the following fuzzy rule:

  If the heating power is low **then** the temperature will rise slow

- *In the system parameters specification*: By employing fuzzy numbers rather than real numbers for parameters, a system can be described by a differential Equation or algebraic. As an illustration, suppose an Equation: $w = \tilde{4}z_1 + \tilde{7}z_2$, where membership functions define $\tilde{4}$ and $\tilde{7}$ as fuzzy numbers "about four" and "about seven", respectively. The uncertainty in the values of parameter is expressed by fuzzy numbers.
- A system's state variables, output, and input may be fuzzy sets. Human perception quantities such as beauty and comfort or unreliable sensors ("noisy" data) can provide fuzzy inputs. While the conventional (crisp) systems cannot process this kind of information, fuzzy systems can do it.

Table 16: Crisp and fuzzy information in systems.

System Description	Input Data	Resulting Output Data	Mathematical Framework
Crisp	Crisp	Crisp	Functional Analysis, Linear Algebra, etc.
Crisp	Fuzzy	Fuzzy	Extension Principle
Fuzzy	Crisp/Fuzzy	Fuzzy	Fuzzy Relational Calculus, Fuzzy Inference

Several of the aforementioned attributes can be employed in a fuzzy system at the same time. A summary of the relationships between the variables and descriptions of crisp and fuzzy systems is given in Table 16. The fuzzily described systems with fuzzy or crisp inputs are discussed in this chapter. The fuzzy systems employ rule-based fuzzy systems, i.e., if-then rules are more common and we will discuss only about these systems in the rest of this chapter. Various goals such as data analysis, modeling, prediction or control can be served as fuzzy systems. For simplicity, without regard to the goal of system, a fuzzy rule-based system is called a fuzzy model in this chapter.

6.2 RULE-BASED FUZZY MODELS

The variables relationships in rule-based fuzzy systems are represented by employing the following common form of fuzzy if–then rules:

If antecedent proposition **then** consequent proposition

The fuzzy proposition type of "$\tilde{z}$ is A" forms the antecedent proposition where A is a linguistic constant (term) and $\tilde{z}$ is a linguistic variable. The amount of similarity between A and $\tilde{z}$ specifies the truth value of proposition which is a real number within one and zero. Two major types of rule-based fuzzy models can be represented based on the consequent's form:

- *Linguistic Fuzzy Model*: both the consequent and the antecedent are fuzzy propositions.
- *Takagi–Sugeno-Kang (TSK) Fuzzy Model*: the consequent is a crisp function but the antecedent is a fuzzy proposition.

In the next subsections, these two types of fuzzy models are discussed more.

6.2.1 LINGUISTIC FUZZY MODEL

In order to obtain available (semi-)qualitative knowledge in the form of if–then rules, the model of linguistic fuzzy has been proposed as following [236],[237]:

$$R_i\text{: If } \tilde{z} \text{ is } A_i \text{ then } \tilde{w} \text{ is } B_i, \qquad i = 1, 2, \ldots, P \tag{37}$$

Here, A_i and B_i are the antecedent and consequent linguistic terms, and $\tilde{z}$ and $\tilde{w}$ are the antecedent and consequent linguistic variables, respectively. The linguistic terms A_i (B_i) and values of $\tilde{z}$ ($\tilde{w}$) are fuzzy sets described in their corresponding base variables' domain: $z \in Z \subset \mathbb{R}^p$ and $w \in W \subset \mathbb{R}^q$. The antecedent (consequent) fuzzy sets' membership functions are based on the mappings: $\mu(z)$: $Z \rightarrow [0, 1]$, $\mu(w)$: $W \rightarrow [0, 1]$. Fuzzy regions in the antecedent space are defined by fuzzy sets A_i while the corresponding consequent propositions hold. Generally, predefined terms such as Medium and Small describe the linguistic terms A_i and B_i. Thus, we have $A_i \in A$ and $B_i \in B$ by denoting these sets by A and B, respectively. The sets A and B, and rule base $R = \{R_i \mid i = 1, 2, \ldots, P\}$ form the linguistic model's knowledge base.

We require an algorithm which permits us to compute the output value for some assigned input values in the interest of utilizing the linguistic model which is called the fuzzy inference algorithm. By employing fuzzy relational calculus, the inference mechanism can be achieved for the linguistic model as discussed in the subsequent subsection.

6.2.1.1 RELATIONAL REPRESENTATION OF A LINGUISTIC MODEL

Fuzzy relation can be employed for each rule in Equation (37): R_i: $(Z \times W) \rightarrow [0, 1]$. Two different primary ways can compute this relation: (1) *Mamdani Method*: utilizing fuzzy conjunctions, and (2) *Fuzzy Logic Method*: utilizing fuzzy implications, as illustrated in [238]. Once the if-then rule in Equation (37) is strictly considered as "*A* implies *B*" ($A_i \rightarrow B_i$) then we can use fuzzy implications. This means that if *A* holds then *B* must hold too in order to obtain a true implication in classical logic. On the other hand, not only we can say nothing about *B* when *A* does not hold but also, we cannot invert the relationship.

The if-then rules' interpretation is "it is true that *A* and *B* hold at the same time" once utilizing a conjunction ($A \wedge B$). We can invert this relationship since this is symmetric. Herein, the minimum ($\wedge$) operator computes the relation *R* as follows:

$$R_i = A_i \times B_i, \text{ that is, } \mu_{R_i}(z, w) = \mu_{A_i}(z) \wedge \mu_{B_i}(w) \tag{38}$$

For all feasible pairs of z and w, the Cartesian product space of Z and W computes the minimum. The disjunction of the *P* individual rule's relations R_i defines the whole model in Equation (37) as follows:

$$R = \bigcup_{i=1}^{P} R_i, \text{ that is, } \mu_R(z, w) = \max_{1 \le i \le P} [\mu_{A_i}(z) \wedge \mu_{B_i}(w)] \tag{39}$$

Then, the relational max-min composition (o) can compute the linguistic model while the fuzzy relation *R* encodes the whole rule base:

$$\widetilde{w} = \tilde{z} \text{ o } R \tag{40}$$

6.2.1.2 MAX-MIN (MAMDANI) INFERENCE

Fuzzy relation can represent a rule base as explained in the previous subsection. Then, the max-min relational composition computes a rule-based fuzzy model's output. We will show that the relational calculus can be by-passed in this subsection. This results in a benefit that we can avoid storing of the relation R and the domains' discretization. In order to prove it, assume the relational composition provides the output value B' for an input fuzzy value $\tilde{z} = A'$:

$$\mu_{B'}(w) = \max_{Z}[\mu_{A'}(z) \wedge \mu_R(z, w)] \tag{41}$$

Below expression is attained by replacing $\mu_R(z, w)$ with Equation (39):

$$\mu_{B'}(w) = \max_{Z}\{\mu_{A'}(z) \wedge \max_{1 \le i \le P}[\mu_{A_i}(x) \wedge \mu_{B_i}(w)]\} \tag{42}$$

The order of max and min operation can be altered as follows since this operation can be taken over different domains:

$$\mu_{B'}(w) = \max_{1 \le i \le P}\{\max_{Z}[\mu_{A'}(z) \wedge \mu_{A_i}(z)] \wedge \mu_{B_i}(w)\} \tag{43}$$

The linguistic model's output fuzzy set is as follows by denoting $\beta_i = \max_{Z}[\mu_{A'}(z) \wedge \mu_{A_i}(z)]$ as the degree of fulfillment of the *i*th rule's antecedent:

$$\mu_{B'}(w) = \max_{1 \le i \le P}[\beta_i \wedge \mu_{B_i}(w)], \quad w \in W \tag{44}$$

Algorithm 6 summarizes the whole algorithm which is called the Mamdani or max-min inference.

Algorithm 6: Mamdani (Max-Min) Inference

1. Compute the degree of fulfillment by: $\beta_i = \max_{Z}[\mu_{A'}(x) \wedge \mu_{A_i}(x)],\ 1 \leq i \leq P$
   Note that for a singleton fuzzy set ($\mu_{A'}(z) = 1$ for $z = z_0$ *and* $\mu_{A'}(z) = 0$ otherwise) the equation for β_i simplifies to $\beta_i = \mu_{A_i}(z_0)$
2. Derive the output fuzzy sets B_i': $\mu_{B'_i}(w) = \beta_i \wedge \mu_{B_i}(w),\ w \in W,\ 1 \leq i \leq P$
3. Aggregate the output fuzzy sets B_i': $\mu_{B'}(w) = \max_{1 \leq i \leq P} \mu_{B'_i}(w),\ w \in W$

6.2.1.3 MULTIVARIABLE SYSTEMS

Generally, the linguistic model was discussed so far includes both the MIMO and SISO cases. By employing multivariate membership functions, vector domains can define all fuzzy sets in the model in the MIMO case. On the other hand, univariate membership functions can write suitably the consequent and antecedent propositions as fuzzy propositions' logical combinations. In order to combine the propositions, we can employ fuzzy logic operators such as the negation, disjunction, and conjunction. Moreover, a set of MISO models can create a MIMO model. Herein, we will write MISO systems' rules for the ease of notation. The conjunctive form of the antecedent is more popular as follows:

$$R_i\text{: If } z_1 \text{ is } A_{i1} \text{ and } z_2 \text{ is } A_{i2} \text{ and } \ldots \text{ and } z_p \text{ is } A_{iq} \text{ then } w \text{ is } B_i, \quad i = 1, 2, \ldots, P \tag{45}$$

We can conclude the aforementioned model is a special format of Equation (37) since the fuzzy set A_i in Equation (37) is achieved as the Cartesian product of fuzzy sets A_{ij}: $A_i = A_{i1} \times A_{i2} \times \ldots \times A_{iq}$. Therefore, the step 1 of Algorithm 6 (degree of fulfillment) can be written as:

$$\beta_i = \mu_{A_{i1}}(z_1) \wedge \mu_{A_{i2}}(z_2) \wedge \ldots \wedge \mu_{A_{iq}}(z_q),\ 1 \leq i \leq P \tag{46}$$

We can utilize the product or other conjunction operators. The input domain can be divided into a lattice of fuzzy hyperboxes by a set of rules in the conjunctive antecedent form. Each Cartesian product-space intersection of the corresponding univariate fuzzy sets is considered as an hyperboxes. For covering the entire domain, the number of rules in the conjunctive form can be obtained as follows:

$$P = \prod_{i=1}^{q} N_i$$

where the number of linguistic terms of the *i*th antecedent variable is N_i and the input space's dimension is q.

6.2.1.4 DEFUZZIFICATION

A crisp output w is suitable in several applications. The output fuzzy set should be defuzzified in the interest of obtaining a crisp value. The center of gravity (COG) defuzzification technique is utilized in the Mamdani inference scheme. The w coordinate of the center of gravity of the area under the fuzzy set B' is computed in this technique as follows:

$$w' = cog\ (B') = \frac{\sum_{j=1}^{F} \mu_{B'}(w_j) w_j}{\sum_{j=1}^{F} \mu_{B'}(w_j)} \quad (47)$$

where the number of elements w_j in *W is F*. In order to compute the center of gravity, continuous domain *W* must be discretized.

6.2.1.5 SINGLETON MODEL

Whenever the consequent fuzzy sets B_i are singleton fuzzy sets then a special case of the linguistic fuzzy model is achieved which is called the singleton model. By employing real numbers b_i, these sets can be denoted easily as the below rules:

$$R_i\text{: If } \tilde{z} \text{ is } A_i \text{ then } w = b_i, \qquad i = 1, 2, \ldots, P \tag{48}$$

We generally can use this model with a simplified inference/defuzzification method which is called the fuzzy mean:

$$w = \frac{\sum_{i=1}^{P} \beta_i b_i}{\sum_{i=1}^{P} \beta_i} \tag{49}$$

The singleton fuzzy model pertains to the basis functions expansion [239] which is a general class of general function approximators:

$$w = \sum_{i=1}^{P} \emptyset_i(x) b_i \tag{50}$$

This systems' class covers almost all nonlinear system identification's structures such as splines, radial basis function networks, or artificial neural networks. [240],[241] investigate connections between these models' types. The constants b_i are the consequents in the singleton model and the degrees of fulfillment of the rule antecedents provides basis functions $\emptyset_i(x)$. We can obtain multilinear interpolation among the rule consequents if

- For each domain element, the membership degrees sum up to one and the antecedent membership functions are trapezoidal, pairwise overlapping
- The connective and logical in the rule antecedents is represented by the product operator.

In addition, any linear mapping of the form can be represented by a singleton model:

$$w = q^T z + v = \sum_{i=1}^{q} q_i x_i + v \tag{51}$$

The antecedent membership functions should be triangular in this case. Through assessing the suitable mapping in Equation (51) for the cores a_{ij} of the antecedent fuzzy sets A_{ij}, we can calculate consequent singletons:

$$b_i = \sum_{j=1}^{q} q_j a_{ij} + v \tag{52}$$

We can benefit from this property by initializing the fuzzy model in a way that it mimics a predefined linear model and then optimize it later.

6.2.2 TAKAGI-SUGENO-KANG MODEL

The introduced linguistic model defines a predefined system by employing linguistic if-then rules with fuzzy proposition in both the consequent and antecedent. However, crisp functions in the consequents are utilized in the Takagi–Sugeno-Kang (TSK) fuzzy model [242]. Therefore, this model is considered as a combination of mathematical and linguistic regression modeling in a way that the antecedents define fuzzy regions for consequent functions which are valid in the input space. The below form shows the TSK rules:

$$R_i\text{: If } z \text{ is } A_i \text{ then } w_i = f_i(z), \quad i = 1, 2, \ldots, P \tag{53}$$

In this model, the input z is a crisp variable opposed to the linguistic model. In the functions f_i, each rule's parameters are only different but generally have the same structure. For the ease of notation, we will consider a scalar f_i in the result but f_i is a vector-valued function in general. By considering the affine form, an affine TSK model is attained based on the following simple rules:

$$R_i\text{: If } z \text{ is } A_i \text{ then } a_i^T z + b_i, \quad i = 1, 2, \ldots, P \tag{54}$$

where a scalar offset is b_i and a parameter vector is a_i. The singleton model in Equation (48) can be achieved if for each *i we have* $a_i = 0$.

6.2.2.1 INFERENCE MECHANISM

A direct extension of the singleton model inference in Equation (49) results in the TSK model's inference formula as follows:

$$w = \frac{\sum_{i=1}^{P} \beta_i w_i}{\sum_{i=1}^{P} \beta_i} = \frac{\sum_{i=1}^{P} \beta_i (a_i^T z + b_i)}{\sum_{i=1}^{P} \beta_i} \tag{55}$$

The TSK model can be considered as a smoothed piece-wise estimation of a nonlinear function when the parameters b_i and a_i correspond to that function's local linearization and the antecedent fuzzy sets consider overlapping regions in the antecedent space and distinct at the same time [235].

6.2.2.2 TSK MODEL AS A QUASI-LINEAR SYSTEMS

By denoting the normalized degree of fulfillment, the affine TSK model can be considered as a quasi-linear system:

$$\gamma_i(z) = \beta_i(z) / \sum_{j=1}^{P} \beta_j(z) \tag{56}$$

In order to emphasize that the TSK model is a quasi-linear model of the below form, we represent $\beta_i(z)$ specifically as a function *z*:

$$w = (\sum_{i=1}^{P} \gamma_i(z) a_i^T) z + \sum_{i=1}^{P} \gamma_i(z) b_i = a^T(z) z + b(z) \tag{57}$$

The convex linear combinations of the consequent parameters b_i and a_i are 'parameters' *b(z)*, *a(z)*:

$$a(z) = \sum_{i=1}^{p} \gamma_i(z) a_i, \quad b(z) = \sum_{i=1}^{p} \gamma_i(z) b_i \tag{58}$$

In the space of a quasi-linear system's parameters, a TSK model can be considered as a mapping from the input (antecedent) space to a polytope region (convex) that provide identical analysis of linear systems and TSK models in a framework. In [243]-[246], several methods have been proposed to design controllers with suitable closed loop characteristics and to evaluate their stability.

