HETEROGENEOUS RECONFIGURABLE FABRICS FOR IN-CIRCUIT TRAINING AND EVALUATION OF NEUROMORPHIC ARCHITECTURES

by

RAMTIN ZAND
M.S. Sharif University of Technology, 2012
B.S. Imam Khomeini International 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

Spring Term
2019

Major Professor: Ronald F. DeMara

ABSTRACT

A heterogeneous device technology reconfigurable logic fabric is proposed which leverages the cooperating advantages of distinct magnetic random access memory (MRAM)-based look-up tables (LUTs) to realize sequential logic circuits, along with conventional SRAM-based LUTs to realize combinational logic paths. The resulting Hybrid Spin/Charge FPGA (HSC-FPGA) using magnetic tunnel junction (MTJ) devices within this topology demonstrates commensurate reductions in area and power consumption over fabrics having LUTs constructed with either individual technology alone. Herein, a hierarchical top-down design approach is used to develop the HSC-FPGA starting from the configurable logic block (CLB) and slice structures down to LUT circuits and the corresponding device fabrication paradigms. This facilitates a novel architectural approach to reduce leakage energy, minimize communication occurrence and energy cost by eliminating unnecessary data transfer, and support auto-tuning for resilience. Furthermore, HSC-FPGA enables new advantages of technology co-design which trades off alternative mappings between emerging devices and transistors at runtime by allowing dynamic remapping to adaptively leverage the intrinsic computing features of each device technology. HSC-FPGA offers a platform for fine-grained Logic-In-Memory architectures and runtime adaptive hardware.

An orthogonal dimension of fabric heterogeneity is also non-determinism enabled by either low-voltage CMOS or probabilistic emerging devices. It can be realized using probabilistic devices within a reconfigurable network to blend deterministic and probabilistic computational models. Herein, consider the probabilistic spin logic p-bit device as a fabric element comprising a crossbar-structured weighted array. The Programmability of the resistive network interconnecting p-bit devices can be achieved by modifying the resistive states of the array's weighted connections. Thus, the programmable weighted array forms a CLB-scale macro co-processing element with bitstream programmability. This allows field programmability for a wide range of classification problems and recognition tasks to allow fluid mappings of probabilistic and deterministic computing approaches. In particular, a Deep Belief Network (DBN) is implemented in the field using recurrent layers of co-processing elements to form an $n \times m_1 \times m_2 \times ... \times m_i$ weighted array as a configurable hardware circuit with an n-input layer followed by $i \geq 1$ hidden layers. As neuromorphic architectures using post-CMOS devices increase in capability and network size, the utility and benefits of reconfigurable fabrics of neuromorphic modules can be anticipated to continue to accelerate.

Dedicated to my dear Golareh, and my wonderful family.

ACKNOWLEDGMENTS

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 the opportunity for me to join the Computer Architecture Lab (CAL) research team, and conduct my research under his supervision. He has kindly supported me throughout my Ph.D. journey, and always been there to ensure that I will reach my professional goals and even go beyond them. I would also like to thank my committee members Dr. Kalpathy Sundaram, Dr. Mingjie Lin, Dr. Annie Wu, and Dr. Deliang Fan for supporting me and my research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION[1]

1.1 Research Motivation

The objective of this dissertation is to architect a next-generation Hybrid Spin- and Charge-based Field Programmable Gate Array (HSC-FPGA) by innovating design techniques, circuit modules, synthesis scripts, and transportable libraries. HSC-FPGAs are reconfigurable logic arrays that leverage the cooperative roles of emerging spintronic and CMOS devices within a runtime adaptable platform. HSC-FPGAs enable a new advance of "*technology co-design*," which trades off alternative mappings between emerging devices and CMOS technology during synthesis-time and also during runtime. The targeted emerging computing devices, and yet-to-be-discovered architectural innovations utilizing them, become readily explored. Many sub-fields of computing including post-CMOS design, advances in Computer-Aided Design (CAD) models and optimizations, and applications become reinvigorated to attain reliable and energy-sparing hardware/software systems at low cost.

Towards the eventual realization of this objective, the HSC-FPGA Configurable Logic Block (H-CLB) shown in Figure 1.1 provides an enabling element.Thus, the dissertation will advance post-CMOS computing by spanning Look Up Tables (LUTs) for two spin-based devices. The potential benefits are summarized in Figure 1.1, which can be realized similarly to how field-programmability had enabled new hardware, CAD, pipelining, soft-cores, and application advances during the CMOS design era, by increasing access to the devices themselves.

[1] ©2017 IEEE. Part of this chapter is reprinted, with permission, from [1].

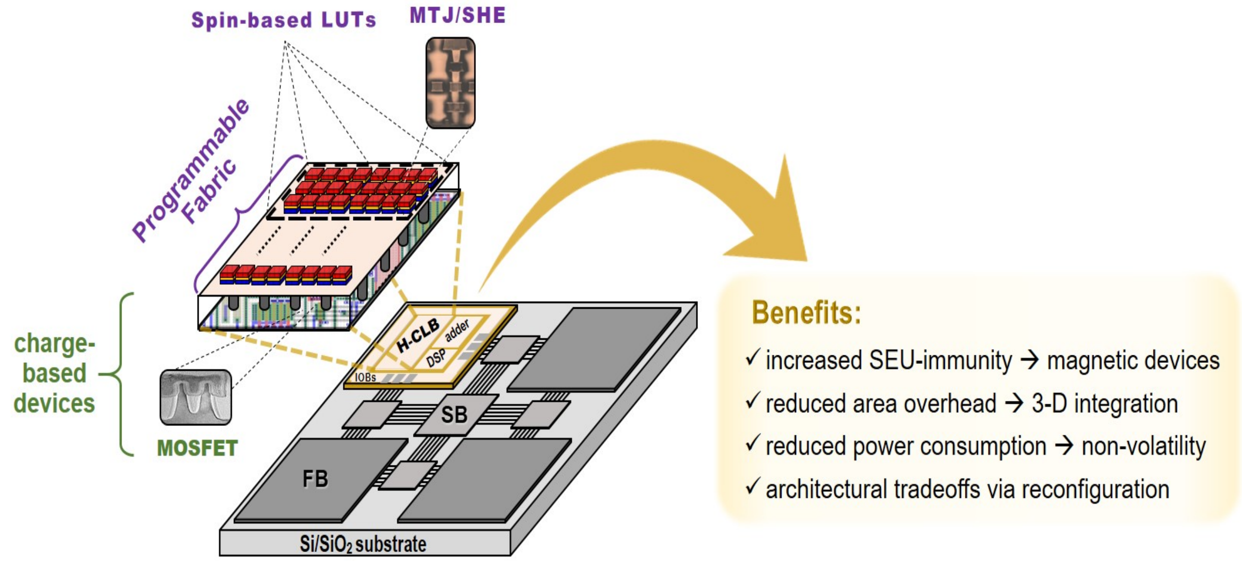

Spin-based LUTs
MTJ/SHE
Programmable Fabric
charge-based devices
MOSFET
H-CLB
adder
DSP
IOBs
SB
FB
Si/SiO2 substrate
Benefits:
✓ increased SEU-immunity → magnetic devices
✓ reduced area overhead → 3-D integration
✓ reduced power consumption → non-volatility
✓ architectural tradeoffs via reconfiguration

Figure 1.1: HSC-FPGA component vision. Physical devices consist of spintronic and CMOS elements within each H-CLB.

Similar to their ASIC counterparts, reconfigurable computing devices strive to surmount the growing technical challenges to improve their logic density, throughput performance, and power profiles. Thus with the geometrical and equivalent scaling trends guided by decades of International Technology Roadmap for Semiconductors (ITRS) projections nearing their end, new pathways towards these goals have been defined in ITRS 2.0 along with the IEEE International Roadmap for Devices and Systems (IRDS) initiative [18]. Two such technical thrusts identified for 2020 onward are leveraging beyond-CMOS devices (ITRS 2.0 theme 5) and utilizing heterogeneous components (ITRS 2.0 theme 4) to realize fundamentally new ways to compute. The perspective taken herein is that a reconfigurable computing paradigm can significantly advance both of these declared ITRS 2.0 themes.

Within the post Moore era, there are several motivations for pursuing novel reconfigurable fabrics of heterogeneous device technologies. Foremost, their one-time design and fabrication model minimizes the recurring engineering effort for post-CMOS devices, while amortizing development costs across multiple applications. Thus, reconfigurable fabrics may offer a more cost-effective approach to utilizing emerging devices. Additionally, post-CMOS ingrained field-programmable fabrics expand the accessibility of emerging devices to vast populations of circuit designers, including the majority of those who lack foundry access. Such a pre-fabrication approach with later field-programmability minimizes the need for extensive post-CMOS circuit design, verification, and validation expertise. Field-programmability also eliminates the computational demands, delays, and inaccuracies of simulation-based modeling associated with emerging devices. Instead, heterogeneous fabrics support rapid and direct realizations in hardware.

As a fundamentally different way to compute, the mapping of operations to device technologies remains fluid. Flexible mappings become possible not only during circuit synthesis, but also during execution-time. Thus when execution demands change, the architecture can adapt by utilizing a preferred device technology within its datapaths via reconfiguration of hardware components. This leverages the complementary characteristics of CMOS and emerging devices by increasing the flexibility in its binding of logic and memory roles to distinct device technologies. This is introduced herein as a post-CMOS era approach referred to as "*technology co-design.*" Overall, the hypothesis is as follows: reconfigurable fabrics of heterogeneous CMOS and spin-based devices offer an orthogonal dimension of technology adaptation to balance throughput, energy consumption, and resilience beyond static emerging device architectures, fixed hybrid emerging/CMOS architectures, and CMOS-only reconfigurable platforms.

1.2 Need for Next Generation Reconfigurable Fabrics

Frequently-cited motivations for embracing reconfigurable fabrics are listed in Table 1.1, including extensions to sub-10nm regimes. Fabric flexibility and accessibility allows realization of logic elements at medium and fine granularities while incurring low Non-Recurring Engineering (NRE) and Time-To-Market (TTM). Reconfigurable fabrics have been demonstrated to provide a viable solution for process-voltage-temperature variation induced problems and could be utilized effectively for fault recovery [31, 32, 33, 34, 35, 36, 37, 38]. They can support circuit synthesis specific to the application at-hand, including localization of data stores. Moreover, highly-scaled devices are expected to increasingly rely on in-situ reconfigurability to mitigate process variation. Thus, their in-field adaptability to intrinsic "as-built" device switching characteristics enables resiliency, as advocated in [39, 40]. Fabrics can also be tuned to meet energy profiles at runtime [39, 41].

Table 1.1: Strengths of Reconfigurable Logic Fabrics over ASIC/CPU/GPU at <10nm regimes.

Attribute	**Typical Benefits**	**Relation to Emerging Devices**	**Prototypes**
Flexibility & Accessibility	- decreased NRE costs & TTM - hardware/software co-design	- usable by designers without foundry access - knowledge/behavior encapsulation	[19, 20, 21, 22]
Energy-sparing Potential	- datapaths synthesized for application - local data stores near data usage	- near-zero standby energy non-volatile designs - increased local capacities in same energy budget	[23, 24]
Bloat-free	- customized accelerator hardware - reduce middleware	- logic-in-memory capabilities - device-physics enabled computing	[25, 26, 27]
Resiliency	- amorphous spares	- true PVT solution and alpha-particle immunity	[28, 29]
Adaptability	- leverage Intrinsic behavior	- tunable to match dynamic requirements	[29, 30]

Extending reconfiguration capabilities using emerging devices is highly desirable. Among promising devices, the 2017 Magnetism Roadmap [42] identifies nanomagnetic devices as capable post-CMOS candidates, of which Magnetic Random Access Memories (MRAMs) are considered feasibly-implemented. Thus, the proposed research is well-motivated by the established aims of academia and industry to:

- Promote emerging devices to surmount scaling challenges of CMOS devices [43, 44]. We will examine and adopt various emerging spintronic devices to further develop novel circuits and architectures to realize higher density with significantly improved power-delay-product (PDP) over CMOS-based FPGAs for identical design rules and power budget. While in this dissertation we will focus on leveraging spintronic devices, alternative methods such as asynchronous switching approaches [45, 46, 47] and Quantum-dot Cellular Automata (QCA) methods [48, 49, 50, 51] also aim towards reduced energy consumption.
- Realize the benefits of Non-Volatile Memory (NVM) to reduce leakage energy for ultra-low-power reconfigurable computing [52, 53, 54]. We propose to utilize the accepted spin-based device models to configure logic arrays, and extend beyond the previous work to include non-volatile functionality.
- Employ the characteristics of the spin-based devices for reliability benefits such as radiation hardness [55, 56]. We will build upon reliability analysis methods for spintronic devices [57, 58] to benefit reliability by conferring the circuit at-hand with a heterogeneous palette of emerging devices that can be configured as needed at runtime.
- Develop transportable libraries to facilitate the wider application of spintronic devices. Our proposed HSC-FPGA platform can enable application-scale study, thus overall benefits can be more readily quantified. Overall, the foundational focus of this research can advance several fields spanning post-CMOS devices, theory, modeling, and circuit innovation.