6.2.3 MODELING DYNAMIC SYSTEMS

As mentioned earlier, by utilizing the concept of the system's state, static functions generally model time-invariant dynamic systems. We can specify what the consequent state will be by using the system's state and input. We can have the following Equation in the discrete-time setting:

$$z(p+1) = f(z(p), u(p)) \tag{59}$$

where the input and the state at time *p are u(p)* and *z(p)*, respectively, and the state-transition function is *f* which is a static function. We can estimate the state-transition function by utilizing fuzzy models of different types. Input-output modeling is often applied while the process' state is usually not measured. The NARX (Nonlinear AutoRegessive with eXogenous input) is the most prevalent model:

$$w(p+1) = f(w(p), w(p-1), \dots, w(p-n_w+1), u(p), u(p-1), \dots, u(p-n_u+1)) \tag{60}$$

where the past model inputs and outputs are denoted by and *u(p), …, u(p-*n_u*+ 1)* and *w(p), …, w(p-*n_w*+ 1)*, respectively, and integers related to the model order are n_u, n_y. As an illustration, the following form's rules may be included in a dynamic system's linguistic fuzzy model:

R_i: **If** *w(p)* is A_{i1} **and** *w(p-1)* is A_{i2} **and**, … w(p-n+1) is A_{in} **and** *u(k)* is B_{i1} **and** *u(k-1)* is B_{i2} **and**, …, *u(k-m+1)* is B_{im} **then** *y(k+1)* is C_i (61)

In this case, external dynamic filters added to the fuzzy system take care of the dynamic behavior. In Equation (61), no output filter is utilized and the input dynamic filter is a straightforward generator of the lagged outputs and inputs.

As mentioned in [247], any smooth function to any accuracy's degree can be estimated by the fuzzy models. Thus, any controllable and observable modes of a large class of discrete-time nonlinear systems can be estimated by type's models in Equation (61) [248].

6.3 BUILDING FUZZY MODELS

The data and prior knowledge are two general information's sources for creating fuzzy models. The prior knowledge normally developed by "experts", i.e., operators, process designers, etc. and can be of a rather approximate nature. In this case, straightforward fuzzy expert systems can define fuzzy models [249].

For most of the processes, data are accesible as records of the special identification experiments or process operation can be implemented to attain the relevant data. Not only approximate reasoning and fuzzy logic can create fuzzy models from data, but also ideas originating from the field of conventional systems identification, data analysis, and neural networks. Fuzzy identification means the tuning or acquisition of fuzzy models by using data.

In a fuzzy model, two fundamental approaches can be defined for the integration of data and knowledge:

1. A set of if–then rules translates the verbal form of expert knowledge that creates a certain model. By employing input-output data, we can fine-tune the parameters in this structure including parameters, consequent singletons, or membership functions. Alike to artificial neural networks that we can apply standard learning algorithms, a fuzzy model can be considered as a layered network at the computational level that is exploited by the certain tuning algorithms. Commonly, this technique is called neuro-fuzzy modeling [240],[250],[251].
2. A fuzzy model is created from data and in order to formulate the rules, no antecedent knowledge about the under-study's system is initially utilized. The membership functions and extracted rules are expected to prepare a posteriori interpretation of the behavior of system. In the interest of obtaining more informative data, an expert can design additional experiments. Moreover, an expert can supply new rules or alter the rules.

Based on the specific application, we can combine these techniques. At the end, we define the primary techniques to fine-tune or extract fuzzy models by utilizing data and the primary choices and steps in the fuzzy models' knowledge-based construction [235].

6.3.1 STRUCTURE AND PARAMETERS

Two fundamentals parts in the fuzzy models' design are recognized: the parameters and the structure of the model. The model's flexibility in approximation mappings is specified by the structure. Then, the parameters are estimated in the interest of fitting the data at hand. While a model with a rich structure has worse generalization properties but has the ability to estimate more sophisticated functions. Good generalization provides this ability that a model can carry out on another data set from the same process as well as the data set that is fitted to.

Structure selection includes the below choices in fuzzy models:

- *Output and input variables*: It is sometimes unclear which variables must be utilized as the model's inputs in complex systems. The system's order should be estimated in the dynamic systems. This means to determine the number of output and input lags n_u and n_w, respectively, for the input-output NARX model in Equation (60). The common information's sources for this choice are the modeling's purpose and insight in the process behavior. In order to compare different choices by considering some performance metric, we can sometimes utilize automatic data-driven selection.
- *The rules' structure*: The antecedent form and the model type (Takagi-Sugeno-Kang, singleton, linguistic) are involved in this choice. The type available knowledge and the purpose of modeling are important aspects.
- *Membership functions' type and number for each variable*: The level of the model's granularity is determined by this choice. This choice will be influenced by the available knowledge's details and the modeling's purpose. Membership functions can be removed or added to the model through data-driven methods.
- *Defuzzification method, connective operators, the inference mechanism's type*: The fuzzy model's type restricts these choices (TSK, Mamdani). Anyway, some freedom such as the conjunction operators' choice remains within these restrictions. Differentiable operators such as sum and product are commonly preferred to the standard min and max operators in the interest of facilitating fuzzy models' data-driven optimization.

By adjusting the parameters, the fuzzy model's performance can be fine-tuned after the structure is fixed. Linguistic models' tunable parameters are the parameters of consequent membership

functions and antecedent, and the rules. Takagi-Sugeno-Kang models have parameters in the consequent functions and in antecedent membership functions.

6.3.2 KNOWLEDGE-BASED DESIGN

The below steps must be followed with the intention of designing a fuzzy model in accordance with available expert knowledge:

1. Select the defuzzification and inference methods, the rules' structure and the output and input variables.
2. Define the corresponding membership functions and decide on the linguistic terms' number for each variable.
3. Utilize fuzzy if-then rules to formulate the available knowledge.
4. Verify the model by utilizing data. Repeat the aforementioned design steps if the model does not fulfill the desired performance.

The quality and extent of the available knowledge, and the prepared problem determine the success of this approach. For certain problems, the knowledge-based design may be an inefficient and very time-consuming procedure, while for other problems it may lead to useful models quickly. As a result, a combination of a data-driven tuning of the model parameters and the knowledge-based design would be more useful. The next subsections discuss various approaches for the fuzzy model parameters' adjustment by using data.

6.3.3 DATA-DRIVEN ACQUISITION/TUNNING OF FUZZY MODELS

In this subsection, we suppose that a set of M input-output data pairs $\{(z_i,w_i)|i = 1, 2, \dots, M\}$ is accessible while w_i are output scalars and $z_i \in \mathbb{R}^q$ are input vectors. Denote $w \in \mathbb{R}^M$ a vector containing the outputs w_p and $Z \in \mathbb{R}^{M\times q}$ a matrix having the vectors z_P^T in its rows:

$$Z = [z_1, \dots, z_M]^T, \quad w = [w_1, \dots, w_M]^T \tag{62}$$

6.3.3.1 LEAST-SQUARE ESTIMATION OF CONSEQUENTS

As shown in Equations (55) and (49), the TSK and singleton models' defuzzification formulas are linear in the consequent parameters (b_i and a_i). By employing least-squares techniques, we can estimate these parameters from the available data. Denote $\Gamma_i \in \mathbb{R}^{M\times M}$, as its pth diagonal element, the diagonal matrix having the normalized membership degree $\gamma_i(z_p)$ of Equation (56). The extended matrix $Z_e = [Z,1]$ can be created by adding a unitary column to Z. Then, the products of matrices Γ_i and Z_e compose Z' the matrix in $\mathbb{R}^{M\times PM}$:

$$Z' = [\Gamma_1 Z_e, \Gamma_2 Z_e, \dots, \Gamma_K Z_e] \tag{63}$$

The single parameter vector $\theta \in \mathbb{R}^{P(q+1)}$ contains the consequent parameters b_i and a_i:

$$\theta = [a_1^T, b_1, a_2^T, b_2, \dots, a_P^T, b_P]^T \tag{64}$$

Now, Equation (55) can be described in a matrix form of $w = Z'\theta + \epsilon$ by utilizing the data Z and w. We can solve this set of Equations for the parameter θ with linear algebra [252]:

$$\theta = [(Z')^T Z']^{-1} (Z')^T w \tag{65}$$

This solution gives us the minimal prediction error as an optimal least-squares solution which is appropriate for prediction models. On the other hand, as local models' parameters, it may bias

the consequent parameters' estimation. We can apply a weighted least-squares approach per rule in the interest of obtaining an accurate approximation of local model parameters:

$$[a_i^T, b_i^T] = [Z_e^T \Gamma_i Z_e]^{-1} Z_e^T \Gamma_i \, w \tag{66}$$

The individual rules' consequent parameters are not "biased" by the rules' interactions since they are approximated separately from each other. Equations (65) and (66) can be directly applied to the singleton model in Equation (48) by deleting a_i for all $1 \leq i \leq P$.

6.3.3.2 TEMPLE-BASED MODELING

By employing this approach, we can easily partition the antecedent variables' domains into a given number of equally shaped and spaced membership functions. Then, all the antecedent terms' combinations are covered by the established rule base. The least-squares method approximates the consequent parameters. As an illustration, assume a first-order difference Equation describes a nonlinear dynamic system as below:

$$w(p+1) = w(p) + u(p)e^{-3|w(p)|} \tag{67}$$

In order to generate a set of 300 input–output data pairs with this Equation, a stepwise inputs signal is used. The below TSK rule structure can be selected by assuming that the system's nonlinearity is only caused by *w* and the system is first order:

$$\textbf{If } w(p) \text{ is } A_i \textbf{ then } w(p+1) = a_i w(p) + b_i u(p) \tag{68}$$

In the domain of *w(p)*, seven triangular membership functions (A_1 to A_7) with equal space are determined by supposing that we have no more prior knowledge.

The combination of known mechanistic models' linearization and local models attained by parameter estimation is facilitated by the TSK model's transparent local structure. The remaining

regions' parameters can be attained by linearizing the process' mechanistic model if measurements are accessible only in specified regions of the process' operating domain. Assume *w = f(z)* predefines this model. The below affine TSK model's parameters in Equation (54) are resulted from linearization around the center c_i of the *i*th rule's antecedent membership function:

$$a_i = \frac{df}{dz} |z = c_i, \;\; b_i = f(c_i) \tag{69}$$

In the template-based approach, the number of rules in the model may rise quickly which is considered as a disadvantage of this approach. Generally, all the antecedent variables are partitioned uniformly if we have no knowledge to show which variables make the system's nonlinearity. This kind of partitioning results in the number of rules' exponential increase.

Some specific regions need almost fine partitioning while other regions can be estimated very well by a single model. This means that the system behavior complexity is usually not uniform. The membership functions should capture the system's non-uniform behavior with the purpose of obtaining an efficient representation with the minimum rules. We often need to form the membership functions by utilizing system measurements, as explained in the next subsections.

6.3.3.3 NEURO-FUZZY MODELING

It is shown that least-squares methods can optimally approximate parameters that are linearly related to the output. We can utilize the known training algorithms from the neural networks' area in the interest of optimizing the parameters which are related in a nonlinear way to the output. Similar to artificial neural networks, a fuzzy model can be represented as a layered structure at the computational level. Therefore, this technique is generally called neuro-fuzzy modeling [240],[241],[250][241].

6.3.3.4 FUZZY CLUSTERING

Fuzzy-clustering-based identification methods employ the graded membership's concept to show the similarity's degree of some typical object and a predefined object. An appropriate distance measure calculates the similarity's degree. Feature vectors can be partitioned based on the similarity in a way that the vectors from different clusters are too dissimilar and vectors within a cluster are too similar.

By employing the Euclidean distance measure, fuzzy clustering groups the data into two clusters which calls v_1 and v_2. The fuzzy partition matrix expresses the data's partitioning in a way that in a fuzzy cluster with prototypes v_j, elements μ_{ij} are degrees of membership of the data points [z_i, w_i].

By projecting the clusters onto the axes, we can extract fuzzy if-then rules. Assume, we have a data set with two associated fuzzy rules and two apparent clusters. The concept of data's similarity to a predefined prototype gives enough space for the choice of the prototype's character itself and of an appropriate distance measure. For instance, the clusters can be ellipsoids with adaptively determined shape [253], or the prototypes can be described as linear subspaces [254]. From such clusters, we can extract the consequent parameters and the antecedent membership functions of the Takagi–Sugeno-Kang model as below [255]:

$$\textbf{If } z \text{ is } A_1 \textbf{ then } w = a_1 z + b_1$$

$$\textbf{If } z \text{ is } A_2 \textbf{ then } w = a_2 z + b_2$$

One rule in the Takagi–Sugeno-Kang model represent each achieved cluster. The partition matrix's point-wise projection onto the antecedent variables generates the membership functions for fuzzy sets A_1 and A_2. Then, an appropriate parametric function estimates these point-wise

defined fuzzy sets. Least-squares estimates in Equations (65) or (66) obtains the consequent parameters for each rule.

6.4 PROPOSED DBN-FUZZY NEURAL NETWORK

A neuro-fuzzy system is an integration of neural networks' learning ability and rule-based fuzzy systems' interpretability [242],[256],[257],[258]. Consequently, an integrated deep neural network fuzzy system can benefit from the advantages of both a deep network and a fuzzy system. In this section, we present a DBN-fuzzy network for image classification on the basis of Takagi-Sugeno-Kang (TSK) system [242]. This system models fuzzy rule-based systems which is composed of several fuzzy rules. A TSK rule set is described as:

$$\text{If } x_1 \text{ is } A_{1j} \text{ and } x_2 \text{ is } A_{2j} \text{ And } \ldots x_d \text{ is } A_{dj} \text{ then } y_j = f_j(x) \tag{70}$$

where $j = 1, 2, \ldots, J$, the number of fuzzy rules is J, the ith input variable x_i, a fuzzy set for ith input of jth rule is A_{ij}, a fuzzy conjunction operator is And, and output of kth rule is $f_j(.)$. The system's output is defined as:

$$y = \frac{\sum_{j=1}^{J} p_j(x).f_k(x)}{\sum_{j=1}^{J} p_k(x)} = \sum_{j=1}^{J} \bar{p}_j(x).f_k(x) \tag{71}$$

where $p_j(x) = \prod_{i=1}^{d} p_{A_{ij}}(x_i)$ and $p_{A_{ij}}(x_i)$ is a membership grade determining similarity's degree of x_i and A_{ik}.

The DBN-Fuzzy neural network structure is shown in Figure 39. In the first phase, the MRAM-based DBN is employed to identify the top recognition results with the highest probability. Figure 39 (a) to (d) shows a 784 × 200 × 10 DBN structure for MNIST digit recognition dataset which is implemented by the PIN-Sim framework [12]. As shown in Figure 39 (c) and (d), weighted connections and activation functions are implemented with memristive crossbars and p-bit respectively. In the second phase, a fuzzy system is utilized to obtain the top-1 recognition results. Herein, we have employed a TSK-based fuzzy system is presented in [259] and modified it to be compatible with our MRAM-based DBN which is implemented in PIN-Sim framework. In this system, a set of subregions of an input image is assessed as the universe of discourse, a specific pattern is examined as a fuzzy set, and the similarity amongst the subregions and the

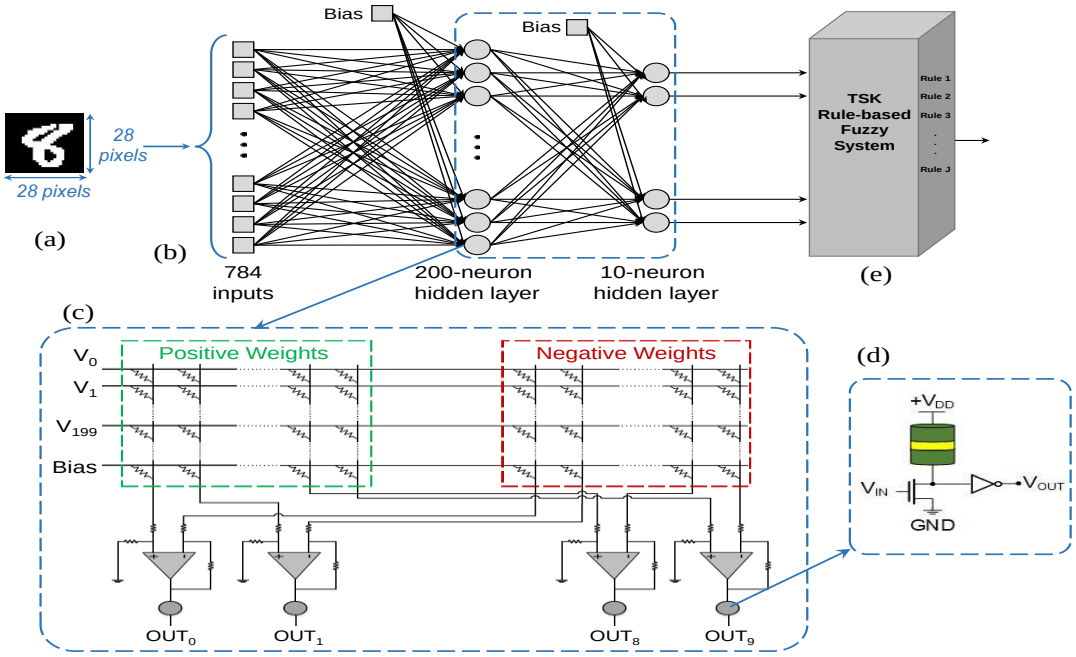


Figure 39: The DBN-Fuzzy system used for application-level simulations. (a) an input image from MNIST dataset, (b) a 784×200×10 DBN developed for MNIST pattern recognition application, (c) hardware implementation of the 784×200×10 DBN using PIN-Sim tool, (d) stochastic MRAM-based neuron (p-bit), and (e) TSK rule-based fuzzy system.

specified pattern is considered as the membership grade. By considering each fuzzy rule take a distinct pattern in the input image, Equation (70) can be reworded as follows:

If subregions and pattern identified by A_j are similar and located in the exact same place then

$$\begin{pmatrix} f_{1j}(x) \\ f_{2j}(x) \\ \vdots \\ f_{nj}(x) \end{pmatrix},\ j= 1,\ \ldots,\ J \tag{72}$$

where number of rules is *J*, number of outputs is *n*, a specific pattern for the *j*th rule defined over set of subregions is A_j, a subregion from set of subregions is *x*, and *i*th output for the *j*th pattern is $f_{ik}(x)$. The final recognition result is made based on the outputs of each rule through the following steps [259], according to Algorithm 7:

I. *Matrix Membership Grades*: In this step, the similarity among a specific pattern (A_j) and a subregion (a_i) in the image is measured for each rule through calculating a matrix of membership grades (P_j) by employing dot product operation as follows:

$$P_j = \left[p_{ij}\right] = \left[a_i . A_j\right] \tag{73}$$

II. *Firing Strength Measurement*: Each rules' firing strength is measured through membership grade matrix's normalization:

$$\overline{P_j} = \frac{P_j}{\sum_{l=1}^{J} P_l} \tag{74}$$

III. *Final Outputs Computation*: The final outputs (y_i) are calculated based on the outputs of each rule (y_{ik}) by multiplying $\overline{P_j}$ to its equivalent subregion ($g_{ij}(x)$):

$$y_i = \sum_{j=1}^{J} y_{ij} = \sum_{j=1}^{J} \overline{p_j}\, g_{ij}(x) \tag{75}$$

Algorithm 7: TSK Rule-based Fuzzy System Algorithm

for all input images **do**
  Get the outputs of MRAM-based DBN
  Find top outputs of MRAM-based DBN
  for each top output **do**
    Calculating membership matrix (P_j) of the input image based on the specified pattern for each top output
    Obtain membership grade matrix's normalization ($\overline{P_j}$) of the input image for each top output
  Outputs final computation (y_i) of the input image for each top output
  end
end

6.5 SIMULATION RESULTS

In this section, we implement the proposed DBN-Fuzzy system for MNIST digit recognition dataset by employing the structure of a 784×10 DBN circuit in the PIN-Sim framework and developing the TSK rule-based fuzzy system using Python scripts. The input of this system is a 28×28 digit image and the size of patterns as the fuzzy set is 7×7. As shown in Figure 40, the input image is divided into 16 subregions which their sizes are identical to the 20 patterns that are identified in blue boxes. For each specified digit, the patterns are chosen by considering which digits are misrecognized with the

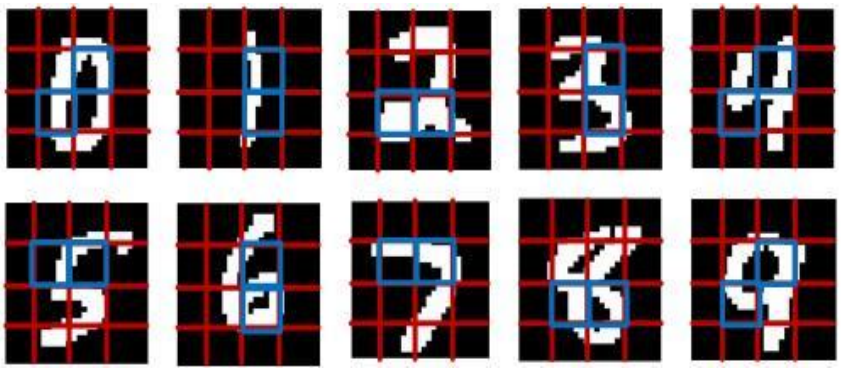


Figure 40: Input image subregions and identified patterns for each input digit in MNIST dataset.

specified correct digit in the MRAM-based DBN. For example, digit “3” is misrecognized with digits “2”, “5”, and “8” by the MRAM-based DBN. For each top output from MRAM-based DBN, a membership matrix is attained by measuring similarity among the pattern and its corresponding subregion utilizing dot product. Then, the strength of each pattern is measured by normalizing each member in the membership matrix with its equivalent member in the other membership matrices. Finally, the output of each pattern is calculated through an element-wise multiplication of the normalized membership grade matrix and its equivalent subregion.

Figure 41 shows the top-1 accuracy of PIN-Sim framework with and without the TSK rule-based fuzzy system for all input digits and each separate digit. As you can see, the proposed DBN-Fuzzy-based PIN-Sim always shows higher accuracy than DBN-based PIN-Sim. Our results show that the accuracy of PIN-Sim is increased from 64% to 82% for all digits by employing the TSK rule-based fuzzy system. As can be seen, each individual digit has at least 7.1% improvement except digits “5”, “6”, and “8”. For each individual digit, the highest top-1 accuracy is obtained for digits “0” and “8” with 100% accuracy, and the lowest top-1 accuracy is obtained for the digit “0” and “8” with 100% accuracy and the lowest top-1 accuracy is obtained for the digit “5” with 57.1% accuracy while for PIN-Sim without TSK rule-based fuzzy system,

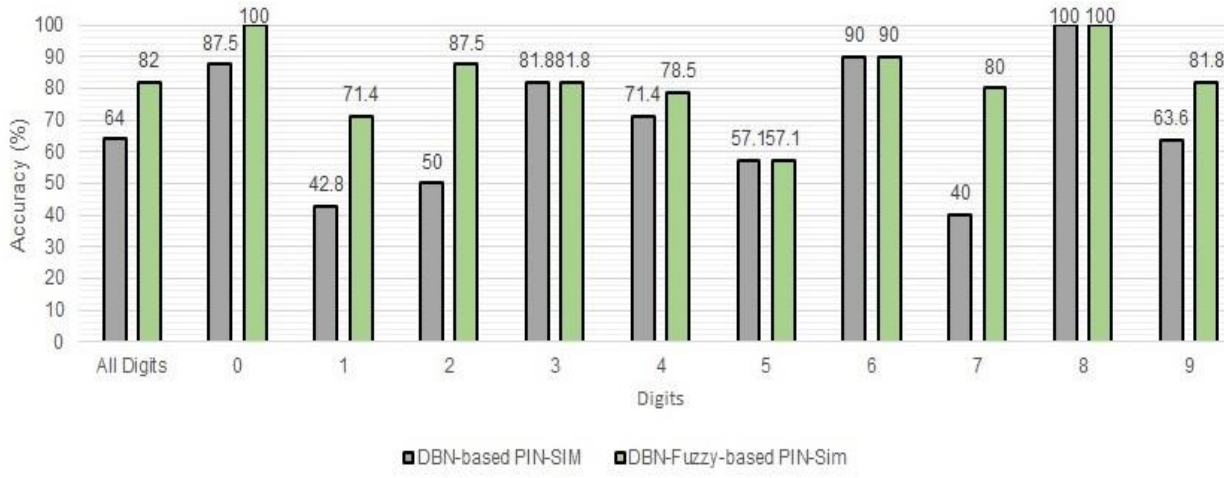


Figure 41: PIN-Sim Top-1 Accuracy for MNIST dataset.

the top-1 accuracy for digits "1", "2", "5", and "7" are below 60%. The highest recognition enhancement is for digit "7" with 40% improvement and the lowest recognition enhancement is for the digit "4" with 7.1% improvement.

Figure 42 shows the top-1 accuracy of the TSK rule-based fuzzy system for all input digits and each individual digit. In other terms, this graph shows the ability of the TSK rule-based fuzzy

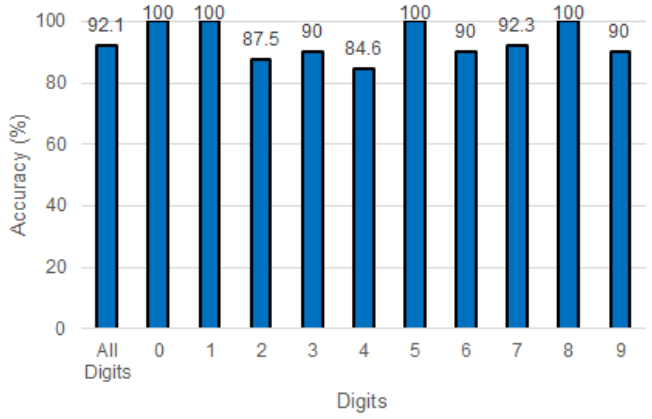


Figure 42: TSK Rule-based Fuzzy System Top-1 Accuracy for MNIST data set.

system to recognize the correct digit in the cases that the correct digit is among the top outputs of MRAM-based DBN. As you can see, the accuracy of this system is 92.1% for all digits. Moreover, the highest top-1 accuracy of 100% is obtained for digits “0”, “1”, “6”, and “8” while at least 84.6% accuracy is achieved for other digits.

Figure 43 illustrates an accuracy comparison between 784 × 10 DBN-Fuzzy neural network and four different DBN topologies for various number of training samples. The results show that an accuracy of 48.8% for a 784×500×500×10 DBN trained by 500 training inputs can be increased to an 83.5% accuracy achieved using 784×10 DBN-Fuzzy which is trained by around 10,000 input training samples. Therefore, the recognition accuracy can be improved by employing the fuzzy system and increasing the number of training samples instead of the number of hidden layers in the network and number of nodes in each layer.

Figure 44 depicts the energy consumption of 784 × 10 DBN-Fuzzy neural network and four different DBN topologies while evaluating a single input image. As shown, a substantial amount

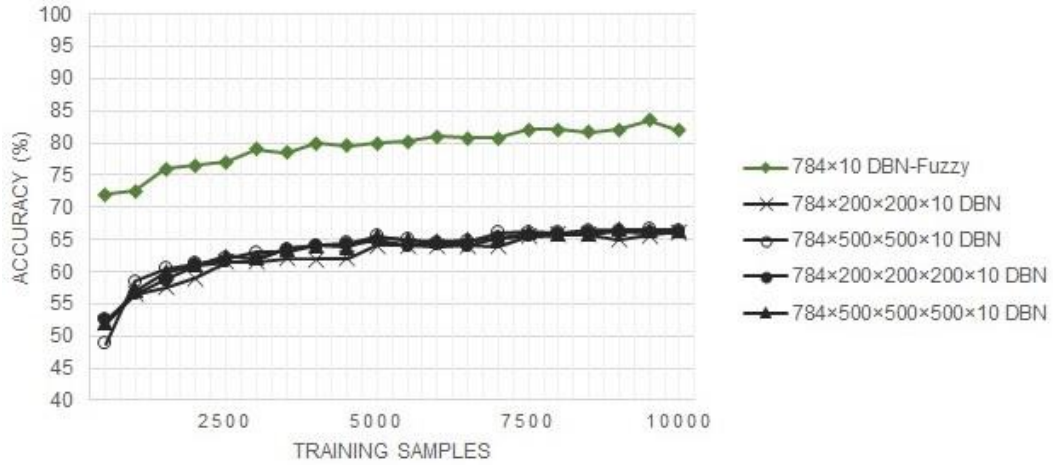


Figure 43: Accuracy of 784 × 10 DBN-Fuzzy neural network and four different DBN topologies for various training samples.

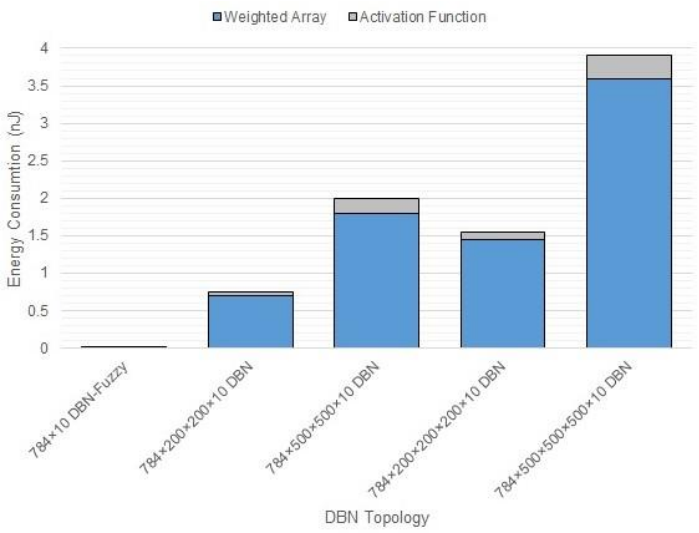


Figure 44: Energy Consumption for 784 × 10 DBN-Fuzzy neural network and four different DBN topologies.

of energy is consumed in the weighted connections, while less than 10% of the total energy is consumed in the neurons. As an illustration, the total energy consumption of a 784×200×200×200×10 DBN is almost equal to 1.55 nJ, only 0.1 nJ of which is dissipated in the neurons. Moreover, the energy consumption of around 3.9 nJ for a 784×500×500×500×10 DBN can be reduced to an 8.3 pJ energy consumption achieved utilizing 784×10 DBN-Fuzzy. This is achieved by employing the proposed TSK Rule-based Fuzzy System to improve the accuracy of MRAM-based DBNs.

Table 17 shows the normalized area consumptions for activation functions and weighted arrays for a 784 × 10 DBN-Fuzzy neural network and four different DBN topologies. Hereby, we have utilized the area consumption of the p-bit neuron as the baseline and relative to the p-bit area consumption, the estimated area values for activation functions and weighted arrays are normalized. According to the MRAM-based neuron's layout design, the MRAM-based neuron's area consumption is almost equal to 32λ×32λ, where λ = 14nm/2 = 7nm for 14nm FinFET

Table 17: Area of weighted array and activation function for 784 × 10 DBN-Fuzzy neural network and four different DBN topologies relative to the area occupied by a single p-bit-based neuron.