1.3 Advancing from Homogeneity towards Heterogeneity in Reconfigurable Computing Fabrics

As depicted in Figure 1.2, FPGA fabrics continue their transition towards embracing the benefits of increased heterogeneity along several cooperating dimensions. Since the inception of the first field-programmable devices, various granularities of general-purpose reconfigurable logic blocks and dedicated function-specific computational units have been added to fabric structures. These have resulted in increased computational functionality compared to homogeneous fabrics [59, 60, 61]. Over the last ten years, reprogrammable fabrics have embraced further highly-dedicated special-purpose co-processing units to handle complex floating-point computations [62]. Some of the standard co-processing units that appear within many contemporary FPGAs are Digital Signal Processing (DSP) blocks [63, 64], Multiplier-Accumulators (MACs) [65], and multi-bit block RAMs [66], as well as processor hardcores which are commonly embedded within the fabric of many leading commercially-available reconfigurable devices.

The upper rightmost corner of Figure 1.2 depicts that emerging devices could advance new transformative opportunities for exploiting technology-specific advantages, which we refer to as Technology Heterogeneity. Technology heterogeneity recognizes the cooperating advantages of CMOS devices for their rapid switching capabilities, while simultaneously embracing emerging devices for their non-volatility, near-zero standby power, high integration density, and radiation-hardness [67, 68]. Realization of technology heterogeneity in a field-programmable fabric enables synthesis-time co-design and dynamic run-time adaptability among device technologies. Thus, we propose exciting feasible research towards utilizing the proven spin-based devices to complement CMOS devices.

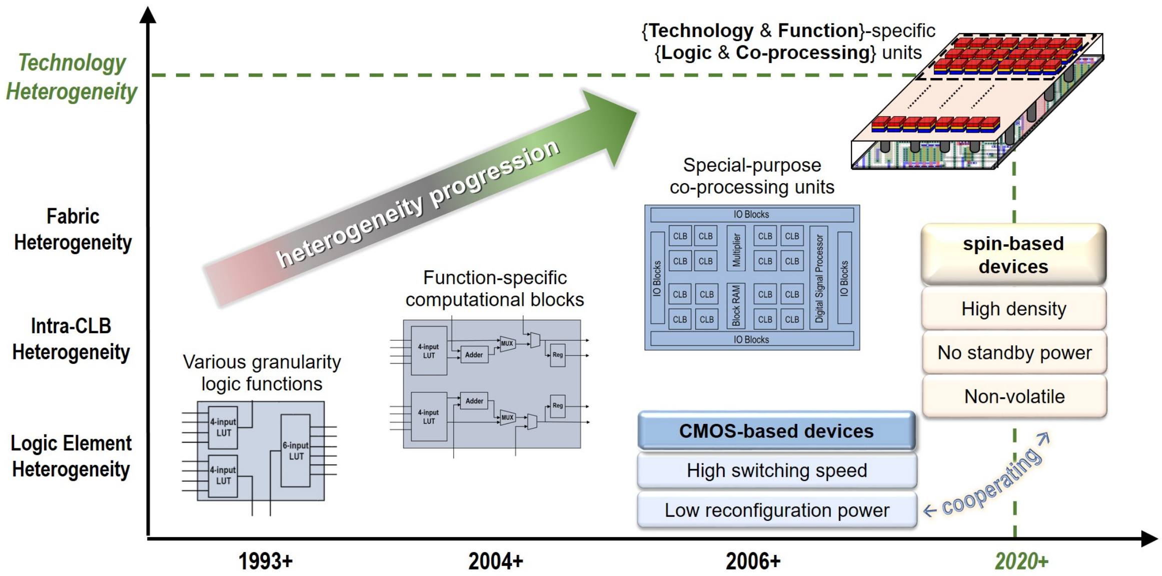


Figure 1.2: Escalation of field-programmable heterogeneity within chronological and structural contexts [1].

1.4 Technology Heterogeneity in Reconfigurable Fabrics

The mentioned motivations for embracing reconfigurable fabrics are achieved at a cost of increased fabric area and power consumption, as well as a decreased performance compared to the application-specific integrated circuits (ASICs). Thus, Innovations using emerging devices within reconfigurable fabrics have been sought to bridge the gaps needed to provide these benefits.

Currently, static random access memory (SRAM) cells are the basis for most of the commercial FPGAs, and can be found in the well-known Xilinx and Intel products. In FPGAs, SRAM cells are employed within programmable switching blocks to control the interconnection between logic building blocks. Moreover, they are utilized in lookup-tables (LUTs) to store the logic function configuration data, which constitute the primary components in reconfigurable fabrics. In particular, LUT is a memory with 2^m cells in which the truth table of an m-input Boolean function is stored. The re-programmability of the SRAM cells, and the fact that they can be implemented by highly-scaled CMOS technology, have made the SRAM-based FPGAs the most popular reconfigurable fabric in market. However, SRAM cells also have some limiting attributes which caused FPGAs to have a niche market share of ASICs.

In [69], Kuon and Rose have provided a comprehensive comparison between SRAM-based FPGAs and ASICs in terms of area, performance, and power consumption. They have reported that in order to achieve a same functionality and performance in an FPGA as an ASIC, FPGA requires significantly larger area while consuming approximately 14 times more power. This is mainly due to the crucial drawbacks of the SRAM cells such as:

- high static power: due to the existence of intrinsic leakage current which is significantly increasing by technology scaling.
- volatility: SRAM is volatile, therefore all functions must be reprogrammed upon each power-up. Consequently, an external non-volatile memory is required to be integrated into the chip either in the same package or on the printed circuit board level.
- low logic density: SRAM consists of six transistors which limits the logic density.

The aforementioned SRAM's drawbacks have motivated exploration of alternative LUT designs, as listed in Table 1.2. One of the introduced alternatives is based on non-volatile flash-based LUTs, however it targets a niche market due to their low reconfiguration endurance [70, 71]. Higher endurance non-volatile LUTs can be enabled by emerging resistive technologies, such as spintronic storage elements [72, 73, 74, 75, 76, 77, 78], resistive random access memory (RRAM) [79, 80, 81, 82], and phase change memory (PCM) [83, 84]. Although PCM can offer non-volatility, its considerable reconfiguration power and high write latency can significantly exceed that of an SRAM LUT. Spintronic devices offer non-volatility, near-zero static power, and high integration density [88, 89]. Two of the spin-based devices, which are previously proposed for use in reconfigurable fabrics are magnetic tunnel junctions (MTJs) [76, 75, 78, 2, 3, 85] and domain wall (DW)-based racetrack memory (RM) [73, 74, 86]. RM is effective for non-volatility and area density, although previous designs can incur significant delay and energy cost due to excessive shift activities to configure the implemented logic function. Hence, MTJ-based LUTs are proposed to be placed at critical points of a large-scale digital circuit to implement various logic functions as a runtime adaptable fabric under middleware control. the magnetic LUTs provides the fabric with sufficient reconfigurability features to mitigate process variations. The fabric will be leveraged for fault detection and recovery using the adaptive self-healing approaches. MTJs comprising the storage elements in the adaptable LUTs are vertically-integrated as a backend process of typical CMOS fabrication, which significantly reduces the area cost of the redundancy.

Moreover, SRAM-based FPGAs, like all CMOS-based fabrics, are susceptible to radiation-induced transient soft faults such as single event upsets (SEUs), which primarily affect SRAM-based storage cells [87]. Thus, research into the design of suitable placements with improved soft error immunity and energy profiles is urgently sought using a number of feasible physical devices including RRAM [80] and magnetic random access memory (MRAM) [55, 4, 88, 89]. This trend has been motivated by aggressive CMOS technology scaling in digital circuits has resulted in significant increase in transient fault rates, as well as timing violations due to process variation (PV) that consequently reduces the performance and reliability of the emerging very large scale integrated (VLSI) circuits. For instance, the probability of single upsets, and more realistically, multiple upsets, is projected to increase several fold at sea-level for sub-10nm technology nodes [90, 91]. By the extensions to sub-10nm regimes, error resiliency has become a major challenge for microelectronics industry, particularly mission critical systems, e.g. space and terrestrial applications. The ability of FPGAs to correctly execute the complicated tasks in harsh environments significantly relies on their fault-handling and radiation hardening techniques.

Table 1.2: Characteristics of enabling LUT technologies. “✓” or “–” indicates strength/limitation relative to SRAM-based LUT.

Design		Baseline	[71]	[80]	[84]	[75]	[56]	[89]
Technology		SRAM	FLASH	RRAM	PCM	MRAM	MRAM	MRAM
Non-Volatile		NO	YES	YES	YES	YES	YES	YES
Endurance		0	– – –	– – –	– – –	–	–	–
Area		0	✓	✓	✓	✓	✓	✓
Read	Power	0	0	✓	0	✓	– –	–
Operation	Delay	0	0	0	0	0	–	– –
Standby Power		0	✓ ✓	✓ ✓	✓	✓ ✓	✓ ✓	✓ ✓
Write	Power	0	–	– – –	– – –	– –	– –	– –
Operation	Delay	0	– –	– – –	– – –	– –	– –	– –
Radiation	SEU	NO	NO	NO	NO	NO	YES	YES
Hardness	DNU	NO	NO	NO	NO	NO	NO	YES

Leveraging MRAMs as storage elements within LUT circuits has the potential to significantly increases their radiation immunity due to the radiation hardness characteristic of spin-based devices. However, the access and sensing circuitry for MRAM still requires transistors, and thus is still susceptible to radiation-induced faults. Therefore, circuit-level innovations are sought to achieve immunity to radiation-induced transient faults such as SEUs and double node upsets (DNUs). In recent years, various radiation hardening techniques are investigated to develop SEU-tolerant MRAM-based LUTs [56, 88]. In particular, in [89] authors have proposed a single-event double-node upset tolerant MRAM-based LUT, which provides multiple upset resiliency at the cost of increased read energy and area consumption with baseline efficacy.

1.5 Logic Paradigm Heterogeneity

The interrelated fields of machine learning (ML), and artificial neural networks (ANN) have grown significantly in previous decades due to the availability of powerful computing systems to train and simulate large scale ANNs within reasonable time-scales, as well as the abundance of data available to train such networks in recent years. The resulting research has realized a bevy of

ANN architectures that have performed incredible feats including a wide range of classification problems, and various recognition tasks.

Most ML techniques in-use today rely on supervised learning, where the systems are trained on patterns with a known desired output, or label. However, intelligent biological systems exhibit unsupervised learning whereby statistically correlated input modalities are associated within an internal model used for probabilistic inference and decision making [92]. One interesting class of unsupervised learning approaches that has been extensively researched is the Restricted Boltzmann machine (RBM) [93]. RBMs can be hierarchically organized to realize deep belief networks (DBNs) that have demonstrated unsupervised learning abilities, such as natural language understanding [94]. Most RBM and DBN research has focused on software implementations, which provides flexibility, but requires significant execution time and energy due to large matrix multiplications that are relatively inefficient when implemented on standard Von-Neumann architectures due to the memory-processor bandwidth bottleneck when compared to hardware-based in-memory computing approaches [95]. Thus, research into hardware-based RBM designs has recently sought to alleviate these constraints.