Topology	Normalized Area	
	Weighted Array	Activation Function
784×10 DBN-Fuzzy	3200×	10×
784×200×200×10 DBN	80000×	400000×
784×500×500×10 DBN	260000×	2500000×
784×200×200×200×10 DBN	96000×	80000000×
784×500×500×500×10 DBN	360000×	1250000000×

technology, thus resulting in the approximate area consumption of 0.05 μm^2 per neuron [12]. For each weight in the weighted array, the well-known 1T-1R structure is utilized. The resistive devices incurring no area overhead since these devices are fabricated on top of the MOS transistors. As a result, this structure allocates one transistor to each weight and the evaluated area consumption for each weight is around 0.02 μm^2 = 0.4X [12]. As it is shown in the table, the area consumption of a 784 × 10 DBN-Fuzzy neural network is significantly smaller than the four DBN topologies while 784 × 10 DBN-Fuzzy neural network has higher accuracy and lower energy consumption.

CHAPTER 7: CONCLUSION

7.1 SUMMARY

The concept of using sampling and count operations to interpret the probabilistic output of a p-bit based neuron offers an intriguing approach to realize a CMOS-based probabilistic interpolation recoder (PIR) for a spin-based stochastic binary neuron. Herein, we proposed a PIR circuit as a replacement for an analog-based approach to interpolate the output of the p-bit based activation functions in the last layer of a DBN circuit. The conventional method involved: first, using an RC circuit to continuously integrate the analog output of the p-bit, next an op-amp based sample and holder is used to sample the output of the RC circuit, finally the analog sampled output is converted to a digital value through an op-AMP based ADC circuit and a priority encoder. Our proposed CMOS-based PIR circuit removes the need for all of area- and energy-consuming analog components existing in conventional circuits such as resistors, capacitors, and opamps, and performs the interpolation operation only by using MOS-transistor based Boolean gates and flip-flops. In addition, the PIR circuits have an inherent single stuck-at fault tolerant features to tolerate either transient or permanent faults at the circuit's output without redundancy or active refurbishment overhead.

Moreover, we will two approaches to mitigate the effects of process variation on the energy barrier of the p-bit based neurons, and their consequent impact on the performance and accuracy of DBNs using p-bit devices as probabilistic sigmoidal neurons [260],[261],[262]. In the first approach, it was shown that an increase in the energy barrier leads to decreased fluctuation speed in the magnetization direction of the p-bit's nanomagnet. Thus, to observe the desired probabilistic sigmoidal behavior in the p-bit based neuron a temporal redundancy is required to

be added to the sampling time of the p-bits output to give it time to have sufficient probabilistic fluctuations. While the temporal redundancy has shown to be an efficient mechanism, it was examined that it can lead to approximately 10-fold higher energy consumption in a 784×200×10 DBN which can tolerate maximum 2 *kT* of energy barrier variations compared to a variation-less DBN. The second variation tolerance mechanism involved implementing p-bit with a negative self-feedback, which significantly increases the probabilistic fluctuation speed of the free layer. In this case, the drain of the NMOS transistor in the p-bit device tracks the magnetization direction of the free layer of the MTJ, and the inverter at the output of the device generates the inverse voltage, hence realizing a negative feedback effect which compensates the variation impacts with only ~10% energy consumption overheads.

Finally, we present an innovative image recognition technique for MNIST dataset on the basis of MRAM-based DBNs and TSK rule-based fuzzy systems. The proposed DBN-fuzzy system is introduced to benefit from low energy and area consumption of MRAM-based DBNs and high accuracy of TSK rule-based fuzzy systems. This system initially recognizes the top results through the MRAM-based DBN and then, the fuzzy system is employed to attain the top-1 recognition results from the obtained top outputs. We have shown that the top-1 accuracy of the DBN-fuzzy neural network is enhanced from 64% to 82% relative to the MRAM-based DBNs for a 784 × 10 network. Simulation results exhibit that a 784 × 10 DBN-Fuzzy neural network not only has lower energy and area consumption than bigger DBN topologies but also has higher accuracy. Neuro-fuzzy systems based on spintronic devices may offer a compact and computationally-efficient architectural approach to machine-based image recognition tasks.

7.2 FUTURE DIRECTIONS

The development of MRAM technology has focused primarily so far on stable binary digital memory applications. However, there is recent work that makes use of MRAM technology in a different type of application space, i.e. Neural Networks (NNs). Recent studies have shown that neurons (one of the two building blocks of an NN) can be built in MRAM cells using e.g. low-barrier magnetic tunnel junctions (MTJs). In particular, Ostwal et al. [263] demonstrated in 2019 for the first time a spin-orbit torque (SOT) based tunable random number generator (TRNG) using an unstable in-plane magnetic (IMA) MTJ stack. A charge current through the SOT material (here tantalum, marked blue in the layer stack) enables manipulation of the free magnetic layer (here CoFeB, marked red in the layer stack). While the device has the free magnetic layer fluctuating between its two magnetic states with a 50:50 probability without any input, a charge current can tune the probability to be either in the parallel (P) or antiparallel (AP) state. Combing this layout with an inverter allows for a neuron with gain that can be assembled into a larger neural network (NN). Note that to build the same tunable random number generator (TRNG) in complementary metal oxide semiconductor (CMOS) technology, about 1,000 transistors would be required, making the above hardware demonstration about 500-fold more compact. Realization of the second building block of NNs, i.e. synapses based on MRAM technology is thus highly desirable, since it would allow for an integrated – all MRAM based – approach towards a compact and highly power efficient neural network. In order to explore the potential of magnetic elements for synapses that consist of an ensemble of binary memory elements, Ostwal et al. [264] recently demonstrated a 4-bit compound synapse that used an array of 16 nanomagnets with perpendicular magnetic anisotropy (PMA) located in the crossbar region of a tantalum film that was used to excerpt spin-orbit torque.

Analog electronic non-volatile memories (eNVMs) have attracted attention in the research community for their potential as synaptic elements [265]. The conductance of such an eNVM can increase or decrease in a continuous analog fashion, mimicking the potentiation or depression of a synapse. However, while Resistive Random Access Memory (RRAM) technology has shown the potential for achieving such analog conductance behavior, the reliable fabrication of analog RRAM devices has remained challenging [266]. Hence, compound synapses that utilize an ensemble of fabricable binary memory elements are proposed. Employing the probabilistic switching of individual memory elements, multilevel operation can be realized in a reproducible fashion. In fact, experimental implementations based on an arrangement of parallel binary RRAM devices and simulations of Convolutional Neural Networks (CNNs) demonstrated multi-level operation of such compound synapse structures [267]. However, RRAM technology is facing challenges in terms of current and voltage scaling and is prone to process variability and instabilities. On the other hand, "MRAM has already found a niche market and is heading toward disruptive growth" according to Bhatti et al. [268]. Spin transfer torque (STT)-MRAM is close to foundry scale production [269] and wafer-scale manufacturability has been shown even for SOT-MRAM [271]. Moreover, we believe STT-MTJs to be less suitable for synapse applications, since the READ path and the WRITE path are identical and an unavoidable constantly changing resistance of the MTJs in series will make the device's switching behavior less controllable.

APPENDIX: COPYRIGHT PERMISSIONS

RightsLink®

Home | Help | Live Chat | Sign in | Create Account

IEEE Requesting permission to reuse content from an IEEE publication

Probabilistic Interpolation Recoder for Energy-Error-Product Efficient DBNs with p-bit Devices

Author: Hossein Pourmeidani

Publication: IEEE Transactions on Emerging Topics in Computing

Publisher: IEEE

Date: Dec 31, 1969

Copyright © 1969, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

© 2021 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Terms and Conditions
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

CCC RightsLink®

Home | Help | Live Chat | Sign in | Create Account

IEEE Requesting permission to reuse content from an IEEE publication

Modular Simulation Framework for Process Variation Analysis of MRAM-based Deep Belief Networks

Conference Proceedings: 2020 SoutheastCon

Author: Paul Wood

Publisher: IEEE

Date: 28 March 2020

Copyright © 2020, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

© 2021 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Terms and Conditions
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

CCC RightsLink®

Home | Help | Live Chat | Sign in | Create Account

IEEE Requesting permission to reuse content from an IEEE publication

Electrically-Tunable Stochasticity for Spin-based Neuromorphic Circuits: Self-Adjusting to Variation

Conference Proceedings:
2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)

Author: Hossein Pourmeidani

Publisher: IEEE

Date: Aug. 2020

Copyright © 2020, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

© 2021 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Terms and Conditions
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

CCC RightsLink®

Home | Help | Live Chat | Sign in | Create Account

IEEE Requesting permission to reuse content from an IEEE publication

Process Variation Sensitivity of Spin-Orbit Torque Perpendicular Nanomagnets in DBNs

Author: Hossein Pourmeidani
Publication: IEEE Transactions on Magnetics
Publisher: IEEE
Date: July 2021

Copyright © 2021, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

© 2021 Copyright - All Rights Reserved | Copyright Clearance Center, Inc. | Privacy statement | Terms and Conditions
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

LIST OF REFERENCES

[1] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[2] R. Sarikaya, G. E. Hinton, and A. Deoras, "Application of deep belief networks for natural language understanding," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 22, no. 4, pp. 778–784, 2014.

[3] S. K. Kim, P. L. McMahon, and K. Olukotun, "A large-scale architecture for restricted boltzmann machines," in Field-Programmable Custom Computing Machines (FCCM), 2010 18th IEEE Annual International Symposium on. IEEE, 2010, pp. 201–208.

[4] D. Le Ly and P. Chow, "High-performance reconfigurable hardware architecture for restricted boltzmann machines," IEEE Transactions on Neural Networks, vol. 21, no. 11, pp. 1780–1792, 2010.

[5] N. Lopes, B. Ribeiro, and J. Goncalves, "Restricted boltzmann machines and deep belief networks on multi-core processors," in Neural Networks (IJCNN), The 2012 International Joint Conference on. IEEE, 2012, pp. 1–7.

[6] M. N. Bojnordi and E. Ipek, "Memristive boltzmann machine: A hardware accelerator for combinatorial optimization and deep learning," in High Performance Computer Architecture (HPCA), 2016 IEEE International Symposium on. IEEE, 2016, pp. 1–13.

[7] A. M. Sheri, A. Rafique, W. Pedrycz, and M. Jeon, “Contrastive divergence for memristor-based restricted boltzmann machine,” Engineering Applications of Artificial Intelligence, vol. 37, pp. 336–342, 2015.

[8] S. B. Eryilmaz, E. Neftci, S. Joshi, S. Kim, M. BrightSky, H.-L. Lung, A. Lam, G. Cauwenberghs, and H.-S. P. Wong, “Training a probabilistic graphical model with resistive switching electronic synapses,” IEEE Transactions on Electron Devices, vol. 63, no. 12, pp. 5004–5011, 2016.

[9] B. Yuan and K. K. Parhi, “Vlsi architectures for the restricted boltzmann machine,” ACM Journal on Emerging Technologies in Computing Systems (JETC), vol. 13, no. 3, p. 35, 2017.

[10] A. Ardakani, F. Leduc-Primeau, N. Onizawa, T. Hanyu, and W. J. Gross, “Vlsi implementation of deep neural network using integral stochastic computing,” IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 25, no. 10, pp. 2688–2699, 2017.

[11] R. Zand, K. Y. Camsari, S. D. Pyle, I. Ahmed, C. H. Kim, and R. F. DeMara, “Low-energy deep belief networks using intrinsic sigmoidal spintronic-based probabilistic neurons,” in Proceedings of the 2018 on Great Lakes Symposium on VLSI, ser. GLSVLSI ’18, 2018, pp. 15–20.

[12] R. Zand, K. Y. Camsari, S. Datta, and R. F. Demara, “Composable probabilistic inference networks using mram-based stochastic neurons,” J. Emerg. Technol. Comput. Syst., vol. 15, no. 2, pp. 17:1–17:22, Mar. 2019. [Online]. Available: http://doi.acm.org/10.1145/3304105

[13] K. Y. Camsari, S. Salahuddin, and S. Datta, “Implementing p-bits with embedded mtj,” IEEE Electron Device Letters, vol. 38, no. 12, pp. 1767–1770, 2017.

[14] W. H. Choi, Y. Lv, H. Kim, J.-P. Wang, and C. H. Kim, “An 8-bit analogto digital converter based on the voltage-dependent switching probability of a magnetic tunnel junction,” in 2015 Symposium on VLSI Technology (VLSI Technology). IEEE, 2015, pp. T162–T163.

[15] Gregg, J.F.; Petej, I.; Jouguelet, E.; Dennis, C. Spin electronics a review. J. Phys. Appl. Phys. 2002, 35, R121–R155.

[16] Baibich, M.N.; Broto, J.M.; Fert, A.; Van Dau, F.N.; Petroff, F.; Etienne, P.; Creuzet, G.; Friederich, A.; Chazelas, J. Giant Magnetoresistance of (001)Fe/(001)Cr Magnetic Superlattices. Phys. Rev. Lett. 1988, 61, 2472–2475.

[17] Ennen, I.; Kappe, D.; Rempel, T.; Glenske, C.; Hütten, A. Giant Magnetoresistance: Basic Concepts, Microstructure, Magnetic Interactions and Applications. Sensors 2016, 16, 904.

[18] Maciel, N., Marques, E., Naviner, L., Zhou, Y., & Cai, H. (2020). Magnetic tunnel junction applications. Sensors, 20(1), 121.

[19] Peng, S.; Zhang, Y.;Wang, M.; Zhang, Y.; Zhao,W. Magnetic Tunnel Junctions for Spintronics: Principles and Applications; Wiley: Hoboken, NJ, USA, 2014; pp. 1–16.

[20] Chappert, C.; Fert, A.; Dau, F.N.V. The emergence of spin electronics in data storage. Nat. Mater. 2007, 6, 813–823.

[21] Qoutb, A.G.; Friedman, E.G. MTJ Magnetization Switching Mechanisms for IoT Applications. In Proceedings of the 2018 on Great Lakes Symposium on VLSI, Chicago, IL, USA, 23–25 May 2018; pp. 347–352.

[22] Cai, H.; Wang, Y.; Naviner, L.A.D.B.; Zhao, W. Robust Ultra-Low Power Non-Volatile Logic-in-Memory Circuits in FD-SOI Technology. IEEE Trans. Circuits Syst. Regul. Pap. 2017, 64, 847–857.

[23] Zarei, A.; Safaei, F. Power and area-efficient design of VCMA-MRAM based full-adder using approximate computing for IoT applications. Microelectron. J. 2018, 82, 62–70.

[24] Salehi, S.; Mashhadi, M.B.; Zaeemzadeh, A.; Rahnavard, N.; DeMara, R.F. Energy-Aware Adaptive Rate and Resolution Sampling of Spectrally Sparse Signals Leveraging VCMA-MTJ Devices. IEEE J. Emerg. Sel. Top. Circuits Syst. 2018, 8, 679–692.

[25] Chakraborty, I.; Agrawal, A.; Roy, K. Design of a Low-Voltage Analog-to-Digital Converter Using Voltage-Controlled Stochastic Switching of Low Barrier Nanomagnets. IEEE Magn. Lett. 2018, 9, 1–5.

[26] Marques, E.C.; Maciel, N.; Naviner, L.; Cai, H.; Yang, J. A Review of Sparse Recovery Algorithms. IEEE Access 2019, 7, 1300–1322.

[27] Zhang, D.; Zeng, L.; Zhang, Y.; Klein, J.O.; Zhao, W. Reliability-Enhanced Hybrid CMOS/MTJ Logic Circuit Architecture. IEEE Trans. Magn. 2017, 53, 1–5.

[28] Deng, E.; Kang, W.; Zhang, Y.; Klein, J.; Chappert, C.; Zhao, W. Design Optimization and Analysis of Multicontext STT-MTJ/CMOS Logic Circuits. IEEE Trans. Nanotechnol. 2015, 14, 169–177.