Previous hardware-based RBM implementations have aimed to overcome software limitations by utilizing FPGAs [96, 97] and stochastic CMOS [98]. In recent years, emerging technologies such as RRAM [99, 100] and PCM [101] are proposed to be leveraged within the DBN architecture as weighted connections interconnecting building blocks in RBMs. While most of the previous hybrid Memristor/CMOS designs focus on improving the synapse behaviors, this dissertation overcomes many of the preceding challenges by utilizing a novel spintronic p-bit device that leverages intrinsic thermal noise within low energy barrier nanomagnets to provide a natural building block for RBMs within a compact and low-energy package. The contribution of this dissertation goes beyond using low-energy barrier MTJs, as has been previously introduced for a neuron in spiking neuromorphic systems [102, 103]. This is the first effort to use MTJs with near-zero energy barriers as neurons within an RBM implementation. Additionally, various parameters of a hybrid CMOS/spin weight array structure are investigated for metrics of power dissipation, and error rate using the MNIST digit recognition benchmarks.

Within the post-Moore era ahead, several design factors and fabrication constraints increasingly emphasize the requirements for in-circuit adaptation to as-built variations. These include device scaling trends towards further reductions in feature sizes [104], the narrow operational tolerances associated with the deployment of hybrid Complementary Metal Oxide Semiconductor (CMOS) and post-CMOS devices [91, 105], and the noise sensitivity limits of analog-assisted neuromorphic computing paradigms [106]. While many recent works have advanced new architectural approaches for the evaluation phase of neuromorphic computation utilizing emerging hardware devices, there have been comparatively fewer works to investigate the hardware-based realization of their training and adaptation phases that will also be required to cope with these conditions. Thus, this dissertation develops one of the first viable approaches to address post-fabrication adaptation and retraining in-situ of resistive weighted-arrays in hardware, which are ubiquitous in post-Moore neuromorphic approaches. Namley, a tractable in-field reconfiguration-based approach is developed to leverage in-field configurability to mitigate the impact of process variation. Reconfigurable fabrics are characterized by their fabric flexibility, which allows realization of logic elements at medium and fine granularities, as well as in-field adaptability, which can be leveraged to realize variation tolerance and fault resiliency as widely-demonstrated for CMOS-based approaches such as [31, 39]. Utilizing reconfigurable computing by applying hardware and time redundancy to the digital circuits offers promising and robust techniques for addressing the above-mentioned reliability challenges. For instance, it is shown in [39] that a successful refurbishment for a circuit with 1,252 LUTs can be achieved with only 10% spare resources to accommodate both soft and hard faults.

1.6 Contributions of the Dissertation

The main focus of this dissertation is to architect a next-generation hybrid spin- and charge-based reconfigurable fabric by innovating design techniques, circuit modules, synthesis scripts, and transportable libraries listed in Table 1.3.

Table 1.3: Research Outcomes of the dissertation spanning Methods (M), Libraries/Models (L), Tools (T), and Publications (P).

Abstraction Level	**Research Outcomes & Work Products**	**Evaluation Metrics**	**Case Studies**	**Deliverables**			
				M	*L*	*T*	*P*
Device	Verilog-A model of MTJ's resistance	- Resistance vs. Experimental Results	N/A		✓		[2] [3] [107] [86]
	MATLAB Model of STT/SHE switching for MTJ devices	- Switching Delay vs. Experimental Results	N/A		✓		
Circuit	STT-MTJ based Adaptive LUT	- Read Power - Read Delay - Read PDP	N/A		✓		[3]
	SHE-MTJ based Fracturable LUT	- Read Energy - Write Energy - Max. CLK Freq.	N/A	✓	✓		[2]
	Radiation-hardened MRAM-LUT	- SEU/DNU Tolerance - Read PDP - Device Count	N/A	✓	✓		[4]
	TG-based MRAM bit-cell	- Write EDP - Area	N/A		✓		[108]
Architecture	HSC-FPGA	- Read Energy -Write Energy - Standby Power - Device Count	MCNC ISCAS99 ITC99	✓			[109]
	SNRA	- Read Power - Standby Power - Device Count	MNIST Dataset	✓			[6] [8]
	Verilog HDL of CD Algorithm	- Error Rate	MNIST Dataset	✓	✓		[6]
	PIN-Sim	- Error Rate - Power Consumption - Device Count	MNIST CIFAR10			✓	[9]

The main contributions of the dissertation are summarized below, which are comprehensively described in the following chapters of the dissertation:

- Developed an approach to model the behavior of 2-terminal STT-MTJs, and 3-terminal SHE-MTJ devices, as described in Chapter 2. In particular, Verilog-A was utilized to model the resistive behavior of the MTJs. Then, the model was leveraged in SPICE circuit simulations to design various hybrid spin/CMOS-based circuits. Moreover, a MATLAB based module was developed to provide details regarding the switching characteristics of the 2-terminal and 3-terminal devices. In addition to the spin/COMS-based circuits designed in this dissertation [2, 4, 3, 6, 108, 109], the developed modeling approach has been also widely-used by other researchers in the related works, such as [85, 107, 86, 110, 111, 112, 113, 114, 115, 116, 117, 118].

- Researched STT-MRAM as a promising alternative for SRAM in reconfigurable fabrics. As described in Section 3.1 of the dissertation, we leveraged physical characteristics of MTJs to design a unique reference MTJ which has a calibrated resistance matching the STT-based LUT (STT-LUT) circuit requirements to provide optimal reading operation. Results obtained show 42% and 70% power-delay product (PDP) improvement over previous MTJ-based LUT designs. Moreover, a four-input adaptive STT-based LUT (A-LUT) is proposed based on the developed STT-LUT, which is configurable to function in seven independent modes. An n-input A-LUT exhibits PDP which can be a fraction of n-input STT-LUT PDP, when performing two-input to $(n-1)$-input Boolean logic functions.

- As described in Section 3.2 of the dissertation, we leveraged SHE-MTJ devices to design an energy-efficient nonvolatile LUT. Functionality of the proposed SHEMTJ-based LUT is validated using SPICE simulation. Our proposed SHEMTJ-based LUT (SHELUT) is compared with the most energy-efficient MTJ-based LUT circuits. The obtained results show more than 6%, 37%,and 67% improvement over three previous MTJ-based designs in terms of read energy consumption. Moreover, the reconfiguration delay and energy of the proposed design is compared with that of the MTJ-based LUTs which utilize the STT switching approach for reconfiguration. The results exhibit that SHELUT can operate at 78% higher clock frequency while achieving at least 21% improvement in terms of reconfiguration energy consumption. The operation-specific clocking mechanisms for managing the SHELUT operations are introduced along with detailed analyses concerning tradeoffs. Results are also extended to design a 6-input fracturable LUT using SHEMTJs.

- Developed a radiation-hardened non-volatile LUT circuit utilizing SHE-MRAM devices, as explained in Section 3.4. The proposed hardening technique is based on using feedback transistors, as well as increasing the radiation capacity of the sensitive nodes. Simulation results show that our proposed LUT circuit can achieve single node upset (SNU) and double node upset (DNU) tolerance with more than 38% and 60% power-delay product improvement as well as 26% and 50% reduction in device count compared to the previous energy-efficient radiation-hardened LUT designs. Finally, we have performed a process variation analysis showing that the MNU immunity of our proposed circuit is realized at the cost of increased susceptibility to transistor and MRAM variations compared to an unprotected LUT design.

- Proposed various energy-efficient write schemes for switching operation of SHE-MTJs. A transmission gate (TG)-based write scheme is proposed, which provides a symmetric and energy-efficient switching behavior. Circuit Simulation results showed that the TG-based write scheme advantages in terms of device count and switching energy. In particular, it can operate at 12% higher clock frequency while realizing at least 13% reduction in energy consumption compared to the most energy-efficient write circuits. We analyzed the performance of the implemented write circuits in presence of process variation (PV) in the transistors' threshold voltage and SHE-MTJ dimensions. Results showed that the proposed

TG-based design is the second most PV-resilient write circuit scheme for SHE-MTJs among the implemented designs. Finally, we proposed the 1TG-1T-1R SHE-based magnetic random access memory (MRAM) bit cell based on the TG-based write circuit. Comparisons with several of the most energy-efficient and variation-resilient SHE-MRAM cells indicate that 1TG-1T-1R delivers reduced energy consumption with 43.9% and 10.7% energy-delay product improvement, while incurring low area overhead.

- Propoesd a hybrid device technology reconfigurable logic fabric which leverages the cooperating advantages of distinct MRAM-based LUTs to realize sequential logic circuits, along with conventional SRAM-based LUTs to realize combinational logic paths. The resulting Hybrid Spin/Charge FPGA (HSC-FPGA) using MTJ devices within this topology demonstrates commensurate reductions in area and power consumption over fabrics having LUTs constructed with either individual technology alone. In Chapter 4 of the dissertation, a hierarchical top-down design approach is used to develop the HSC-FPGA starting from the configurable logic block (CLB) and slice structures down to LUT circuits and the corresponding device fabrication paradigms. The Xilinx ISE Design Suite was used to implement, and evaluate resource utilization contributing to HSC-FPGA's fabric-level simulation that yields 70% and 30% reductions in standby and read power, respectively, for various ISCAS-89 and ITC-99 benchmark circuits. To address the MTJ fabrication process and challenges, a circuit-level modular redundancy based method is developed to increase the resiliency of the MRAM-LUTs against process variation. The corresponding power consumption and area utilization are analyzed to formulate extensive device tradeoffs resulting in recommendations towards future multi-device based reconfigurable fabrics.

- In Chapter 5, a low-energy hardware implementation of deep belief network (DBN) architecture is developed using near-zero energy barrier probabilistic spin logic devices (p-bits), which are modeled to realize an intrinsic sigmoidal activation function. A CMOS/spin based weighted array structure is designed to implement a restricted Boltzmann machine (RBM). Device-level simulations based on precise physics relations are used to validate the sigmoidal relation between the output probability of a p-bit and its input currents. Characteristics of the resistive networks and p-bits are modeled in SPICE to perform a circuit-level simulation investigating the performance, area, and power consumption tradeoffs of the weighted array. In the application-level simulation, a DBN is implemented in MATLAB for digit recognition using the extracted device and circuit behavioral models. The MNIST data set is used to assess the accuracy of the DBN using 5,000 training images for five distinct network topologies. The results indicate that a baseline error rate of 36.8% for a 784×10 DBN trained by 100 samples can be reduced to only 3.7% using a $784 \times 800 \times 800 \times 10$ DBN trained by 5,000 input samples. Finally, Power dissipation and accuracy tradeoffs for probabilistic computing mechanisms using resistive devices are identified.

- Developed a spintronic neuromorphic reconfigurable Array (SNRA) to fuse together power-efficient probabilistic and in-field programmable deterministic computing during both training and evaluation phases of restricted Boltzmann machines (RBMs), as described in Chapter 6. First, probabilistic spin logic devices are used to develop an RBM realization which is adapted to construct deep belief networks (DBNs) having one to three hidden layers of size 10 to 800 neurons each. Second, we designed a hardware implementation for the contrastive divergence (CD) algorithm using a four-state finite state machine capable of unsupervised training in $N+3$ clocks where N denotes the number of neurons in each RBM. The functionality of our proposed CD hardware implementation is validated using ModelSim simulations. We synthesize the developed Verilog HDL implementation of our proposed test/train control circuitry for various DBN topologies where the maximal RBM dimensions yield resource utilization ranging from 51 to 2,421 lookup tables (LUTs). Next, we leverageed SHE-MTJ based non-volatile LUT circuits to form a reconfigurable fabric. Finally, we compare the performance of our proposed SNRA with SRAM-based configurable fabrics focusing on the area and power consumption induced by the LUTs used to implement both CD and evaluation modes. The results obtained indicate more than 80% reduction in combined dynamic and static power dissipation, while achieving at least 50% reduction in device count.

- Leveraged MRAM technologies with thermally unstable nanomagnets to develop an intrinsic stochastic neuron as a building block for RBMs. The embedded MRAM-based neuron is modeled using precise physics equations. A probabilistic inference network simulator (PIN-Sim) is developed to realize a circuit-level model of an RBM utilizing resistive crossbar arrays along with differential amplifiers to implement the positive and negative weight values. The PIN-Sim is composed of five main blocks to train a DBN, evaluate its accuracy, and measure its power consumption. The MNIST dataset is leveraged to investigate the energy and accuracy tradeoffs of seven distinct network topologies in SPICE using the 14nm HP-FinFET technology library with the nominal voltage of 0.8V, in which an MRAM-based neuron is used as the activation function. The software and hardware level simulations indicate that a $784\times200\times10$ topology can achieve less than 5% error rates with $\sim 400pJ$ energy consumption. The error rates can be reduced to 2.5% by using a $784\times500\times500\times500\times10$ DBN at the cost of $\sim 10\times$ higher energy consumption and significant area overhead. Finally, the effects of specific hardware-level parameters on power dissipation and accuracy tradeoffs are identified via the developed PIN-Sim framework.

CHAPTER 2: FUNDAMENTALS AND MODELING OF MAGNETIC TUNNEL JUNCTIONS[1]

Figure 2.1 depicts the vertical structure of an MTJ [5, 119], which consist of two ferromagnetic (FM) layers: *(1) Fixed Layer*, that is magnetically-pinned and utilized as a reference layer, and *(2) Free Layer*, that its magnetic orientation can be switched. These two FM layers are separated by a thin oxide barrier, e.g. MgO [44]. The FM layers can have two different magnetization configurations called *parallel (P)* and *antiparallel (AP)*, according to which the MTJ's resistance changes between R_P and R_{AP}, respectively. The MTJ resistance is determined by the angle (θ) between the magnetization orientations of fixed layer and free layer due to the tunnel magnetoresistance (TMR) effect. The MTJ resistance in P (θ=0), and AP (θ=180) states is expressed by the following equations [11, 120, 10]:

$$R(\theta) = \frac{2R_{MTJ}(1+TMR)}{2+TMR+TMR\times\cos\theta} = \begin{cases} R_P = R_{MTJ}, & \theta = 0 \\ R_{AP} = R_{MTJ}(1+TMR), & \theta = \pi \end{cases} \tag{2.1}$$

$$R_{MTJ} = \frac{t_{ox}}{Factor \times Area \times \sqrt{\varphi}}\, exp(1.025 \times t_{ox} \times \sqrt{\varphi}) \tag{2.2}$$

$$TMR(T, V_b) = \frac{2P^2(1-\alpha_{sp}T^{3/2})^2}{1-P^2(1-\alpha_{sp}T^{3/2})^2} . \frac{1}{1+(\frac{V_b}{V_0})^2} \tag{2.3}$$

[1] © 2017 IEEE. Part of this chapter is reprinted, with permission, from [2].

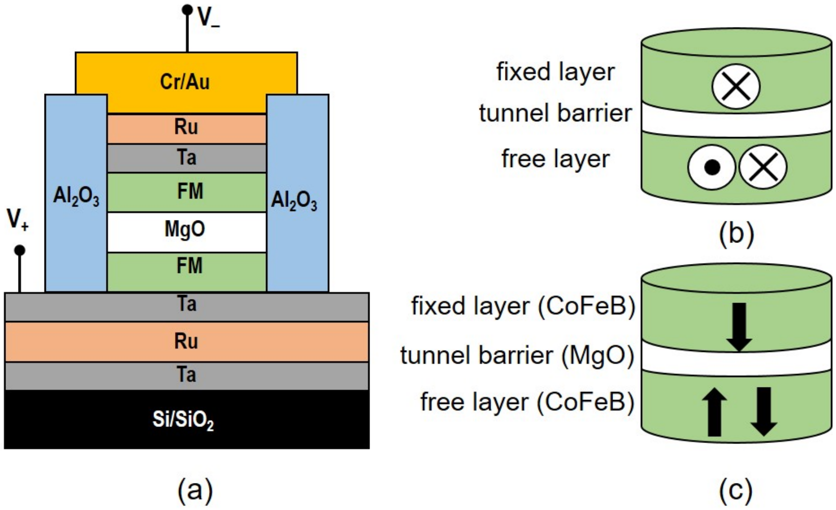


Figure 2.1: (a) MTJ vertical structure, (b) In-plane MTJ (IMTJ), and (c) Perpendicular MTJ (PMTJ) [2].

In the above equations, t_{ox} is the oxide thickness of MTJ, *Factor* is obtained from the resistance-area product (RA) value of the MTJ that relies on the material composition of its layers, *Area* is the surface area of the MTJ, φ is the oxide layer energy barrier height, TMR is the tunneling magnetoresistance, which relies on temperature (T) and bias voltage (V_b). P is the spin polarization factor, V_0 is a fitting parameter, and α_{sp} is a material-dependent constant.

The energy barrier between P and AP configurations of MTJ is in a range such that it can switch between configurations, while also retaining thermal stability. The magnetic direction of MTJ layers can be in the film plane or out of the film plane referred to as in-plane MTJ (IMA) and perpendicular MTJ (PMA) structure, respectively.

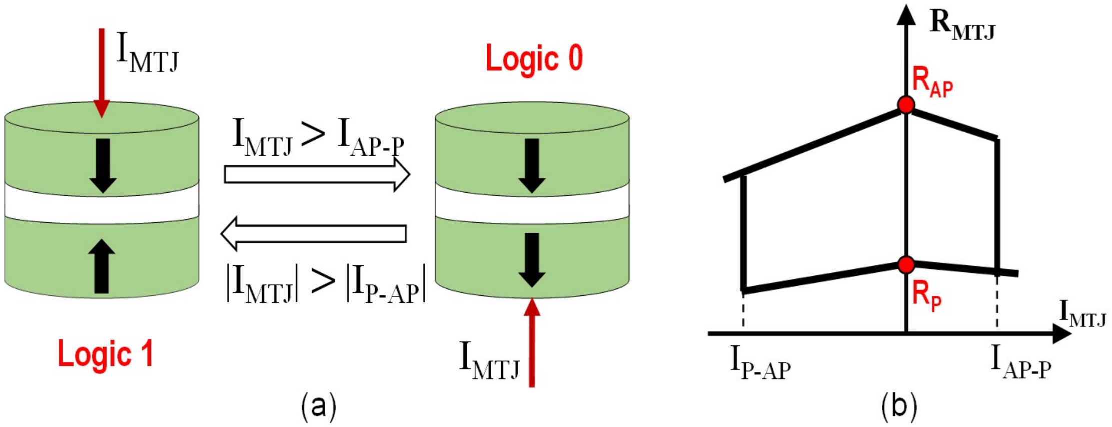


Figure 2.2: MTJ state change from AP to P due to the positive current $I_{MTJ} > I_{AP-P}$ condition, and vice versa. (b) MTJ resistance hysteresis curve relative to the I_{MTJ} [2].

Two of the conventional switching methods used for changing the magnetization orientation of free layers are Field-induced magnetic switching (FIMS) [121] and thermally assisted switching (TAS) [122]. In the mentioned approaches, a current source with an amplitude in range of milliampere (mA) was required to generate the magnetic field, which should be applied to switch the MTJ state. Thus, these approaches are not appropriate for low power integrated circuits, due to the significantly high switching energy consumption. In 1996, Slonczewski [123] proposed Spin Transfer Torque (STT) switching method, which is known as a promising alternative for changing the MTJ states.

2.1 Spin Transfer Torque (STT) Switching Approach

Based on the STT approach, a bidirectional spin-polarized current (I_{MTJ}) is required for switching MTJ nanomagnet configuration, as shown in Figure 2.2. Electrons that flow through the MTJ free layer will experience an exchange field which aligns the spin of the electron with the magnetization direction of the nanomagnet. This phenomenon is called *spin-filtering effect*. The conservation of the angular momentum results in the exertion of an opposite sign torque with equal magnitude on the free layer which eventually change its magnetization direction. The P or AP configuration of the MTJ is determined by the direction of the current that flows through it. The required bidirectional current could be produced by means of simple MOS-based circuits. Due to the vertical structure of the MTJ, it can be readily integrated at the back-end process of the CMOS fabrication [124, 125].

STT switching behavior can be categorized into two main regions based on the relation between I_{MTJ} and the switching critical current (I_C): (1) *precessional region* ($I_{MTJ} > I_C$) described by Sun model [126], where MTJ experiences a rapid precessional switching, and (2) *thermal activation region* ($I_{MTJ} < I_C$) defined by Brown model [127], in which the switching can occur with a long input current pulse due to the thermal activation. The switching duration in the precessional and thermal activation regions are described by Equations 2.4 and 2.5, respectively [125]:

$$\frac{1}{\tau_{STT}} = [\frac{2}{C + ln(\pi^2 \Delta)}].\frac{\mu_B P}{em(1+P^2)}(I_{MTJ} - I_C) \tag{2.4}$$

$$\frac{1}{\tau_{STT}} = \tau_0 e^{\Delta(1-\frac{I_{MTJ}}{I_C})} \tag{2.5}$$

where τ_{STT} is the mean switching duration, $C = 0.577$ is the Euler's constant, $\Delta = E/4k_B T$ is the thermal stability factor, m is the free layer magnetic moment, and τ_0 is the attempt period. In practice, MTJ is normally designed to work in precessional region with an input current amplitude larger than critical current to achieve high switching speed.

Equations 2.6 and 2.7 express the switching critical current for IMA (I_{c-IMA}) [128] and PMA

(I_{c-PMA}) [5] MTJ devices, respectively:

$$I_{c-IMA} = 2\alpha e M_S V (H_C + \frac{H_{eff}}{2})/g(\theta) P\bar{h} \tag{2.6}$$

$$I_{c-PMA} = \alpha\gamma e M_S V H_k/\mu_B g(\theta) \tag{2.7}$$

where α is the magnetic damping constant, μ_B is the Bohr magneton, γ is the gyromagnetic ratio, e is the electric charge, V is the volume of the free layer, M_S is the saturation magnetization, $\bar{h}$ is the reduced Planck's constant, H_C is the in-plane coercive field, H_{eff} is the effective out-of-plane demagnetization field, and H_k is the anisotropy field. The effective demagnetization field in IMA is approximately equal to the saturation magnetization, which is normally larger than anisotropy field in PMA. Thus, switching current for PMA is smaller than that of the IMA devices according to the Equations 2.6 and 2.7. Moreover, spin polarization efficiency factor, $g(\theta)$, is a function of the angle between free layer and fixed layer magnetization directions (θ), and is obtained by the Equations 2.8 [129] and 2.9 [11] for IMA and PMA MTJ devices, respectively.

$$g_{IMA} = [-4 + (P^{1/2} + P^{-1/2})(3 + cos\theta)/4]^{-1} \tag{2.8}$$

$$g_{PMA} = g_{SV} \pm g_{tunnel} = [-4 + (P^{1/2} + P^{-1/2})^3(3 + cos\theta)/4]^{-1} \pm [\frac{P}{2(1 + P^2 cos\theta)}] \tag{2.9}$$

where P is the spin polarization percentage of the tunnel current, g_{SV} is the spin polarization efficiency in a spin valve and g_{tunnel} is the spin polarization efficiency in tunnel junction nanopillars.

The dynamics of the magnetic moment of the free layer (m) in an STT-MTJ device is described by the Landau-Lifshitz-Gilbert (LLG) equation [130]:

$$\frac{(1+\alpha^2)}{\gamma}.\frac{d\hat{m}}{dt} = -\hat{m} \times \vec{H} - \alpha.\hat{m} \times (\hat{m} \times \vec{H}) + c_{STT}.\hat{m} \times (\hat{m} \times \hat{m}_p) \tag{2.10}$$

where $\hat{m}$ and $\hat{m}_p$ are the unit vectors of the free layer and pinned layer magnetizations, respectively. H is the effective perpendicular anisotropy field, α is the damping coefficient, and γ is the gyromagnetic ratio. The STT coefficient c_{STT} equals $\frac{\hbar PJ}{2et_f M_S}$, where $\hbar$ is the reduced Planck's constant, J is the switching current density, e is the electron charge, t_f is the free layer thickness, and M_S is the saturation magnetization.

While STT-MTJ offers significant advantages in terms of read energy and speed, a significant incubation delay due to the pre-switching oscillation [131, 132] incurs high switching energy. Consequently, Spin Hall Effect (SHE) and Rashba effect are investigated to achieve an alternative low power switching approach [133, 134, 135]. Recently, SHE-MTJ is introduced as an alternative for 2-terminal MTJs, which provides separate paths for read and write operations, while expending significantly less switching energy [136, 13].