[29] Kang, W.; Deng, E.; Klein, J.; Zhang, Y.; Zhang, Y.; Chappert, C.; Ravelosona, D.; Zhao, W. Separated Precharge Sensing Amplifier for Deep Submicrometer MTJ/CMOS Hybrid Logic Circuits. IEEE Trans. Magn. 2014, 50, 1–5.

[30] Cai, H.; Wang, Y.; de Barros Naviner, L.A.; Yang, J.; Zhao, W. Exploring Hybrid STT-MTJ/CMOS Energy Solution in Near-/Sub-Threshold Regime for IoT Applications. IEEE Trans. Magn. 2018, 54, 1–9.

[31] Berger, L. Emission of spin waves by a magnetic multilayer traversed by a current. Phys. Rev. B 1996, 54, 9353–9358.

[32] Senni, S.; Torres, L.; Sassatelli, G.; Gamatie, A.; Mussard, B. Exploring MRAM Technologies for Energy Efficient Systems-On-Chip. IEEE J. Emerg. Sel. Top. Circuits Syst. 2016, 6, 279–292.

[33] Wang, Y.; Zhang, Y.; Deng, E.; Klein, J.O.; Naviner, L.A.B.; Zhao, W. Compact model of magnetic tunnel junction with stochastic spin transfer torque switching for reliability analyses. Microelectron. Reliab. 2014, 54, 1774–1778.

[34] Cai, H.; Wang, Y.; de Barros Naviner, L.A.; Zhao, W. Low Power Magnetic Flip-Flop Optimization With FDSOI Technology Boost. IEEE Trans. Magn. 2016, 52, 1–7.

[35] Kang,W.; Ran, Y.; Zhang, Y.; Lv,W.; Zhao,W. Modeling and Exploration of the Voltage-Controlled Magnetic Anisotropy Effect for the Next-Generation Low-Power and High-Speed MRAM Applications. IEEE Trans. Nanotechnol. 2017, 16, 387–395.

[36] Maruyama, T.; Shiota, Y.; Nozaki, T.; Ohta, K.; Toda, N.; Mizuguchi, M.; Tulapurkar, A.A.; Shinjo, T.; Shiraishi, M.; Mizukami, S.; et al. Large voltage-induced magnetic anisotropy change in a few atomic layers of iron. Nat. Nanotechnol. 2009, 4, 158–161.

[37] Wang, W.; Li, M.; Hageman, S.; Chien, C. Electric-field-assisted switching in magnetic tunnel junctions. Nat. Mater. 2012, 11, 64–68.

[38] Barnes, S.E.; Ieda, J.I.; Maekawa, S. Rashba spin-orbit anisotropy and the electric field control of magnetism. Sci. Rep. 2014, 4, 1–5.

[39] Velev, J.P.; Jaswal, S.S.; Tsymbal, E.Y. Multi-ferroic and magnetoelectric materials and interfaces. Philos. Trans. Math. Phys. Eng. Sci. 2011, 369, 3069–3097.

[40] Cai, H.; Wang, Y.; Kang, W.; Naviner, L.; Shan, W.; Yang, J.; Zhao, W. Enabling Resilient Voltage-Controlled MeRAM Using Write Assist Techniques. In Proceedings of the 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, Italy, 27–30 May 2018; pp. 1–5.

[41] Kang, W.; Chang, L.; Zhang, Y.; Zhao, W. Voltage-controlled MRAM for working memory: Perspectives and challenges. In Proceedings of the Design, Automation Test in Europe Conference Exhibition (DATE), Lausanne, Switzerland, 27–31 March 2017; pp. 542–547.

[42] Apalkov, D.; Dieny, B.; Slaughter, J.M. Magnetoresistive Random Access Memory. Proc. IEEE 2016, 104, 1796–1830.

[43] Wang, Z.; Li, Z.;Wang, M.;Wu, B.; Zhu, D.; Zhao,W. Field-free spin–orbit-torque switching of perpendicular magnetization aided by uniaxial shape anisotropy. Nanotechnology 2019, 30, 375202.

[44] Dieny, B.; Sousa, R.; Herault, J.; Papusoi, C.; Prenat, G.; Ebels, U.; Houssameddine, D.; Rodmacq, B.; Auffret, S.; Buda-Prejbeanu, L.; et al. Spin-Transfer Effect and its Use in Spintronic Components. Int. J. Nanotechnol. 2010, 7.

[45] Åkerman, J. Toward a Universal Memory. Science 2005, 308, 508–510.

[46] Hosomi, M.; Yamagishi, H.; Yamamoto, T.; Bessho, K.; Higo, Y.; Yamane, K.; Yamada, H.; Shoji, M.; Hachino, H.; Fukumoto, C.; et al. A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM. In Proceedings of the IEEE International Electron Devices Meeting, 2005, IEDM Technical Digest, Washington, DC, USA, 5–7 December 2005; pp. 459–462.

[47] Wang, Z.; Zhang, L.; Wang, M.; Wang, Z.; Zhu, D.; Zhang, Y.; Zhao, W. High-Density NAND-Like Spin Transfer Torque MemoryWith Spin Orbit Torque Erase Operation. IEEE Electron Device Lett. 2018, 39, 343–346.

[48] Lin, C.J.; Kang, S.H.; Wang, Y.J.; Lee, K.; Zhu, X.; Chen, W.C.; Li, X.; Hsu, W.N.; Kao, Y.C.; Liu, M.T.; et al. 45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell. In Proceedings of the 2009 IEEE International Electron Devices Meeting (IEDM), Baltimore, MD, USA, 7–9 December 2009; pp. 1–4.

[49] Cubukcu, M.; Boulle, O.; Drouard, M.; Garello, K.; Onur Avci, C.; Mihai Miron, I.; Langer, J.; Ocker, B.; Gambardella, P.; Gaudin, G. Spin-orbit torque magnetization switching of a three-terminal perpendicular magnetic tunnel junction. Appl. Phys. Lett. 2014.

[50] Fong, X.; Kim, Y.; Venkatesan, R.; Choday, S.H.; Raghunathan, A.; Roy, K. Spin-Transfer Torque Memories: Devices, Circuits, and Systems. Proc. IEEE 2016, 104, 1449–1488.

[51] Shiota, Y.; Nozaki, T.; Bonell, F.; Murakami, S.; Shinjo, T.; Suzuki, Y. Induction of coherent magnetization switching in a few atomic layers of FeCo using voltage pulses. Nat. Mater. 2012, 11, 39–43.

[52] Sharmin, S.; Jaiswal, A.; Roy, K. Modeling and design space exploration for bit-cells based on voltage-assisted switching of magnetic tunnel junctions. IEEE Trans. Electron Devices 2016, 63, 3493–3500.

[53] Wang, S.; Lee, H.; Ebrahimi, F.; Amiri, P.K.; Wang, K.L.; Gupta, P. Comparative Evaluation of Spin-Transfer-Torque and Magnetoelectric Random Access Memory. IEEE J. Emerg. Sel. Top. Circuits Syst. 2016, 6, 134–145.

[54] Zhang, H.; Kang, W.; Wang, Z.; Deng, E.; Zhang, Y.; Zhao, W. High-Density and Fast-Configuration Non-Volatile Look-Up Table Based on NAND-Like Spintronic Memory. In Proceedings of the 2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Chengdu, China, 26–30 October 2018; pp. 382–385.

[55] Tsou, Y.; Chiu, J.; Shih, H.; Liu, C.W. Write Margin Analysis of Spin-Orbit Torque Switching Using Field-Assisted Method. IEEE J. Explor.-Solid-State Comput. Devices Circuits 2019, 1.

[56] Garello, K.; Yasin, F.; Hody, H.; Couet, S.; Souriau, L.; Sharifi, S.H.; Swerts, J.; Carpenter, R.; Rao, S.; Kim, W.; et al. Manufacturable 300mm platform solution for Field-Free Switching SOT-MRAM. In Proceedings of the 2019 Symposium on VLSI Technology, Kyoto, Japan, 9–14 June, 2019; pp. T194–T195.

[57] Jan, G.; Thomas, L.; Le, S.; Lee, Y.; Liu, H.; Zhu, J.; Iwata-Harms, J.; Patel, S.; Tong, R.; Sundar, V.; et al. Demonstration of Ultra-Low Voltage and Ultra Low Power STT-MRAM designed for compatibility with 0x node embedded LLC applications. In Proceedings of the 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, USA, 18–22 June 2018; pp. 65–66.

[58] Garello, K.; Yasin, F.; Kar, G.S. Spin-Orbit Torque MRAM for ultrafast embedded memories: From fundamentals to large scale technology integration. In Proceedings of the 2019 IEEE 11th International MemoryWorkshop (IMW), Monterey, CA, USA, 12–15 May 2019; pp. 1–4.

[59] Li, X.; Lee, A.; Razavi, S.;Wu, H.;Wang, K. Voltage-controlled magnetoelectric memory and logic devices. MRS Bull. 2018, 43, 970–977.

[60] Deng, E. Design and Development of Low-Power and Reliable Logic Circuits Based on Spin-Transfer Torque Magnetic Tunnel Junctions. Ph.D. Thesis, Université Grenoble Alpes, Grenoble, France, 10 February 2017.

[61] Kuon, I.; Tessier, R.; Rose, J. FPGA Architecture: Survey and Challenges; Foundations and Trends in Electronic Design Automation Series; Foundations and Trends R: Delft, The Netherlands, 2008.

[62] Salehi, S.; Zand, R.; DeMara, R.F. Clockless Spin-based Look-Up Tables with Wide Read Margin. In Proceedings of the 2019 on Great Lakes Symposium on VLSI, Tysons Corner, VA, USA, 9–11 May 2019; pp. 363–366.

[63] Zand, R.; Roohi, A.; Salehi, S.; DeMara, R.F. Scalable Adaptive Spintronic Reconfigurable Logic Using Area-Matched MTJ Design. IEEE Trans. Circuits Syst. II Express Briefs 2016, 63, 678–682.

[64] Zhao, W.; Belhaire, E.; Chappert, C.; Dieny, B.; Prenat, G. TAS-MRAM-Based Low-Power High-Speed Runtime Reconfiguration (RTR) FPGA. ACM Trans. Reconfigurable Technol. Syst. 2009, 2, 8:1–8:19.

[65] Huang, K.; Ha, Y.; Zhao, R.; Kumar, A.; Lian, Y. A Low Active Leakage and High Reliability Phase Change Memory (PCM) Based Non-Volatile FPGA Storage Element. IEEE Trans. Circuits Syst. Regul. Pap. 2014, 61, 2605–2613.

[66] Zand, R.; DeMara, R.F. Radiation-hardened MRAM-based LUT for non-volatile FPGA soft error mitigation with multi-node upset tolerance. J. Phys. Appl. Phys. 2017, 50, 505002.

[67] Attaran, A.; Sheaves, T.D.; Mugula, P.K.; Mahmoodi, H. Static Design of Spin Transfer Torques Magnetic Look Up Tables for ASIC Designs. In Proceedings of the 2018 on Great Lakes Symposium on VLSI, Chicago, IL, USA, 23–25 May 2018; pp. 507–510.

[68] Zhao, W.; Belhaire, E.; Javerliac, V.; Chappert, C.; Dieny, B. Evaluation of a Non-Volatile FPGA based on MRAM technology. In Proceedings of the 2006 IEEE International Conference on IC Design and Technology, Padova, Italy, 1–4 May 2006; pp. 1–4.

[69] Zhao, W.; Belhaire, E.; Chappert, C. Spin-MTJ based Non-volatile Flip-Flop. In Proceedings of the 2007 7th IEEE Conference on Nanotechnology (IEEE NANO), Hong Kong, China, 2–5 August 2007; pp. 399–402.

[70] Montesi, L.; Zilic, Z.; Hanyu, T.; Suzuki, D. Building Blocks to Use in Innovative Non-volatile FPGA Architecture Based on MTJs. In Proceedings of the 2012 IEEE Computer Society Annual Symposium on VLSI, Amherst, MA, USA, 19–21 August 2012; pp. 302–307.

[71] Onizawa, N.; Hanyu, T. Redundant STT-MTJ-based nonvolatile flip-flops for low write-error-rate operations. In Proceedings of the 2016 14th IEEE International New Circuits and Systems Conference (NEWCAS), Vancouver, BC, Canada, 26–29 June 2016; pp. 1–4.

[72] Iyengar, A.S.; Ghosh, S.; Jang, J. MTJ-Based State Retentive Flip-Flop With Enhanced-Scan Capability to Sustain Sudden Power Failure. IEEE Trans. Circuits Syst. Regul. Pap. 2015, 62, 2062–2068.

[73] Meng, H.; Wang, J.G.; Wang, J.P. A spintronics full adder for magnetic CPU. IEEE Electron Device Lett. 2005, 26, 360–362.

[74] Matsunaga, S.; Hayakawa, J.; Ikeda, S.; Miura, K.; Hasegawa, H.; Endoh, T.; Ohno, H.; Hanyu, T. Fabrication of a Nonvolatile Full Adder Based on Logic-in-Memory Architecture Using Magnetic Tunnel Junctions. Appl. Phys. Express 2008, 1.

[75] Gang, Y.; Zhao, W.; Klein, J.; Chappert, C.; Mazoyer, P. A High-Reliability, Low-Power Magnetic Full Adder. IEEE Trans. Magn. 2011, 47, 4611–4616.

[76] B.N. Engel, N.D. Rizzo, J. Janesky, J.M. Slaughter, R. Dave, M. DeHerrera, M. Durlam, S. Tehrani, The science and technology of magnetoresistive tunneling memory, IEEE Tran. Nanotech. 99 (1) (2002) 32–38.

[77] A. Lyle, J. Harms, S. Patil, X. Yao, D.J. Lilija, J.-P. Wang, Direct communication between magnetic tunnel junctions for nonvolatile logic fan-out architecture, Appl. Phys. Lett. 97 (15) (2010) 152504–1-3.

[78] M.K. Ho, C.H. Tsang, R.E. Fontana Jr., S.S.P. Parkin, K.J. Carey, T. Pan, S. MacDonald, P.C. Arnett, J.O. Moore, Study of magnetic tunnel junction read sensors, IEEE Trans. Magn. 37 (4) (2001) 1691–1694.

[79] X. Zhu, J.-G. Zhu, Domain wall pinning and corresponding energy barrier in percolated perpendicular medium, IEEE Trans. Magn. 43 (6) (2007) 2349–2353.

[80] K. Kaisha Toshiba, Random number generator. US Patent, 2012/0026, 784, 2012-02-22.

[81] S.V. Dijken, C. Jiang, S.S. Parkin, Room temperature operation of a high output current magnetic tunnel transistor, Appl. Phys. Lett. 80 (18) (2002) 3364–3366.

[82] Y. Shuto, R. Nakane, W. Wang, H. Sukegawa, S. Yamamoto, M. Tanaka, K. Inomata, S. Sugahara, A new spin-functional metal–oxide–semiconductor field-effect transistor based on magnetic tunnel junction technology: pseudo-spin-MOSFET, Appl. Phys. Exp. 3 (1) (2010) 013003-1-3.

[83] P. Krzysteczko, G. Reiss, A. Thomas, Memristive switching of MgO based magnetic tunnel junctions, Appl. Phys. Lett. 95 (11) (2009) 112508–1-3.

[84] J.P. Singh, B. Kaur, S. Gautam, W.C. Lim, K. Asokan, K.H. Chae, Chemical Effects at the interfaces of Fe/MgO/Fe magnetic tunnel junction, Superlattices Microstruct. 100 (2016) 560–586.

[85] X.F. Han, S.S. Ali, S.H. Liang, MgO(001) barrier based magnetic tunnel junctions and their device applications, Sci. China Phys. Mech. Astron. 56 (1) (2013) 29–60.

[86] J.D.R. Buchanan, T.P.A. Hase, B.K. Tanner, N.D. Hughes, R.J. Hicken, Determination of the thickness of Al2O3 barriers in magnetic tunnel junctions, Appl. Phys. Lett. 81 (4) (2002) 751–753.