2.2 Spin Hall Effect (SHE)-based Switching Approach

As mentioned, spin-polarized currents can be utilized to generate the torque required for switching the magnetization directions of the free layer in MTJs. In [133], Liu et al. have shown that passing a charge current (I_c) through a heavy metal (HM) such as β-tantalum can generate a spin-polarized current (I_s) using the spin Hall Effect (SHE). This can switch the magnetization direction of the free layer in an MTJ with in-plane magnetic anisotropy.

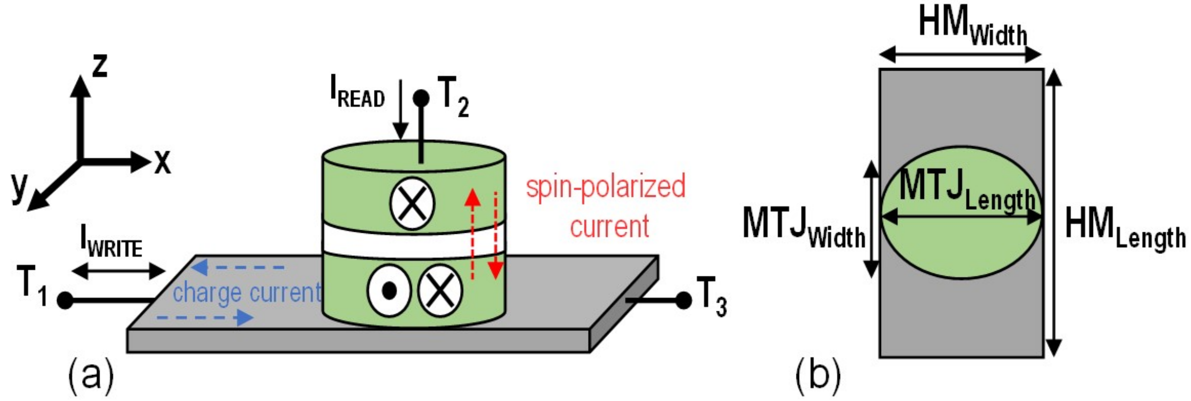


Figure 2.3: (a) SHE-MTJ vertical structure. Positive current along $+x$ induces a spin injection current $+z$ direction. The injected spin current produces the required spin torque for aligning the magnetic direction of the free layer in $+y$ directions, and vice versa. (b) SHE-MTJ Top view. [2].

In [13], Manipatruni et al. have provided the physical equations of the three-terminal SHE-MTJ device behavior. Figure 3.1 shows the structure of the SHE-MTJ device, in which the magnetic orientation of the free layer changes by passing a charge current through a heavy metal (HM). MTJ free layer is directly connected to HM which is normally made of β-tantalum [133], β-tungsten [135]. The spin-orbit coupling in HM deflects the electrons with different spins in opposite directions, which results in a spin injection current (I_s) transverse to the applied charge current (I_c). The injected current produces the required spin torque for aligning the magnetic direction of the free layer. The ratio of the generated spin current to the applied charge current is defined as below [13]:

$$SHIE = \frac{I_s}{I_c} = \frac{\pi.w_{MTJ}.l_{MTJ}}{4.t_{HM}.w_{HM}}\theta_{SHE}\left[1 - sech(\frac{t_{HM}}{\lambda_{sf}})\right] \tag{2.11}$$

where w_{MTJ} is the width of the MTJ, l_{MTJ} is the length of the MTJ, t_{HM} is thickness of the HM, w_{HM} is the width of the HM, λ_{sf} is the spin flip length in HM and θ_{SHE} is the spin Hall angle. The spin hall injection efficiency (SHIE) value is normally greater than one. Therefore, SHE-MTJs can achieve equivalent switching delays with lower write current amplitudes compared to STT-MTJs, resulting in lower power consumption for write operation. The critical spin current required for switching the free layer magnetization orientation is expressed by Equation 2.12 [137]. Thus, the SHE-MTJ's critical charge current (I_C) can be calculated using Equations 2.11 and 2.12.

$$I_s = 2\alpha e M_S V (H_k + 2\pi M_S)/\bar{h} \tag{2.12}$$

The relation between the switching time (τ_{SHE}) and the applied charge current (I_{SHE}) is shown in 2.13, in which v_c is the critical switching voltage, τ_0 is the characteristic time, and R_{HM} is the HM resistance, which are given by 2.14, 2.15, and 3.1, respectively [13]:

$$\tau_{SHE} = [\tau_0 ln(\pi/2\theta_0)]/[(\frac{R_{HM} I_{SHE}}{v_c}) - 1] \tag{2.13}$$

$$v_c = 8\rho I_C [\pi \theta_{SHE} l_{HM} (1 - sech(\frac{t_{HM}}{\lambda_{sf}})] \tag{2.14}$$

$$\tau_0 = M_S V_{HM} e / I_C P \mu_B \tag{2.15}$$

$$R_{HM} = \frac{\rho_{HM}.l_{HM}}{w_{HM}.t_{HM}} \tag{2.16}$$

where θ_0 is the effect of stochastic variation, l_{HM} is the length of the HM, and I_C is the critical charge current for spin-torque induced switching. The magnetization dynamics of the free layer in SHE-MTJ device can be captured by the modified Landau-Lifshitz-Gilbert (LLG) equation [138]:

$$(1 + \alpha^2).\frac{d\hat{m}}{dt} = -\gamma \hat{m} \times \vec{H} - \alpha\gamma(\hat{m} \times \hat{m} \times \vec{H}) + \frac{\hat{m} \times \vec{I_s} \times \hat{m}}{qN} + \frac{\alpha(\hat{m} \times \vec{I_s})}{qN} \tag{2.17}$$

where $N = M_S V/\mu_B$ is the total number of the spins in the volume (V) of free layer nanomagnet, in which μ_B is the Bohr magneton.

Figure A.4 shows the block diagram of an approach proposed by authors in [2, 107] to model the behavior of STT-MTJ and SHE-MTJ devices, in which a Verilog-AMS model is developed using the aforementioned equations. Then, the model is leveraged in SPICE circuit simulator to design hybrid CMOS/spin-based circuits and validate their functionality using experimental parameters such as the ones listed in Table 2.1.

Table 2.1: Parameters of STT/SHE-MTJ devices [2].

Parameter	Description	Value
HM Volume	$HM_{Length} \times HM_{Width} \times HM_{Thickness}$	100×60×3 nm^3
MTJ Area	$MTJ_{Length} \times HM_{Width} \times \pi/4$	60×30×$\pi/4$ nm^2
MTJ Area	Reference MTJ Surface	50×25×$\pi/4$ nm^3
$I_{\text{C-SHE}}$	SHE-MTJ Critical Curren	108 μA
$I_{\text{P-AP}}$	STT-MTJ Critical Current for P to AP Switching	37 μA
$I_{\text{AP-P}}$	STT-MTJ Critical Current for AP to P Switching	18 μA
θ_{SHE}	Spin Hall Angle	0.3
ρ_{HM}	Resistivity	200 $\mu\Omega.cm$
ϕ	Potential Barrier Height	0.4 V
t_{ox}	Thickness of oxide barrier	0.85 nm
α	Gilbert Damping factor	0.007
M_s	Saturation magnetization	200 7.8e5 A.m^{-1}
μ_B	Bohr Magneton	9.27 e-24 J.T^{-1}
P	Spin Polarization	0.52
γ	Gyromagnetic Ratio	1.76e7 (Oe.s)$^{-1}$
R_{AP}, R_P	MTJ Resistances	2.8 KΩ, 5.6 KΩ
R_P	Reference MTJ Resistance	4.12 KΩ
TMR_0	TMR ratio	100%
H_k	Anisotropy Field	80 Oe
μ_0	Permeability of Free Space	1.25663e-6 T.m/A
θ_{SHE}	Spin Hall Angle	0.3
ρ_{HM}	HM Resistivity	200 $\mu\Omega$.cm
ϕ	Potential Barrier Height	0.4 V
Λ_{sf}	Spin Flip Length	1.5nm
e	Electric charge	1.602e-19 C
$\hbar$	Reduced Planck's Constant	6.626e-34/2π J.s

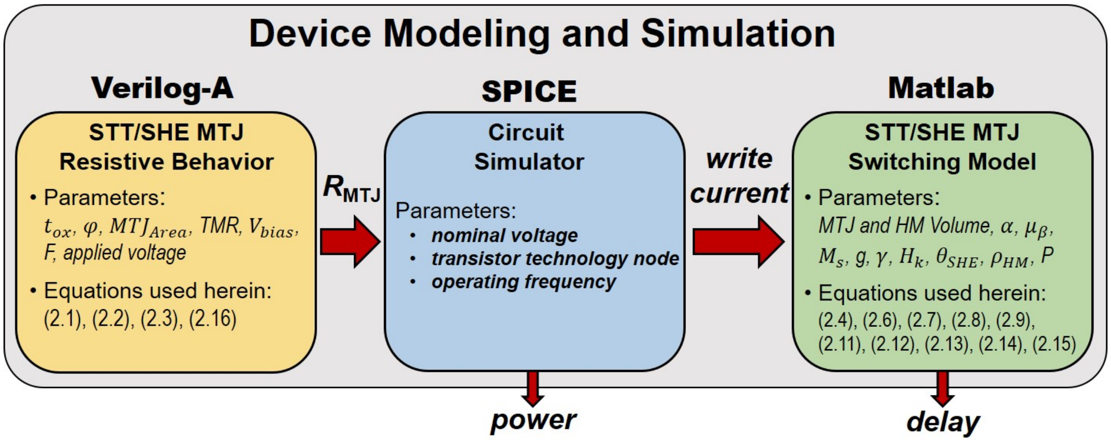


Figure 2.4: Modeling and simulation process of STT/SHE MTJ devices [2].

Figures 4.10 (a) and (b) show the CMOS-based bitcell of the 2-terminal STT-MTJ and SHE-MTJ, respectively. In SHE-MTJ device, the spin current can be significantly larger than the applied charge current. Therefore, the transistor utilized in the bitcell of the 2-terminal MTJ should be larger than that of the SHE-MTJ to be able to provide equal switching delay. Thus, although SHE-MTJ bitcell requires two MOS transistors, its integration density is comparable to that of the 2-terminal MTJs. Increasing the transistor size in 2-terminal MTJs may also impacts the reliability of tunneling oxide barrier, which is improved in 3-terminal SHE-MTJ devices, since the current does not flow through it during the write operation [124].

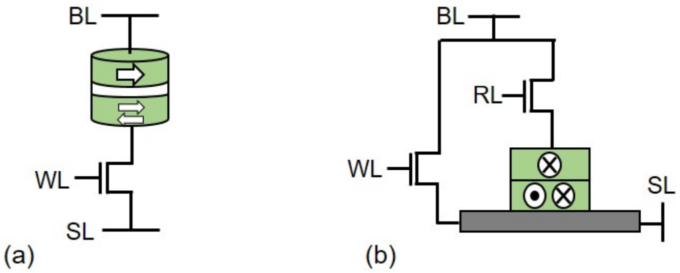


Figure 2.5: (a) 2-terminal MTJ bitcell, (b) SHE-MTJ bitcell [2].

CHAPTER 3: MAGNETIC RANDOM ACCESS MEMORY BASED LOOK-UP TABLE (LUT) CIRCUIT DESIGN[1][2]

A Look-Up Table (LUT) circuit is the building block of reconfigurable computing fabrics, which includes a $2^m \times 1$ memory block to store the configuration data of an m-input Boolean logic function. Currently, static random access memory (SRAM)-based LUTs are primary constituents for logic realization in most reconfigurable fabrics. However, SRAM's drawbacks such as high static power consumption, volatility, and restricted logic density [139, 140] have motivated exploration of alternative LUT designs. One of the introduced alternatives is based on non-volatile flash-based LUTs, however it targets a niche market due to their low reconfiguration endurance [70]. Higher endurance non-volatile LUTs can be enabled by emerging resistive technologies, such as spintronic storage elements [72, 73, 74, 75, 76, 77, 78], resistive random access memory (RRAM) [79, 80, 81, 82], and phase change memory (PCM) [83, 84]. Although PCM can offer non-volatility, its considerable reconfiguration power and high write latency can significantly exceed that of an SRAM LUT.

Spintronic devices offer non-volatility, near-zero static power, and high integration density [68, 67]. Two of the spin-based devices, which are previously proposed for use in reconfigurable fabrics are magnetic tunnel junctions (MTJs) [76, 75, 78, 2, 3, 85] and domain wall (DW)-based racetrack memory (RM) [73, 74]. RM is effective for non-volatility and area density, although previous designs can incur significant delay and energy cost due to excessive shift activities to configure the implemented logic function. Hence, MTJ-based LUTs are proposed herein to be utilized within the reconfigurable fabrics to implement various logic functions as a runtime adaptable fabric under middleware control. Moreover, Radiation immunity of MTJ devices decrease the susceptibility of

[1] © 2016 IEEE. Part of this chapter is reprinted, with permission, from [3].

[2] © 2017 IEEE. Part of this chapter is reprinted, with permission, from [2].

the design to radiation-induced errors [4], as will be described in this chapter.

Three types of energy consumption profiles can be identified in FPGA LUTs. First, an initial write energy consumption incurred at LUT configuration time. Second, the LUTs comprising active logic paths will consume read energy, which may constitute only certain sub areas within high gate equivalent capacity of contemporary FPGA chips. Third, the standby energy consumed by the remaining significant quantity of the LUTs comprising the fabric that may be inactive. It is not possible to power-gate LUT islands, as they must retain the stored configuration. It has been estimated in [75] that if the combined effect of these three modes can be mitigated with a suitable SRAM alternative, then typical power consumption can be reduced up to 81% under representative applications based on measurements of fabricated devices. In [75], Suzuki et al. have fabricated a nonvolatile FPGA with 3000 6-input STT-MTJ based LUTs under 90nm CMOS and 75nm perpendicular MTJ technologies. They have utilized the LUT designs introduced in [76, 78], and in addition to the mentioned energy savings they also achieved 56% area reduction. Herein, we will study two of the MTJ-based LUT designs developed by the author that can realize even more power reduction.

3.1 Spin Transfer Torque (STT)-Magnetic Tunnel Junction (MTJ)-based LUT Circuits

In this section, a 4-input STT-MTJ LUT [3] is introduced which consists of read and write circuits as shown in Figure 4.1. The write circuit includes two transmission gates (TGs) which provide the desired charge current for STT switching [108], while the read circuit is comprised of a pre-charge Sense Amplifier (SA) [141, 142], a TG-based Multiplexer (MUX), and a reference tree. Each MTJ cell of LUT could be accessed according to the input signals, A, B, C, and D, through MUX which employs TGs instead of Pass Transistors (PTs). TGs have near optimal full-swing switching behavior which results in less delay. In addition, TG-based circuits are more resilient to process variation comparing to PT-based designs [143, 108].

The reference tree in read circuit is designed to provide SA with required reference resistance to properly sense each MTJ cell state. Reference tree consists of four TGs in series configuration to compensate for the select tree active resistance. Reference MTJ resistance is designed in a manner such that its value in parallel configuration is between low resistance (R_P) and high resistance (R_{AP}) of the LUT MTJ cells as shown in equation below:

$$R_{P-referenceMTJ} \cong \frac{1}{2}(R_{AP-LUTMTJ} + R_{P-LUTMTJ}) \tag{3.1}$$

In [144] the first prototype of a two input MTJ-based LUT is simulated. It contains four MTJs to store data, and a separate SA and write circuit for each MTJ which lead to significant area overhead and power consumption. In [78], Suzuki et al. has proposed an optimized STT-MTJ based LUT. They reported a 44% reduction in active power, for a 4-input XOR operation, comparing to the LUT designed in [144]. They employed a single SA for the whole LUT circuit instead of using one for each memory cell which results in area and active power reduction. Herein, the proposed STT-MTJ based LUT circuit is implemented using both TG-based and PT-based select and reference trees. The performence of the developed STT-MTJ LUT circuit is compared with the above mentioned MTJ-based LUTs, as listed in Table 3.1. The proposed STT-MTJ LUT provides high speed and ultra-low power circuits with improved power-delay product (PDP) values shown in seventh row of the table. Furthermore, TG-based STT-LUT exhibits least PDP value while it leverages larger number of MOS transistors comparing to PT-based STT-LUT which is the optimum choice from the area efficiency point of view.

In order to evaluate the scalability of the STT-MTJ LUT circuit, PDP values are calculated for 2-input to 6-input STT-MTJ LUTs, considering the worst case scenario, i.e. MTJ state is zero. Figure 3.2 exhibits that PDP and number of LUT inputs are linearly proportional with a low slope which validates the STT-MTJ LUT scalability. This capability led to the proposition of a 4-input adaptive STT-MTJ LUT (A-LUT), as shown in Figure 3.3, which is compatible with the Altera's adaptive LUT structure [145].

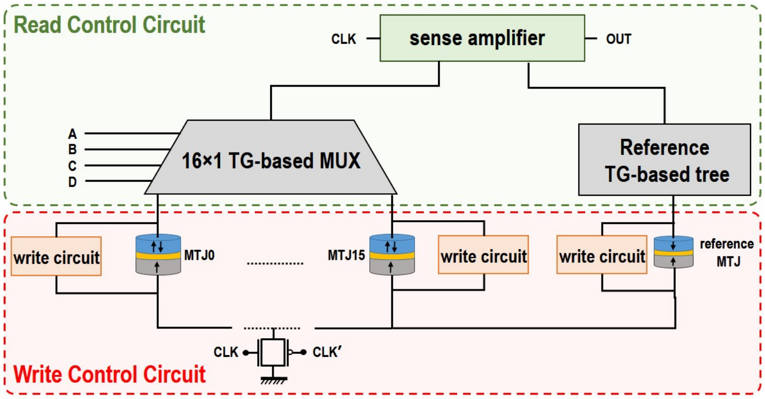


Figure 3.1: A 4-input STT-MTJ LUT functional diagram [3].

Table 3.1: Performance comparison for 4-input NAND operation [3].

Features	[144]	[78]	**PT based STT-LUT**	**TG based STT-LUT**
NO. of MTJs	32	36	17	17
NO. of MOSs	154	74	59	112
Delay$^{(\dagger)}$ (ps)	88	81	94	83
Active Power$^{(*)}$ (μW)	13.40	7.58	4.30	4.27
PDP ($ps\times$ μW)	1179.2	613.98	404.20	354.41
Standby Power	0	0	0	0
PDP [144]	—	48%	65.7%	70%
Improvement [78]	—	—	34%	42%

(†) Worst case delay, switching delay is not included.

(∗) Average power dissipation, switching power is not included.

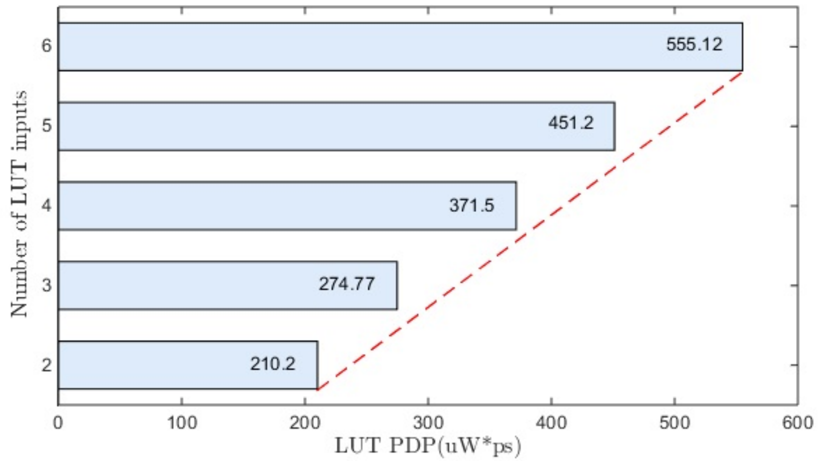


Figure 3.2: PDP growth of STT-MTJ LUT in terms of input widths [3].

Table 3.2: Configuration specifications and MTJ usage for 2-input to 4-input LUT organization [3].

	S21	S22	S23	S24	S31	S32	S4	RS2	RS3	RS4	bitstream	MTJs	Description
mode 0	1	0	0	0	0	0	0	1	0	0	**10’h204**	0-3	2-input LUT
mode 1	0	1	0	0	0	0	0	1	0	0	**10’h104**	4-7	2-input LUT
mode 2	0	0	1	0	0	0	0	1	0	0	**10’h84**	8-11	2-input LUT
mode 3	0	0	0	1	0	0	0	1	0	0	**10’h44**	12-15	2-input LUT
mode 4	0	0	0	0	1	0	0	0	1	0	**10’h22**	0-7	3-input LUT
mode 5	0	0	0	0	0	1	0	0	1	0	**10’h12**	8-15	3-input LUT
mode 6	0	0	0	0	0	0	1	0	0	1	**10’h9**	0-15	4-input LUT

3.1.1 Adaptive STT-MTJ LUT Circuit

The proposed 4-input A-LUT could be configured to operate as different LUTs in seven independent modes: four 2-input STT-MTJ LUTs, two 3-input STT-MTJ LUTs, and one 4-input STT-MTJ LUT. Output of each configuration is individually connected to SA through a mode selector which includes PTs to choose between different operational modes, described in Table 3.2. For example, $bitstream = 10'h104$ configures A-LUT to operate as a 2-input STT-MTJ LUT based on the logic function stored in MTJ4 to MTJ7.

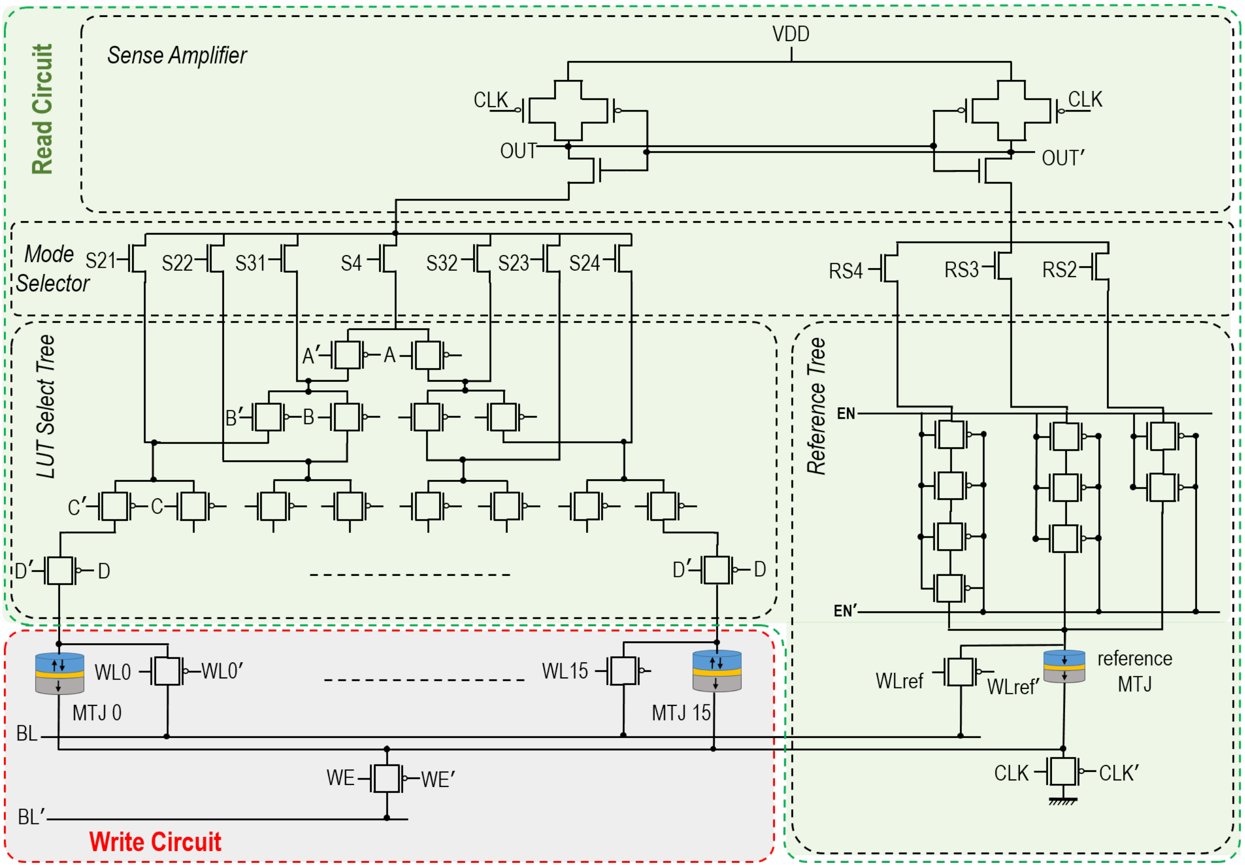


Figure 3.3: The circuit view of A-LUT schematic. [3].