[87] M. Nakazumi, D. Yoshioka, H. Yanagihara, E. Kita, T. Koyano, Fabrication of magnetic tunneling junctions with NaCl barriers, Jpn. J. Appl. Phys. 46 (10A) (2007) 6618–6620.

[88] Z. Yang, Q. Zhan, X. Zhu, Y. Liu, H. Yang, B. Hu, J. Shang, L. Pan, B. Chen, R.-W. Li, Tunneling magnetoresistance induced by controllable formation of Co filaments in resistive switching Co/ZnO/ Fe structures, Europhys. Lett. 108 (5) (2014) 58004-p1-p6.

[89] D.A. Stewart, New type of magnetic tunnel junction based on spin filtering through a reduced symmetry oxide: FeCo|Mg3B2O6|FeCo, Nano Lett. 10 (1) (2010) 263–267.

[90] N.M. Caffrey, T. Archer, I. Rungger, S. Sanvito, Prediction of large bias-dependent magnetoresistance in all-oxide magnetic tunnel junctions with a ferroelectric barrier, Phys. Rev. B 83 (12) (2011) 125409-1-5.

[91] N. Kumar, P. Misra, R.K. Kotnala, A. Gaur, R.S. Katiyar, Room temperature magnetoresistance in Sr2FeMoO6/SrTiO3/Sr2FeMoO6 trilayer devices, J. Phys. D Appl. Phys. 47 (6) (2014) 065006 (5 pp.).

[92] A.V. Ramos, M.-J. Guittet, J.-B. Moussy, C. Gatel, Room temperature spin filtering in epitaxial cobalt-ferrite tunnel barriers, Appl. Phys. Lett. 91 (12) (2007) 122107-1-3.

[93] E. Cobas, A.L. Friedman, O.M.J. van't Erve, J.T. Robinson, B.T. Jonker, Graphene as a tunnel barrier: graphene-based magnetic tunnel junctions, Nano Lett. 12 (6) (2012) 3000–3004.

[94] W. Li, L. Xue, H.D. Abruna, D.C. Ralph, Magnetic tunnel junctions with single-layer-graphene tunnel barriers, Phys. Rev. B 89 (18) (2014) 184418-1-5.

[95] E.J. Tsymbol, O.N. Myrasov, P.R. LeClair, Spin dependent tunneling in magnetic tunnel junction, J. Phys. Condens. Matter. 15 (4) (2003) R109–R142.

[96] P.J. Chen, G. Feng, R.D. Shull, Use of half metallic heusler alloys in CoFeB/MgO/heusler alloy tunnel junctions, IEEE Trans. Magn. 49 (7) (2013) 4379–4382.

[97] T. Marukame, T. Kasahara, K. Matsuda, T. Uemura, M. Yamamoto, Fabrication of fully epitaxial magnetic tunnel junctions using full-Heusler alloy Co2Cr0.6Fe0.4Al thin film and MgO tunnel barrier, Jpn. J. Appl. Phys. 44 (17) (2005) L521–L524.

[98] T. Hatori, H. Ohmori, M. Tada, S. Nakagawa, MTJ elements with MgO barrier using RE-TM amorphous layers for perpendicular MRAM, IEEE Tran. Magn. 43 (6) (2007) 2331–2333.

[99] K. Zhang, Y.I. Cao, Y.W. Fang, Q. Li, J. Zhang, C. Duan, S.-S. Yan, Y. Tian, R. Huang, R.-K. Zheng, S.-S. Kang, Y.-X. Chen, G.-L. Liu, L-M. Mei, Electrical control of memristance and magnetoresistance in oxide magnetic tunnel junctions, Nanoscale 7 (14) (2015) 6334–6339.

[100] Te Velthuis Suzanne G. E., Liu Y., Freeland J. W., Zhernenkov M., Fitzsimmons M. R., Visani C., Bibes M., Barthélémy A., Cuellar F., Sefrioui Z., Leon C., Santamaria J. Magnetic behavior of complex oxide magnetic tunnel junctions. American Physical Society, APS March Meeting 2012, February 27-March 2, 2012, abstract #T9.009.

[101] Singh, J. P., Bhardwaj, R., Sharma, A., Kaur, B., Won, S. O., Gautam, S., & Chae, K. H. (2019). Fabrication of magnetic tunnel junctions. In Advanced Applications in Manufacturing Enginering (pp. 53-77). Woodhead Publishing.

[102] X. Chen, Y. Lei, Electrical conductivity measurement of ferromagnetic metallic materials using pulsed eddy current method, NDT & E Int. 75 (2015) 33–38.

[103] Kittel C. Introduction to Solid State Physics, 8th Edition (2005).

[104] C. Heiliger, P. Zahn, I. Mertig, Microscopic origin of magnetoresistance, Mater. Today 9 (11) (2006) 46–54.

[105] S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, K. Ando, Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions, Nat. Mater. 3 (12) (2004) 868–871.

[106] W.H. Butler, X.-G. Zhang, T.C. Schulthess, J.M. MacLaren, Spin-dependent tunneling conductance of Fe|MgO|Fe sandwiches, Phys. Rev. B. 63 (5) (2001) 054416–1-11.

[107] J. Mathon, A. Umerski, Theory of tunneling magnetoresistance of an epitaxial Fe/MgO/Fe(001) junction, Phys. Rev. B. 63 (22) (2001) 220403(R)-1-4.

[108] S. Ikeda, J. Hayakawa, Y. Ashizawa, Y.M. Lee, K. Miura, H. Hasegawa, M. Tsunoda, F. Matsukura, H. Ohno, Tunnel magnetoresistance of 604% at 300K by suppression of Ta diffusion in CoFeB/MgO/ CoFeB pseudo-spin-valves annealed at high temperature, Appl. Phys. Lett. 93 (8) (2008) 082508-1-3.

[109] G. Kim, Y. Sakuraba, M. Oogane, Y. Ando, T. Miyazaki, Tunneling magnetoresistance of magnetic tunnel junctions using perpendicular magnetization L10-CoPtL10-CoPt electrodes, Appl. Phys. Lett. 92 (17) (2008) 172502.

[110] H. Boeve, R.J.M. van de Veerdonk, B. Dutta, J. De Boeck, J.S. Moodera, G. Borghs, Area scaling of planar ferromagnetic tunnel junctions: from shadow evaporation to lithographic microfabrication, J. Appl. Phys. 83 (1998) 6700.

[111] S.A. Rishton, Y. Lu, R.A. Altman, A.C. Marley, X.P. Bian, C. Jahnes, R. Viswanathan, G. Xiao, W.J. Gallagher, S.S.P. Parkin, Magnetic tunnel junctions fabricated at tenth-micron dimensions by electron beam lithography, Microelectr. Eng. 35 (1–3) (1997) 249–252.

[112] K. Yakushiji, K. Noma, T. Saruya, H. Kubota, A. Fukushima, T. Nagahama, S. Yuasa, K. Ando, High magnetoresistance ratio and low resistance–area product in magnetic tunnel junctions with perpendicularly magnetized electrodes, Appl. Phys. Express 3 (5) (2010) 053003.

[113] Z.M. Zeng, P.K. Amiri, G. Rowlands, H. Zhao, I.N. Krivorotov, J.-P. Wang, J.A. Katine, J. Langer, K. Galatsis, K.L. Wang, H.W. Jiang, Effect of resistance-area product on spin-transfer switching in MgO-based magnetic tunnel junction memory cells, Appl. Phys. Lett. 98 (7) (2011) 072512.

[114] J.M. Daughton, Magnetic tunneling applied to memory, J. Appl. Phys. 81 (8) (1997) 3758–3763.

[115] H. Boeve, C. Bruynseraede, J. Das, K. Dessein, G. Borghs, J. De Boeck, R.C. Sousa, L.V. Melo, P.P. Freitas, Technology assessment for the implementation of magnetoresistive elements with semiconductor components in magnetic random access memory (MRAM) architectures, IEEE Trans. Magn. 35 (5) (1999) 2820–2825.

[116] R. Coehoorn, S.R. Cumpson, J.J.M. Ruigrok, P. Hidding, The electrical and magnetic response of yoke-type read heads based on a magnetic tunnel junction, IEEE Trans. Magn. 35 (5) (1999) 2586– 2588.

[117] J.J. Sun, V. Soares, P.P. Freitas, Low resistance spin-dependent tunnel junctions deposited with a vacuum break and radio frequency plasma oxidized, Appl. Phys. Lett. 74 (3) (1999) 448–450.

[118] K. Matsuda, A. Kamijo, T. Mitsuzuka, H. Tsuge, Exchange-biased magnetic tunnel junctions fabricated with in situ natural oxidation, J. Appl. Phys. 85 (8) (1999) 5261–5263.

[119] P.K. Wong, J.E. Evetts, M.G. Blamire, High conductance small area magnetoresistive tunnel junctions, J. Appl. Phys. 84 (3) (1998) 384–386.

[120] H. Boeve, J. De Boeck, G. Borghs, Low-resistance magnetic tunnel junctions by in situ natural oxidation, J. Appl. Phys. 89 (1) (2001) 482–487.

[121] H. Tsuge, E. Mitsuzuka, Magnetic tunnel junctions with in situ naturally-oxidized tunnel barrier, Appl. Phys. Lett. 71 (22) (1997) 3296–3298.

[122] J. O'Donnell, A.E. Andrus, S. Oh, E.V. Colla, J.N. Eckstein, Colossal magnetoresistance magnetic tunnel junctions grown by molecular-beam epitaxy, Appl. Phys. Lett. 76 (14) (2000) 1914–1916.

[123] G.X. Miao, J.Y. Chang, M.J. van Veenhuizen, K. Thiel, M. Seibt, G. Eilers, M. Münzenberg, J.S. Moodera, Epitaxial growth of MgO and Fe/MgO/Fe magnetic tunnel junctions on (100)Si by molecular beam epitaxy, Appl. Phys. Lett. 93 (14.) (2008) 142511-1-3.

[124] S. Yuasa, A. Fukushima, T. Nagahama, K. Ando, Y. Suzuki, High tunnel magnetoresistance at room temperature in fully epitaxial Fe/MgO/Fe tunnel junctions due to coherent spin-polarized tunneling, Jap. J. Appl. Phys. 43 (4B) (2004) L588–L590.

[125] C. Tusche, H.L. Meyerheim, N. Jedrecy, G. Renaud, A. Ernst, J. Henk, P. Bruno, J. Kirschner, Oxygen-induced symmetrization and sructural coherency in Fe/MgO/Fe(001) magnetic tunnel junctions, Phys. Rev. Lett. 95 (17) (2005) 176101/1-4.

[126] J.O. Hauch, M. Fonin, M. Fraune, P. Turban, R. Guerrero, F.G. Aliev, J. Mayer, U. Rüdiger, G. Güntherodt, Fully epitaxial Fe(110)/MgO(111)/Fe(110) magnetic tunnel junctions: growth, transport, and spin filtering properties, Appl. Phys. Lett. 93 (8.) (2008) 083512-1-3.

[127] S. Mitani, T. Moriyama, K. Takanashi, Fe/MgO/FeCo(100) epitaxial magnetic tunnel junctions prepared by using in situ plasma oxidation, J. App. Phys. 93 (10) (2003) 8041–8043.

[128] S. Andrieu, F. Bonell, T. Hauet, F. Montaigne, L. Calmels, E. Snoeck, P. Lefevr, F. Bertran, Magnetotransport in MgO-based magnetic tunnel junctions grown by molecular beam epitaxy, J. App. Phys. 115 (17) (2013) 172610-1-6.

[129] E. Popova, J. Faure-Vincent, C. Tiusan, C. Bellouard, H. Fischer, M. Hehn, F. Montaigne, M. Alnot, S. Andrieu, A. Schuhl, E. Snoeck, V. da Costa, Epitaxial MgO layer for low-resistance and coupling-free magnetic tunnel junctions, Appl. Phys. Letter. 81 (6) (2002) 1035–1037.

[130] Y.T. Takahashi, Y. Shiota, S. Miwa, F. Bonell, N. Mizuochi, T. Shinjo, Y. Suzuki, Fabrication of Fe/MgO/Gd magnetic tunnel junctions, IEEE Trans. Magn. 49 (7) (2013) 4417–4420.

[131] N. Tezuka, N. Ikeda, F. Mitsuhashi, S. Sugimoto, Improved tunnel magnetoresistance for magnetic tunnel junctions with heusler Co2FeAl05Si0.5 electrodes fabricated by molecular beam epitaxy system, J. Appl. Phys. 94 (16.) (2009) 162504-1-3.

[132] G.X. Miao, J.S. Moodera, Magnetic tunnel junctions with MgO-EuO composite tunnel barriers, Phys. Rev. B 85 (14) (2012) 144424-1-5.

[133] S. Yuasa, A. Fukushima, H. Kubota, Y. Suzuki, K. Ando, Giant tunneling magnetoresistance up to 410% at room temperature in fully epitaxial Co/MgO/Co magnetic tunnel junctions with bcc Co(001) electrodes, Appl. Phys. Lett. 89 (4) (2006) 042505-1-3.

[134] J.P. Singh, K. Asokan, D. Kabiraj, M. Raju, S. Chaudhary, S.R. Anhilash, D. Kanjilal, Magnetization in Fe/MgO/Fe based multilayers fabricated by e-beam evaporation method, AIP Conf. Proc. 1447 (1) (2012) 749–750.

[135] J.P. Singh, S. Gautam, B.B. Singh, S. Chaudhary, D. Kabiraj, D. Kanjilal, K.H. Chae, R. Kotnala, J.-M. Lee, J.M. Chen, K. Asokan, Magnetic, electronic structure and interface study of Fe/MgO/Fe multilayer, Adv. Mat. Lett. 5 (7) (2014) 372–377.

[136] V. Singh, S.R. Abhilash, B.R. Behera, D. Kabiraj, Fabrication of thin self-supporting platinum targets using evaporation techniques, Nucl. Inst. Meth. Phys. Res. A 635 (1) (2011) 20–23.

[137] S. Gautam, K. Asokan, J.P. Singh, Chang Fan-Hsiu, Lin Hong-Ji, K.H. Chae, Electronic structure of Fe/MgO/Fe multilayer stack by X-ray magnetic circular dichroism, J. Appl. Phys. 115 (17) (2014) 17C109-1-4.

[138] J.P. Singh, S. Gautam, W.C. Lim, K. Asokan, B.B. Singh, M. Raju, S. Chaudhary, D. Kabiraj, D. Kanjilal, J.-M. Lee, J.M. Chen, K.H. Chae, Electronic structure of magnetic Fe/MgO/Fe/Co multilayer structure by NEXAFS spectroscopy, Vacuum 138 (4) (2017) 48–54.

[139] J.P. Singh, W.C. Lim, S. Gautam, K. Asokan, K.H. Chae, Swift heavy ion irradiation induced effects in Fe/MgO/Fe/Co Multilayer, Mater. Design. 101 (2016) 72–79.

[140] J.P. Singh, W.C. Lim, K.H. Chae, Atomic diffusion processes in MgO/Fe/MgO multilayer, Superlattices Microstruct. 88 (2015) 609–619.

[141] Y.-C. Weng, C.W. Cheng, G. Chern, Interlayer exchange coupling and perpendicular magnetic anisotropy in CoFeB/MgO/CoFeB tunnel junction structures, IEEE Trans. Magn. 49 (7) (2013) 4425–4428.

[142] C.C. Tsai, C.-W. Cheng, Y.-C. Weng, G. Chern, The dipolar interaction in CoFeB/MgO/CoFeB perpendicular magnetic tunnel junction, J. Appl. Phys. 115 (17) (2014) 17C720-1-3.