The reference tree of an n-input STT-LUT could be implemented by n TGs and a reference MTJ in series configuration, which provides a resistance equal to $R_{ReferenceTree} = n.R_{TG} + R_{P-referenceMTJ}$. Thus, different number of LUT inputs, only affects the number of TGs which must be utilized in reference tree, and modification to the dimensions of the reference tree MTJ is not required to keep the optimized sensing behavior of SA. Hence, the A-LUT reference tree includes three different branches in parallel configuration that are serially connected to a single MTJ. As shown in Figure 3.3, each of the branches contains two, three, and four TGs which are used for 2-input, 3-input, and 4-input A-LUT configurations, respectively. Figure 3.4 shows the layout of the A-LUT which occupies a cell area of $13.5\mu m \times 15.75\mu m$ in 90nm process. A five metal layer design is depicted. The MTJ cell has a vertical structure which could be readily integrated at the backend process of CMOS fabrication.

The proposed A-LUT circuit is examined using SPICE simulation in 90nm technology. Figure 3.5 elaborates the functionality of the proposed A-LUT for a 4-input NAND operation when ABCD= “1111” and ABCD= “0000” inputs are applied, respectively. The former set of inputs selects $MTJ15$ which has a parallel configuration that denotes logic “0”, while the latter input selects MTJ0 with anti-parallel configuration representing logic “1”. Herein, mode selector’s bitstream is equal to $10'h9$, which selects the sixth mode, i.e. A-LUT fuctioning as 4-input STT-MTJ LUT.

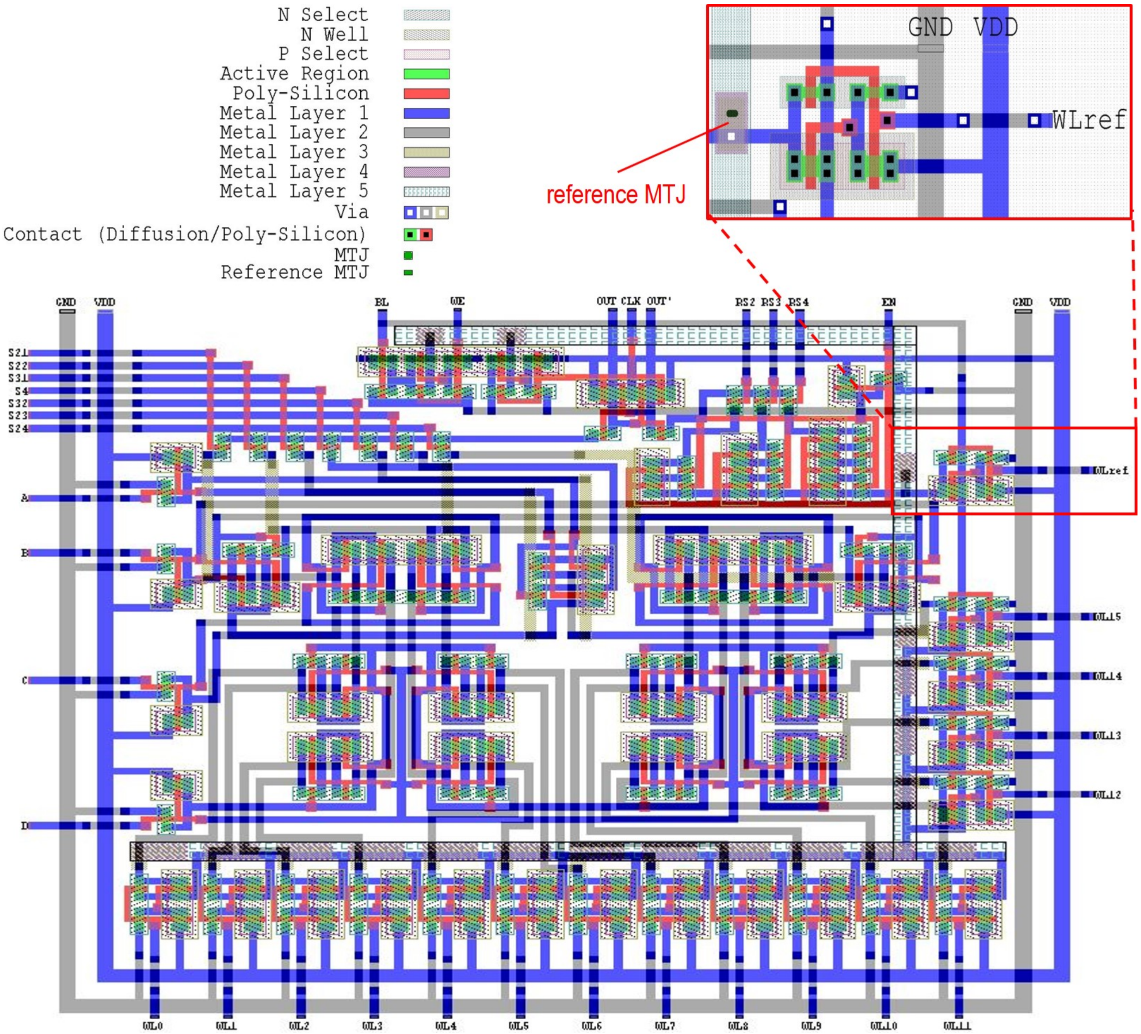


Figure 3.4: A $13.5\mu m \times 15.75\mu m$ 4-input A-LUT layout [3].

Table 3.3: PDP values for STT-MTJ LUT and A-LUT designs *($ps\times\mu W$)* [3].

Boolean Function Inputs	**8-input STT-MTJ LUT**	**8-input A-LUT**		
	Power × Delay $(ps \times \mu W)$ 5.934×138.14	Power (μW)	Delay (ps)	PDP $(ps \times \mu W)$
2	819.72	3.826	70.51	269.8
3	819.72	4.260	83	353.58
4	819.72	4.734	94.92	449.35
5	819.72	5.138	107.05	549.98
6	819.72	5.571	120.98	673.97
7	819.72	5.960	134	798.64
8	819.72	6.307	146.84	926.1

Herein, a comprehensive PDP analysis is performed to evaluate the performance of A-LUT. Therefore, an 8-input A-LUT and 8-input STT-MTJ LUT are examined to implement 2-input to 8-input Boolean logic functions. The PDP results are extracted for a worst case NAND operation utilizing 1.2V nominal voltage (VDD) and 1GHz circuit clock (CLK) frequency. As listed in Table 3.3, an n-input A-LUT PDP is smaller than n-input STT-MTJ LUT PDP, when performing 2-input to *(n-1)*-input Boolean functions.

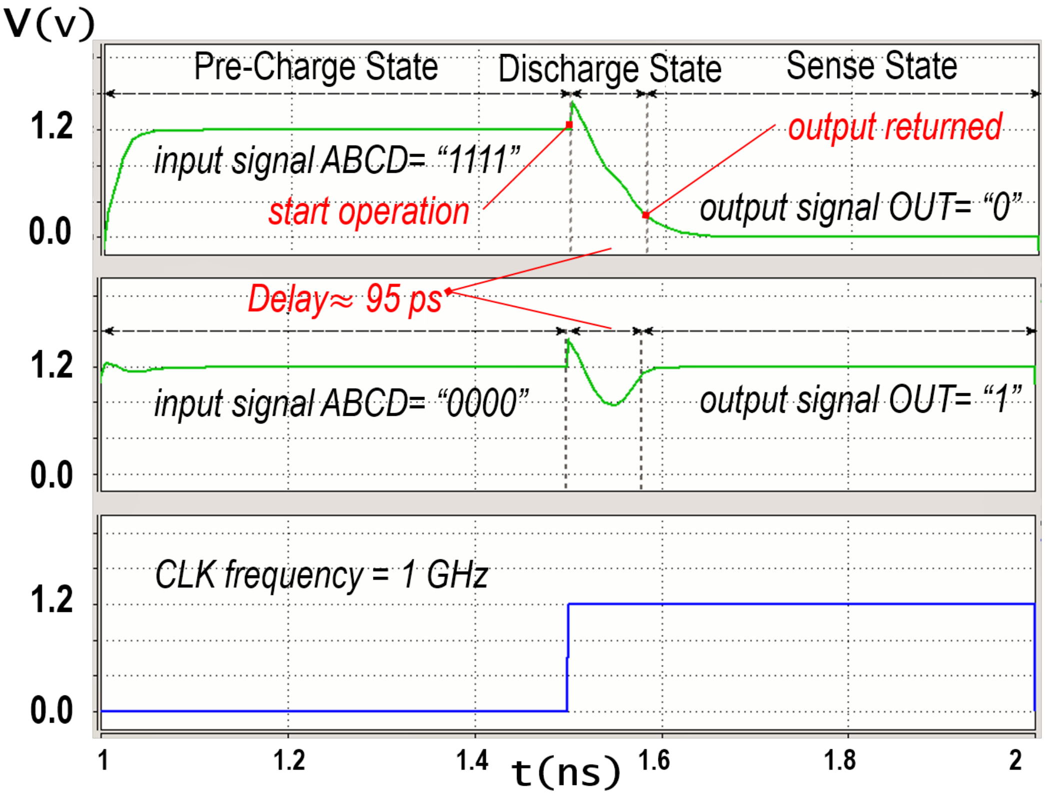


Figure 3.5: Transient response of A-LUT for 4-input NAND operation for ABCD= “1111” (top), and ABCD= “0000” (middle) [3].

Despite the mentioned advantages of conventional STT-MTJ devices, their main challenge is relatively high delay and power consumption for write operation. Moreover, two-terminal MTJ devices can experience occasional read/write disturbances due to sharing a common path for read and write operations. Recently, 3-terminal spin Hall effect (SHE)-based MTJ has been introduced as an alternative for conventional 2-terminal MTJs. SHE-MTJ provides separate paths for read and write operations, while expending significantly less switching energy [133, 13]. In next section, we develop a nonvolatile LUT circuit using SHE-MTJ devices, and provide a detailed comparison between the SHE-MTJ-based LUT (SHE-LUT) and 2-terminal MTJ-based LUTs including the reconfiguration energy consumption and delay.

3.2 Spin Hall Effect (SHE)-Magnetic Tunnel Junction (MTJ)-based LUT Circuits

In this section, a non-volatile LUT circuit is developed based on the SHE-MTJ devices. SHE-LUT structure includes two main parts: write circuit and read circuit. Designing the read and write circuits requires considering important details which can significantly influence the energy consumption and delay of the LUT circuit.

Herein, we have utilized SHE-MTJ device as a storage element in the LUT circuit as shown in Figure 3.6. In general, data is stored in resistive memory cells in form of different resistance levels, e.g. high resistance state stands for logic 1 and vice versa. Therefore, a sense amplifier (SA) is required to distinguish the resistive state of the memory cell. In [141], Zhao et al. studied various SAs which could be leveraged for sensing the magnetic configuration of the MTJs. They have proposed a Pre-Charge Sense Amplifier (PCSA) consisting of seven MOS transistors and a reference MTJ, which could provide a low power and high speed read operation.

Figure 3.6 shows the PCSA circuit which includes four PMOS transistors connected to the VDD, two NMOS transistors which connects the PMOS transistors to the select trees and data storage cells, and one NMOS transistor which connects the circuit to ground (GND). Moreover, a TG-based reference tree including four TGs in series configuration is utilized in our designs to compensate for the select tree resistance. Reference MTJ dimensions are designed in a manner such that its resistance value in parallel configuration is between low resistance, R_{Low}, and high resistance, R_{High}, of the SHE-based MTJ cells as shown in Figure 3.7.