[143] S.S.P. Parkin, C. Kaiser, A. Panchula, P.M. Rice, B. Hughes, M. Samant, S.H. Yang, Giant tunnelling magnetoresistance at room temperature with MgO (100) tunnel barriers, Nat. Mater. 3 (12) (2004) 862–867.

[144] D.D. Djayaprawira, K. Tsunekawa, M. Nagai, H. Maehara, S. Yamagata, N. Watanabe, S. Yuasa, Y. Suzuki, K. Ando, 230% room-temperature magnetoresistance in CoFeB/MgO/CoFeB magnetic tunnel junction, Appl. Phys. Lett. 86 (2005) 092502-1-3.

[145] J. Hayakawa, S. Ikeda, Y.M. Lee, F. Matsukuraand, H. Ohno, Effect of high annealing temperature on giant tunnel magnetoresistance ratio of CoFeB/MgO/CoFeB magnetic tunnel junctions, Appl. Phys. Lett. 89 (2006) 232510-1-3.

[146] C.-Y. Yang, S.-J. Chang, M.-H. Lee, K.H. Shen, S.-Y. Yang, H.-J. Lin, Y.-C. Tseng, Competing anisotropy-tunneling correlation of the CoFeB/MgO perpendicular magnetic tunnel junction: an electronic approach, Sci. Rep. 5 (2015) 17169.

[147] S. Ikeda, J. Hayakawa, Y.M. Lee, R. Sasaki, T. Meguro, F. Matsukura, H. Ohno, Dependence of tunnel magnetoresistance in MgO based magnetic tunnel junctions on Ar pressure during MgO sputtering, Jpn. J. Appl. Phys. 44 (48.) (2005) L1442-L1445.

[148] W. Shen, B.D. Schrag, A. Girdhar, M.J. Carter, H. Sang, G. Xiao, Effects of superparamagnetism in MgO based magnetic tunnel junctions, Phy. Rev. B. 79 (1) (2009) 014418-1-3.

[149] S. Ikeda, K. Miura, H. Yamamoto, K. Mizunuma, H.D. Gan, M. Endo, S. Kanai, J. Hayakawa, F. Matsukura, H. Ohno, A perpendicular-anisotropy CoFeB–MgO magnetic tunnel junction, Nat. Mater. 9 (9) (2010) 721–724.

[150] S. Cardoso, R.J. Macedo, R. Ferreira, A. Augusto, P. Wisniowski, P.P. Freitas, Ion beam assisted deposition of MgO barriers for magnetic tunnel junctions, J. Appl. Phys. 103 (7) (2008) 07A905-1-3.

[151] S. Wang, C. Antonakos, C. Bordel, D.S. Bouma, P. Fischer, F. Hellman, Ultrathin IBAD, MgO films for epitaxial growth on amorphous substrates and sub-50 nm membranes, Appl. Phys. Lett. 109 (19) (2016) 191603-1-3.

[152] B.B. Singh, S. Chaudhary, Effect of MgO spacer and annealing on interface and magnetic properties of ion beam sputtered NiFe/Mg/MgO/CoFe layer structures, J. Appl. Phys. 112 (6) (2012) 063906-1-7.

[153] B.B. Singh, S. Chaudhary, Effect of annealing on the temperature dependence of inelastic tunneling contributions vis-à-vis tunneling magnetoresistance and barrier parameters in CoFe/MgO/NiFe magnetic tunnel junctions, J. Appl. Phys. 115 (8) (2014) 083904-1-9.

[154] B.B. Singh, S. Chaudhary, Study of angular dependence of exchange bias and misalignment in uniaxial and unidirectional anisotropy in NiFe(111)/FeMn(111)/CoFeB(amorphous) stack, J. Magn. Magn. Mater. 385 (2015) 166–174.

[155] A. Zaleski, W. Skowronski, M. Czapkiewicz, J. Kanak, T. Stobiecki, R. Macedo, S. Cardoso, P.P. Freitas, Reduction of critical current in magnetic tunnel junctions with CoFeB/Ru/CoFeB synthetic free layer, J. Phys. Conf. Ser. 200 (2010) 052035.

[156] J. Orna, L. Morellon, P.A. Algarabel, J.A. Pardo, S. Sangiao, C. Magen, E. Snoeck, J.M. De Teresa, M.R. Ibarra, FeO/MgO/Fe heteroepitaxial structures for magnetic tunnel junctions, IEEE Trans. Magn. 44 (11) (2008) 2862–2864.

[157] A.M. Sánchez, L. Äkäslompolo, Q.H. Qin, S. van Dijken, Towards all-oxide magnetic tunnel junctions: epitaxial growth of SrRuO3/CoFe2O4/La2/3Sr1/3MnO3 trilayers, Cryst. Growth Des. 12 (2) (2012) 954–959.

[158] X. Liu, J. Shi, Magnetic tunnel junctions with Al2O3 tunnel barriers prepared by atomic layer deposition, Appl. Phys. Lett. 102 (20.) (2013) 202401-1-3.

[159] M.-B. Martin, B. Dlubak, R.S. Weatherup, H. Yang, C. Deranlot, K. Bouzehouane, F. Petroff, A. Anane, S. Hofmann, J. Robertson, A. Fert, P. Seneor, Sub-nanometer atomic layer deposition for spintronics in magnetic tunnel junctions based on graphene spin-filtering membranes, ACS Nano 8 (8) (2014) 7890–7895.

[160] R. Mantovan, S. Vangelista, B. Kutrzeba-Kotowska, A. Lamperti, N. Manca, L. Pellegrino, M. Fanciulli, Fe3−δO4/MgO/Co magnetic tunnel junctions synthesized by full in situ atomic layer and chemical vapour deposition, J. Phys. D Appl. Phys. 47 (10) (2014) 102002-1-4.

[161] S.-H. Han, W.-C. Jeong, J.-S. Lee, B.D. Kim, S.-K. Joo, Formation of tunnel barrier using a pseudo-atomic layer deposition method and its application to spin-dependent tunneling junction, Appl. Phys. A 81 (3) (2005) 611–615.

[162] L.F. Thompson, An introduction to lithography, Introduction to Microlithography, American Chemical Society, Washington, DC, (1983).

[163] M. Razeghi, Fundamentals of Solid State Engineering, 2nd ed., Springer Science + Business Media, (2006).

[164] C.M. Garner, Lithography for enabling advances in integrated circuits and devices, Phil. Trans. R. Soc. A 370 (1973) (2012) 4015–4041.

[165] V. AZlobin, Electron beam technology for power semiconductor device fabrication, 1993 Fifth European Conference on Power Electronics and Applications, 1993Brighton, UK. vol. 2.

[166] C.S. Wu, Y. Makiuchi, CD Chen, High-energy electron beam lithography for nanoscale fabrication, in: M. Wang (Ed.), Lithography, INTECH, Croatia, 2010, pp. 656 ISBN 978-953-307-064-3.

[167] M. Itano, F.W. Kern, M. Miyashita, T. Ohmi, Particle removal from silicon wafer surface in wet cleaning process, IEEE Trans. Semicond. Manuf. 6 (3) (1993) 258.

[168] V.M. Martinez, T.F. Edgar, Control of lithography in semiconductor manufacturing, IEEE Control Syst. 26 (6) (2006) 46–55.

[169] J. Srlund, M. Ritala, M. Leskela, E. Siponmaa, Zilliacus, Characterization of etching procedure in preparation of CdTe solar cells, Sol. Energy Mater Sol. Cells 44 (2) (1996) 177–190.

[170] C. Cardinaud, M.-C. Peignon, P.-Y. Tessier, Plasma etching: principles, mechanisms, application to micro- and nano-technologies, Appl. Surf. Sci. 164 (1–4) (2000) 72–83.

[171] H. Jansen, H. Gardeniers, M. de Boer, M. Elwenspoek, J. Fluitman, A survey on the reactive ion etching of silicon in microtechnology, J. Micromech. Microeng. 6 (1) (1996) 14–28.

[172] P.F.A. Alkemade, E. van Veldhoven, Deposition, milling, and etching with a focused helium ion beam, in: M. Stepanova, S. Dew (Eds.), Nanofabrication, Springer-Verlag, Wien, 2012, pp. 275–300.

[173] T. Miyazaki, S. Kumagai, T. Yaoi, Spin tunneling in Ni–Fe/Al2O3/Co junction devices, J. Appl. Phys. 81 (8) (1997) 3753–3757.

[174] X.F. Han, T. Daibou, M. Kamijo, K. Yaoita, K. Kubota, Y. Ando, T. Miyazaki, High-magnetoresistance tunnel junctions using Co75Fe25 ferromagnetic electrodes, Jpn. J. Appl. Phys. 39 (5B) (2000) L 439–L 441.

[175] G.X. Miao, Y.J. Park, J.S. Moodera, M. Seibt, G. Eilers, M. Münzenberg, Disturbance of tunneling coherence by oxygen vacancy in epitaxial Fe/MgO/Fe magnetic tunnel junctions, Phys. Rev. Lett. 100 (24) (2008) 246803.

[176] J.-Y. Chen, Y.-C. Lau, J.M.D. Coey, M. Li, J.-P. Wang, High performance MgO-barrier magnetic tunnel junctions for flexible and wearable spintronic applications, Sci. Rep. 7 (2017) 42001.

[177] J.S. Moodera, L.R. Kinder, T.M. Wong, R. Meservey, Large magnetoresistance at room temperature in ferromagnetic thin film tunnel junctions, Phys. Rev. Lett. 74 (16) (1995) 3273–3276.

[178] J.S. Moodera, G. Mathon, Spin polarized tunneling in ferromagnetic junctions, J. Magn. Magn. Mater. 200 (1–3) (1999) 248–273.

[179] M. Julloere, Tunneling between ferromagnetic films, Phys. Lett. 54 (3) (1975) 225–226.

[180] P. LeClair, J.S. Moodera, R. Meservey, Ferromagnetic-ferromagnetic tunneling and the spin filter effect, J. Appl. Phys. 76 (10) (1994) 6546–6548.

[181] Y. Ootuka, K. Ono, H. Shimada, S.-H. Kobayashi, A new fabrication method of ultra-small tunnel junctions, Physica B 227 (1–4) (1996) 307–309.

[182] C. Barraud, C. Deranlot, P. Seneor, R. Mattana, B. Dlubak, S. Fusil, K. Bouzehouane, D. Deneuve, F. Petroff, A. Fert, Magnetoresistance in magnetic tunnel junctions grown on flexible organic substrates, Appl. Phys. Lett. 96 (7) (2010) 072502-1-3.

[183] A. Bedoya-Pinto, M. Donolato, Gobbi Marco, L.E. Hueso, P. Vavassori, Flexible spintronic devices on Kapton, Appl. Phys. Lett. 104 (6) (2014) 062412-1-3.

[184] A. Persson, G. Thorne, H. Nguyen, Rapid prototyping of magnetic tunnel junctions with focused ion beam processes, J. Micromech. Microeng. 20 (5) (2010) 055039.

[185] J.R. Maldonado, M. Peckerar, X-ray lithography: some history, current status and future prospects, Microelectron. Eng. 161 (2016) 87–93.

[186] S.G. Keens, B. Rossa, M. Frei, Free-electron lasers for 13nm EUV lithography: RF design strategies to minimize investment and operational costs. Proceedings of SPIE 9776, Extreme Ultraviolet (EUV) Lithography VI. 97760T. 2016.

[187] A. Joshi-Imre, S. Bauerdick, Direct-write ion beam lithography, J. Nanotechnol. 2014 (2014) 170415 26 pages.

[188] Y. Yamakoshi, N. Atoda, K. Shimizu, T. Sato, Y. Shimizu, X-ray lithography system: analysis and an optimum construction, Appl. Opt. 25 (6) (1986) 922–927.

[189] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines," Cognitive science, vol. 9, no. 1, 1985.

[190] M. A. Carreira-Perpinan and G. E. Hinton, "On contrastive divergence learning." in Aistats, vol. 10, 2005, pp. 33–40.

[191] W. H. Choi, Y. Lv, J. Kim, A. Deshpande, G. Kang, J.-P. Wang, and C. H. Kim, "A magnetic tunnel junction based true random number generator with conditional perturb and real-time output probability tracking," in Electron Devices Meeting (IEDM), 2014 IEEE International. IEEE, 2014, pp. 12–5.

[192] B. Sutton, K. Y. Camsari, B. Behin-Aein, and S. Datta, "Intrinsic optimization using stochastic nanomagnets," Scientific reports, vol. 7, p. 44370, 2017.

[193] P. Debashis, R. Faria, K. Y. Camsari, J. Appenzeller, S. Datta, and Z. Chen, "Experimental demonstration of nanomagnet networks as hardware for ising computing," in Electron Devices Meeting (IEDM), 2016 IEEE International. IEEE, 2016, pp. 34–3.

[194] K. Y. Camsari, R. Faria, B. M. Sutton, and S. Datta, "Stochastic p-bits for invertible logic," Physical Review X, vol. 7, no. 3, p. 031014, 2017

[195] A. Sengupta, P. Panda, P. Wijesinghe, Y. Kim, and K. Roy, "Magnetic tunnel junction mimics stochastic cortical spiking neurons," Scientific reports, vol. 6, p. 30039, 2016.

[196] A. Sengupta, M. Parsa, B. Han, and K. Roy, "Probabilistic deep spiking neural systems enabled by magnetic tunnel junction," IEEE Transactions on Electron Devices, vol. 63, no. 7, pp. 2963–2970, July 2016.

[197] B. Behin-Aein, V. Diep, and S. Datta, "A building block for hardware belief networks," Scientific reports, vol. 6, 2016.

[198] V. Ostwal, P. Debashis, R. Faria, Z. Chen, and J. Appenzeller, "Spintorque devices with hard axis initialization as stochastic binary neurons," Scientific reports, vol. 8, no. 1, p. 16689, 2018.

[199] A. Sengupta, A. Banerjee, and K. Roy, "Hybrid spintronic-cmos spiking neural network with on-chip learning: Devices, circuits, and systems," Phys. Rev. Applied, vol. 6, p. 064003, Dec 2016.

[200] S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, and K. Ando, "Giant room-temperature magnetoresistance in single-crystal fe/mgo/fe magnetic tunnel junctions," Nature materials, vol. 3, no. 12, p. 868, 2004.

[201] M. Tanaka and M. Okutomi, "A novel inference of a restricted boltzmann machine," in 2014 22nd International Conference on Pattern Recognition, Aug 2014, pp. 1526–1531.

[202] J.-s. Seo, B. Brezzo, Y. Liu, B. D. Parker, S. K. Esser, R. K. Montoye, A. Rajendran, J. A. Tierno, L. Chang, D. S. Modha et al., "A 45nm cmos neuromorphic chip with a scalable architecture for learning in networks of spiking neurons," in 2011 IEEE Custom Integrated Circuits Conference (CICC). IEEE, 2011, pp. 1–4.

[203] Y. Kim, Y. Zhang, and P. Li, “A reconfigurable digital neuromorphic processor with memristive synaptic crossbar for cognitive computing,” ACM Journal on Emerging Technologies in Computing Systems (JETC), vol. 11, no. 4, p. 38, 2015.

[204] Y. Wang, T. Tang, L. Xia, B. Li, P. Gu, H. Yang, H. Li, and Y. Xie, “Energy efficient rram spiking neural network for real time classification,” in Proceedings of the 25th edition on Great Lakes Symposium on VLSI. ACM, 2015, pp. 189–194.

[205] P. Wijesinghe, A. Ankit, A. Sengupta, and K. Roy, “An all-memristor deep spiking neural computing system: A step toward realizing the low-power stochastic brain,” IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, no. 5, pp. 345–358, 2018.

[206] A. Sengupta, A. Ankit, and K. Roy, “Performance analysis and benchmarking of all-spin spiking neural networks (special session paper),” in 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017, pp. 4557–4563.

[207] R. P. Cowburn, D. K. Koltsov, A. O. Adeyeye, M. E. Welland, and D. M. Tricker, “Single-domain circular nanomagnets,” Phys. Rev. Lett., vol. 83, pp. 1042–1045, Aug 1999. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.83.1042

[208] P. Debashis, R. Faria, K. Y. Camsari, and Z. Chen, “Design of stochastic nanomagnets for probabilistic spin logic,” IEEE Magnetics Letters, vol. 9, pp. 1–5, 2018.

[209] J. C. Sankey, Y.-T. Cui, J. Z. Sun, J. C. Slonczewski, R. A. Buhrman, and D. C. Ralph, “Measurement of the spin-transfer-torque vector in magnetic tunnel junctions,” Nature Physics, vol. 4, no. 1, p. 67, 2008.

[210] C. Lin, S. Kang, Y. Wang, K. Lee, X. Zhu, W. Chen, X. Li, W. Hsu, Y. Kao, M. Liu et al., “45nm low power cmos logic compatible embedded stt mram utilizing a reverse-connection 1t/1mtj cell,” in 2009 IEEE International Electron Devices Meeting (IEDM). IEEE, 2009, pp. 1–4.

[211] N. Gougol, “Cmos operational amplifier design,” University of California at Berkeley Technical Report No. UCB/EECS-2016-223, December 31, 2016.

[212] A. Stillmaker and B. Baas, “Scaling equations for the accurate prediction of cmos device performance from 180 nm to 7 nm,” Integration, vol. 58, pp. 74–81, 2017.

[213] M. Habibi and H. Pourmeidani, “A hierarchical defect repair approach for hybrid nano/cmos memory reliability enhancement,” Microelectronics Reliability, vol. 54, no. 2, pp. 475–484, 2014.

[214] H. Pourmeidani, S. Sheikhfaal, R. Zand, and R. F. DeMara, "Probabilistic Interpolation Recoder for Energy-Error-Product Efficient DBNs with p-bit Devices," IEEE Transactions on Emerging Topics in Computing, 2020.

[215] J. L. Drobitch and S. Bandyopadhyay, “Reliability and Scalability of p-Bits Implemented With Low Energy Barrier Nanomagnets,” IEEE Magnetics Letters, vol. 10, pp. 1–4, 2019.

[216] M. A. Abeed and S. Bandyopadhyay, “Sensitivity of the power spectra of thermal magnetization fluctuatuions in low barrier nanomagnets proposed for stochadtic computing to in-plane barrier height variations and structural defects,” SPIN, 2050001, World Scientific Publishing Company, 2019.

[217] M. A. Abeed and S. Bandyopadhyay. “Low energy barrier nanomagnet design for binary stochastic neurons: Design challenges for real nanomagnets with fabrication defects”, IEEE Magnetics Letters, vol. 10, 4504405, 2019. [Online] Available: DOI: 10.1109/LMAG.2019.2929484.

[218] J. Kaiser, A. Rustagi, K. Y. Camsari, J. Z. Sun, S. Datta, and P. Upadhyaya, “Subnanosecond Fluctuations in Low-Barrier Nanomagnets,” Phys. Rev. Applied 12, 054056, 2019.

[219] O. Hassan, R. Faria, K. Y. Camsari, J. Z. Sun, and S. Datta, “Low barrier magnet design for efficient hardware binary stochastic neurons,” IEEE Magnetics Lett., vol. 10, 4502805, 2019. [Online]. Available: doi: 10.1109/LMAG.2019.2910787.

[220] K. Y. Camsari, B. M. Sutton, and S. Datta, “P-bits for probabilistic spin logic,” Appl. Phys. Rev., vol. 6, no. 1, 2019.

[221] J. L. Drobitch and S. Bandyopadhyay, “Spin electronics robustness and scalability of p-bits implemented with low energy barrier nanomagnets,” pp. 1–5, 2019.

[222] Debashis, Punyashloka, Rafatul Faria, Kerem Y. Camsari, Supriyo Datta, and Zhihong Chen. "Correlated fluctuations in spin orbit torque coupled perpendicular nanomagnets." Physical Review B 101, no. 9 (2020): 094405.

[223] Vodenicarevic, Damir, Nicolas Locatelli, Alice Mizrahi, Joseph S. Friedman, Adrien F. Vincent, Miguel Romera, Akio Fukushima et al. "Low-energy truly random number generation with superparamagnetic tunnel junctions for unconventional computing." Physical Review Applied 8, no. 5 (2017): 054045.

[224] Debashis, Punyashloka, and Zhihong Chen. "Tunable Random Number Generation Using Single Superparamagnet with Perpendicular Magnetic Anisotropy." In 2018 76th Device Research Conference (DRC), pp. 1-2. IEEE, 2018.

[225] Lopez-Diaz, L., L. Torres, and E. Moro. "Transition from ferromagnetism to superparamagnetism on the nanosecond time scale." Physical Review B 65, no. 22 (2002): 224406.

[226] Debashis, Punyashloka, and Zhihong Chen. "Electrical annealing and stochastic resonance in low barrier perpendicular nanomagnets for oscillatory neural networks." In 2019 77th Device Research Conference (DRC), pp. 85-86. IEEE, 2019.

[227] K. Garello, F. Yasin, S. Couet, L. Souriau, J. Swerts, S. Rao, S. Van Beek, W. Kim, E. Liu, S. Kundu et al., “SOT-MRAM 300mm integration for low power and ultrafast embedded memories,” in 2018 IEEE Symposium on VLSI Circuits. IEEE, 2018, pp. 81–82.

[228] T. Devolder, J. Hayakawa, K. Ito, H. Takahashi, S. Ikeda, P. Crozat, N. Zerounian, J.-V. Kim, C. Chappert, and H. Ohno, “Single-shot time-resolved measurements of nanosecond-scale spin-transfer induced switching: Stochastic versus deterministic aspects,” Physical review letters, vol. 100, no. 5, p. 057206, 2008.

[229] I. M. Miron, K. Garello, G. Gaudin, P.-J. Zermatten, M. V. Costache, S. Auffret, S. Bandiera, B. Rodmacq, A. Schuhl, and P. Gambardella, “Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection,” Nature, vol. 476, no.

7359, pp. 189–193, 2011.

[230] L. Liu, C.-F. Pai, Y. Li, H. Tseng, D. Ralph, and R. Buhrman, “Spin-torque switching with the giant spin Hall effect of tantalum,” Science, vol. 336, no. 6081, pp. 555–558, 2012.

[231] S. Ikeda, K. Miura, H. Yamamoto, K. Mizunuma, H. Gan, M. Endo, S. Kanai, J. Hayakawa, F. Matsukura, and H. Ohno, “A perpendicular-anisotropy CoFeB–MgO magnetic tunnel junction,” Nature materials, vol. 9, no. 9, pp. 721–724, 2010.

[232] S. Miura, K. Nishioka, H. Naganuma, T. A. Nguyen, H. Honjo, S. Ikeda, T. Watanabe, H. Inoue, M. Niwa, T. Tanigawa et al., “Scalability of quad interface p-MTJ for 1x nm STT-MRAM with 10-ns low power write operation, 10 years retention and endurance > 1011,” IEEE Transactions on Electron Devices, vol. 67, no. 12, pp. 5368–5373, 2020.

[233] Z. Li and S. Zhang, “Thermally assisted magnetization reversal in the presence of a spin-transfer torque,” Physical Review B, vol. 69, no. 13, p. 134416, 2004.

[234] W. Rippard, R. Heindl, M. Pufall, S. Russek, and A. Kos, “Thermal relaxation rates of magnetic nanoparticles in the presence of magnetic fields and spin-transfer effects,” Physical Review B, vol. 84, no. 6, p. 064439, 2011.

[235] Babuška, Robert. "Fuzzy systems, modeling and identification." Delft University of Technology, Department of Electrical Engineering Control Laboratory, Mekelweg 4

(1996).

[236] Zadeh, L.A. (1973). Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Systems, Man, and Cybernetics 1, 28–44.

[237] Mamdani, E.H. (1977). Application of fuzzy logic to approximate reasoning using linguistic systems. Fuzzy Sets and Systems 26, 1182–1191.

[238] Driankov, D., H. Hellendoorn and M. Reinfrank (1993). An Introduction to Fuzzy Control. Springer, Berlin.

[239] Friedman, J.H. (1991). Multivariate adaptive regression splines. The Annals of Statistics 19(1), 1–141.

[240] Jang, J.-S.R. and C.-T. Sun (1993). Functional equivalence between radial basis function networks and fuzzy inference systems. IEEE Transactions on Neural Networks 4(1), 156–159.

[241] Brown, M. and C. Harris (1994). Neurofuzzy Adaptive Modelling and Control. New York: Prentice Hall.

[242] Takagi, T. and Sugeno, M., 1985. Fuzzy identification of systems and its applications to modeling and control. IEEE transactions on systems, man, and cybernetics, (1), pp.116-132.

[243] Tanaka, K. and M. Sugeno (1992). Stability analysis and design of fuzzy control systems. Fuzzy Sets and Systems 45(2), 135–156.

[244] Zhao, J. (1995). Fuzzy logic in modeling and control. PhD dissertation,

CESAME, Louvain la Neuve, Belgium.

[245] Tanaka, K., T. Ikeda and H.O. Wang (1996). Robust stabilization of a class of uncertain nonlinear systems via fuzzy control: Quadratic stability, H1 control theory and linear matrix inequalities. IEEE Transactions on Fuzzy Systems 4(1), 1–13.

[246] Filev, D.P. (1996). Model based fuzzy control. In Proceedings Fourth European Congress on Intelligent Techniques and Soft Computing EUFIT'96, Aachen, Germany.

[247] Wang, L.-X. (1992). Fuzzy systems are universal approximators. In Proc. IEEE Int. Conf. on Fuzzy Systems 1992, San Diego, USA, pp. 1163–1170.

[248] Leonaritis, I.J. and S.A. Billings (1985). Input-output parametric models for non-linear systems. International Journal of Control 41, 303–344.

[249] Zimmermann, H.-J. (1987). Fuzzy Sets, Decision Making and Expert Systems. Boston: Kluwer Academic Publishers.

[250] Jang, J.-S.R. (1993). ANFIS: Adaptive-network-based fuzzy inference systems. IEEE Transactions on Systems, Man & Cybernetics 23(3), 665–685.

[251] Pedrycz, W. (1995). *Fuzzy Sets Engineering*. Boca Raton, Fl.: CRC Press.

[252] Strang, G. (1976). *Linear Algebra and Its Applications*. New York, U.S.A.: Academic Press.

[253] Gustafson, D.E. and W.C. Kessel (1979). Fuzzy clustering with a fuzzy covariance matrix. In Proc. IEEE CDC, San Diego, CA, USA, pp. 761–766.

[254] Bezdek, J.C. (1981). Pattern Recognition with Fuzzy Objective Function. Plenum

Press, New York.

[255] Babuˇska, R. and H.B. Verbruggen (1995). Identification of composite linear models via fuzzy clustering. In Proceedings European Control Conference, Rome, Italy, pp. 1207–1212.

[256] Zheng, Y.J., Sheng, W.G., Sun, X.M. and Chen, S.Y., 2016. Airline passenger profiling based on fuzzy deep machine learning. IEEE transactions on neural networks and learning systems, 28(12), pp.2911-2923.

[257] Zhang, X., Pan, X. and Wang, S., 2017, November. Fuzzy DBN with rule-based knowledge representation and high interpretability. In 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) (pp. 1-7). IEEE.

[258] Shuang, F. and Chen, C.P., 2017, October. Fuzzy restricted boltzmann machine and deep belief network: a comparison on image reconstruction. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 1828-1833). IEEE.

[259] Yazdanbakhsh, O. and Dick, S., 2019. A Deep Neuro-Fuzzy Network for Image Classification. arXiv preprint arXiv:2001.01686.

[260] P. Wood, H. Pourmeidani, and R. F. DeMara, “Modular simulation framework for process variation analysis of MRAM-based deep belief networks,” 2020 SoutheastCon, 2020, pp. 1-2, doi: 10.1109/SoutheastCon44009.2020.9368299.

[261] H. Pourmeidani, P. Debashis, Z. Chen, R. F. DeMara, and R. Zand, “Electrically-tunable stochasticity for spin-based neuromorphic circuits: self-adjusting to variation,” in

2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS). IEEE, 2020, pp. 81–84.

[262] H. Pourmeidani, P. Debashis, Z. Chen, and R. F. DeMara, " Process Variation Sensitivity of Spin Orbit Torque Perpendicular Nanomagnets in DBNs," IEEE Transactions on Magnetics, 2021.

[263] V. Ostwal and J. Appenzeller, Spin-orbit torque controlled Magnetic Tunnel Junction with low thermal stability for tunable random number generation, IEEE Magnetics Letters 10, 4503305-1 – 4503305-5 (2019).

[264] V. Ostwal, R. Zand, R. DeMara, and J. Appenzeller, A novel compound synapse using probabilistic spin-orbit-torque switching for MTJ based deep neural networks, IEEE Journal on Exploratory Solid-State Computational Devices and Circuits, 5(2), pages 182-187, Dec. 2019.

[265] Shimeng Yu, “Neuro-Inspired Computing With Emerging Nonvolatile Memory,” Proc. IEEE , 106(2), 2018.

[266] G. C. Adam, A. Khiat, and T. Prodromakis, “Challenges hindering memristive neuromorphic hardware from going mainstream,” Nat. Commun., 9(1), pages 2–5, 2018.

[267] D. Garbin, E. Vianello, O. Bichler, Q. Rafhay, C. Gamrat, G. Ghibaudo, B. DeSalvo, and L. Perniola. “HfO2-Based OxRAM Devices as Synapses for Convolutional Neural Networks,” IEEE Trans. Electron Devices, 62(8), pages 2494–2501, 2015.

[268] S. Bhatti, R. Sbiaa, A. Hirohata, H. Ohno, S. Fukami, and S. N. Piramanayagam, “Spintronics based random access memory: a review,” Mater. Today, 20(9), pages 530–

548, 2017.

[269] Y. Song, J. Lee, S. Han, H. Shin, K. Lee, K. Suh, D. Jeong, G. H. Koh, S. Oh, J. Park, S. Park, B. Bae, O. Kwon, K. Hwang, B. Seo, Y. Lee, S. Hwang, D. Leee, Y. Ji, K. Park, G. Jeong, H. Hong, K. Lee, H. Kang, and E. Jung. "Demonstration of Highly Manufacturable STT-MRAM Embedded in 28nm Logic," IEEE Int. Electron Devices Meet., pages 18–2, 2019.

[270] L. Wei, J. G. Alzate, U. Arslan, J. Brockman, N. Das, K. Fischer, T. Ghani, O. Golonzka, P. Hentges, R. Jahan, P. Jain, B. Lin, M. Meterelliyoz, J. O'Donnell, C. Puls, P. Quintero, T. Sahu, M. Sekhar, A. Vangapaty, C. Wiegand, F. Hamzaoglu."A 7Mb STT-MRAM in 22FFL FinFET Technology with 4ns Read Sensing Time at 0.9V Using Write-Verify-Write Scheme and Offset-Cancellation Sensing Technique," IEEE Int. Solid-State Circuits Conf., pages 480–481, 2019.

[271] K. Garello1, F. Yasin, S. Couet, L. Souriau, J. Swerts, S. Rao, S. Van Beek, W. Kim, E. Liu, S. Kundu, D. Tsvetanova, N. Jossart, K. Croes, E. Grimaldi, M. Baumgartner, D. Crotti1, A. Furnémont, P. Gambardella, G.S. Kar. "SOT-MRAM 300mm integration for low power and ultrafast embedded memories K.," IEEE Trans. Magn., 54(4), pages 81–82, 2018.

[272] J. Castro, M. Georgiopoulos, R. Demara, and A. Gonzalez. "Data-partitioning using the hilbert space filling curves: Effect on the speed of convergence of Fuzzy ARTMAP for large database problems," Neural networks, 18(7), pages 967-984, 2005.