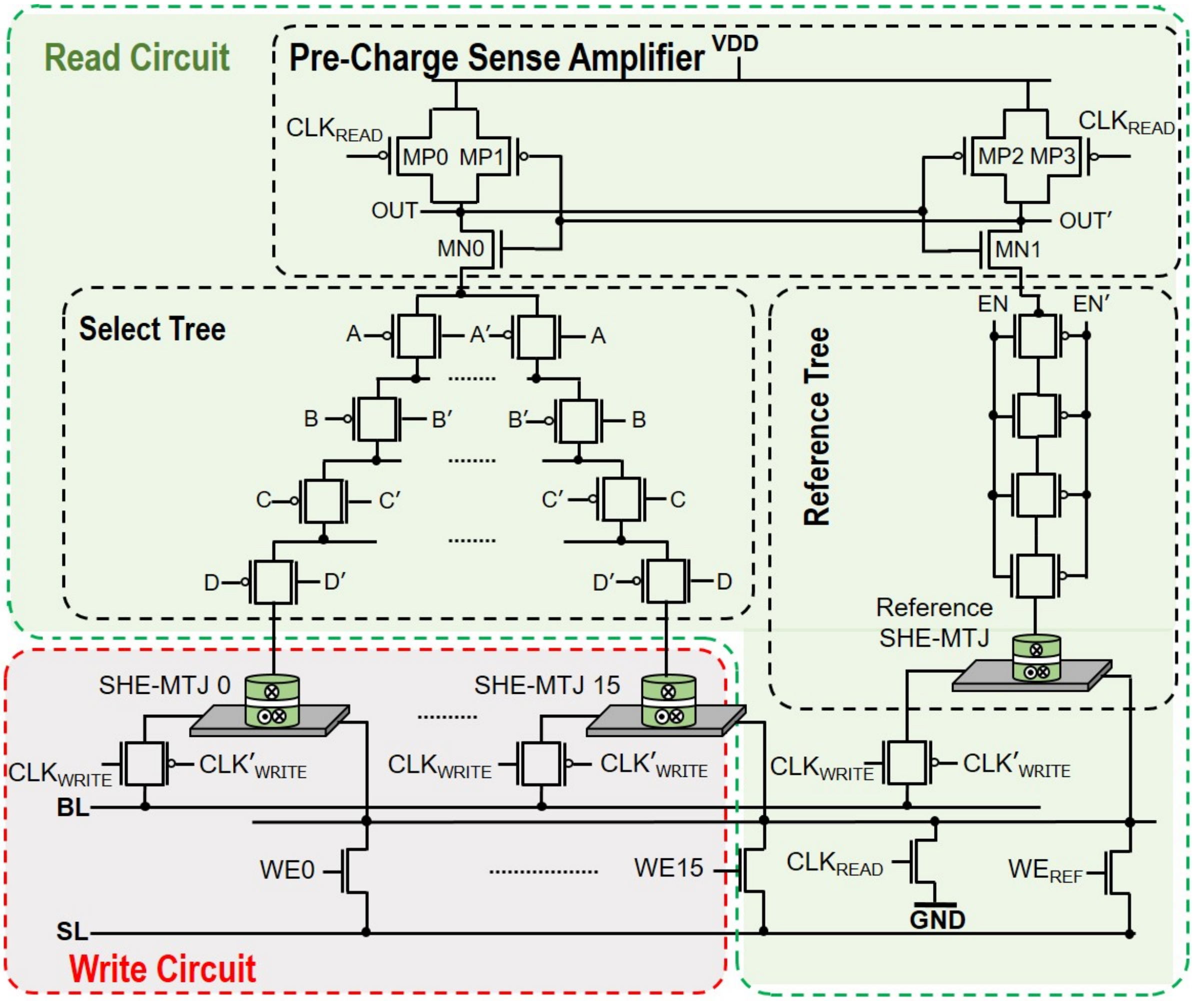


Figure 3.6: The circuit level design of the proposed SHE-LUT [2].

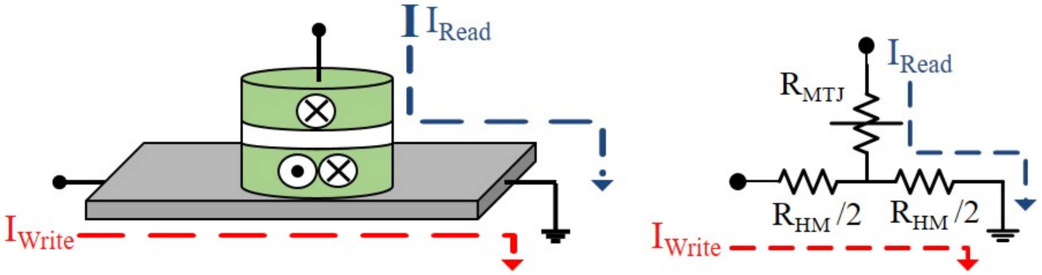


Figure 3.7: SHE-MTJ read and write path equivalent resistances [2].

Sensing with PCSA requires two operating phases which could be performed in a single clock (CLK) period: *pre-charge phase* and *discharge phase*. During the pre-charge phase, CLK signal is equal to zero, therefore MP0 and MP3 transistors, shown in Figure 3.6, are ON and the drains of the MN0 (OUT) and the MN1 (OUT') transistors are charged to VDD. In the discharge phase, CLK is equal to VDD and all the PMOS transistors are OFF. Consequently, the voltage source (VDD) is disconnected from the circuit and the pre-charged nodes, i.e. OUT and OUT', begins to discharge. The discharge speed in each of the branches of the PCSA is different due to the difference between the resistances of the resistive storage elements, and the reference SHE-MTJ. For example, assume that SHE-MTJ0 with AP configuration is the storage element that is being sensed. Since it has higher resistance than the reference MTJ, the branch connected to it discharges slower than the reference SHE-MTJ branch, thus the voltage drops faster in OUT' node. Since OUT' is connected to the gate of the MP1 transistor, the voltage drop results in a voltage difference between source and gate of the MP1 transistors that is higher than threshold voltage. Consequently, MP1 will be ON and the OUT node which is connected to gate of the MP2 transistor will be charged to VDD. This causes the MP2 transistor to remain OFF, and as the result OUT' node will be completely discharged to GND.

In practice, an external synchronizer circuit can be utilized to ensure that the input signals are synchronous to the local clock signal of the SA, as shown in Figure 3.8. The synchronizer circuit for an n-input LUT includes n flip-flops that samples the inputs at each clock cycle. The pre-charge state of the SA should be sufficiently long to meet the required setup and hold times of the flip flops to avoid metastability. The probability of synchronization failure caused by staying of a flip-flop in the metastable state exponentially decreases with time [146].

In this work, we have utilized a TG in the SHE-MTJ write circuit, as shown in Figure 3.9 (a). TGs are composed of one NMOS and one PMOS transistor, and characterized by their near optimal full-swing switching behavior [108]. TG-based write circuit provides a symmetric switching behavior, i.e. the generated write current amplitude for P to AP switching equals the current amplitude produced for switching from AP to P state. Moreover, TGs are capable of producing a current amplitude larger than the switching critical currents of both 2-terminal MTJ and SHE-MTJ devices, which are listed in Table 2.1. The produced current amplitude is sufficiently large to ensure the complete switching of the MTJ devices utilized herein.

Figure 3.9 (b) shows the TG-based write circuit layout view. To address the feasibility of integrating SHE-MTJ with TGs, the three-dimensional (3D) cross-sectional view is provided in Figure 3.9 (c), which depicts the SHE-MTJ integration at the back-end process of CMOS fabrication. The required current for switching the SHE-MTJ passes through the heavy metal structure, which is built in the second metal layer. The MTJ stack is integrated between the second and forth metal layers, and occupies the space for the third via and metal layer as well as the fourth via. Although, TG-based designs necessitate the availability of both CLK and inverse CLK signals, it is common and reasonable to assume access to both signal conditions within typical integrated circuits.

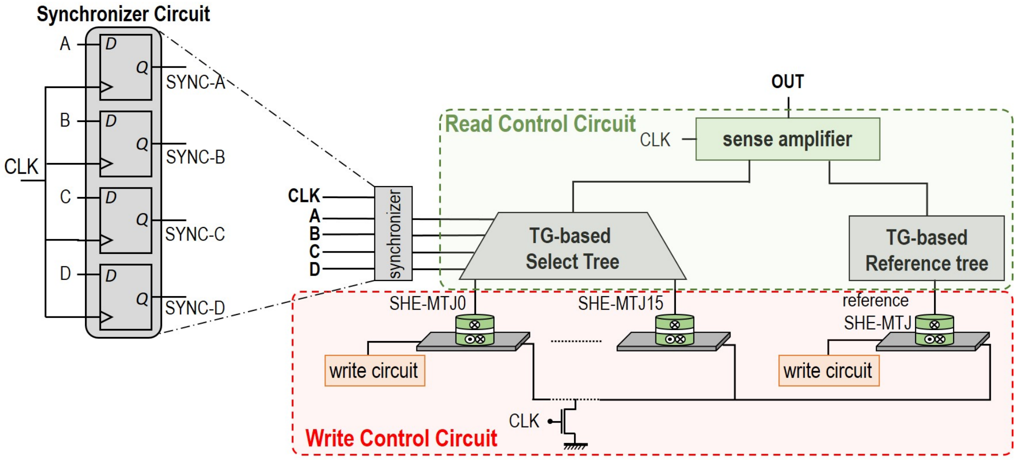


Figure 3.8: Schematic of 4-input SHE-based MTJ-LUT along with an external synchronizer circuit [2].

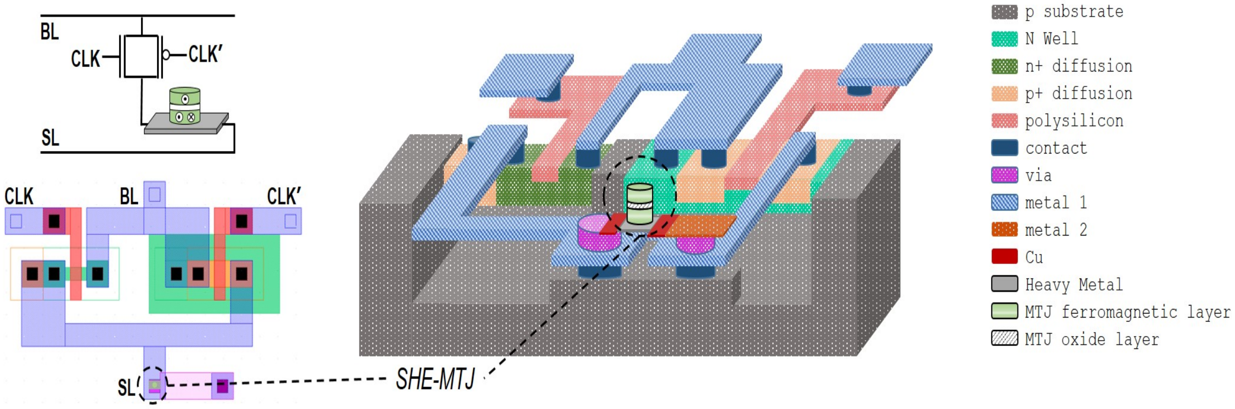


Figure 3.9: (a) Proposed Transmission Gate-based Write Circuits. (b) TG-based write circuit layout view. (c) Three-dimensional (3D) cross-sectional schematic of SHE-MTJ integration at the back-end process of TG fabrication [2].

3.2.1 Fracturable 6-input SHE-MTJ LUT design

The proposed 4-input SHE-LUT circuit can be readily extended to LUT designs with greater number of inputs. Most of the modern FPGAs utilize fracturable 6-input LUTs in their design [147, 148]. These LUTs have six independent inputs and two separated outputs. The fracturable 6-input LUT can implement any six-input Boolean functions, as well as two five-input Boolean functions with common inputs [148]. Herein, we have designed a fracturable 6-input LUT using SHE-MTJ devices, in which two PCSAs and two reference trees are utilized to ensure the independence of the outputs. Five NMOS transistors and two select signals, i.e. S5 and S6, are added to the LUT circuit to control the 5-input and 6-input operation modes of the SHE-based fracturable LUT. The structure of the proposed 6-input SHE-based fracturable LUT is shown in Figure 3.10. It provides significantly higher functional flexibility at the expense of slightly more area and power consumption.

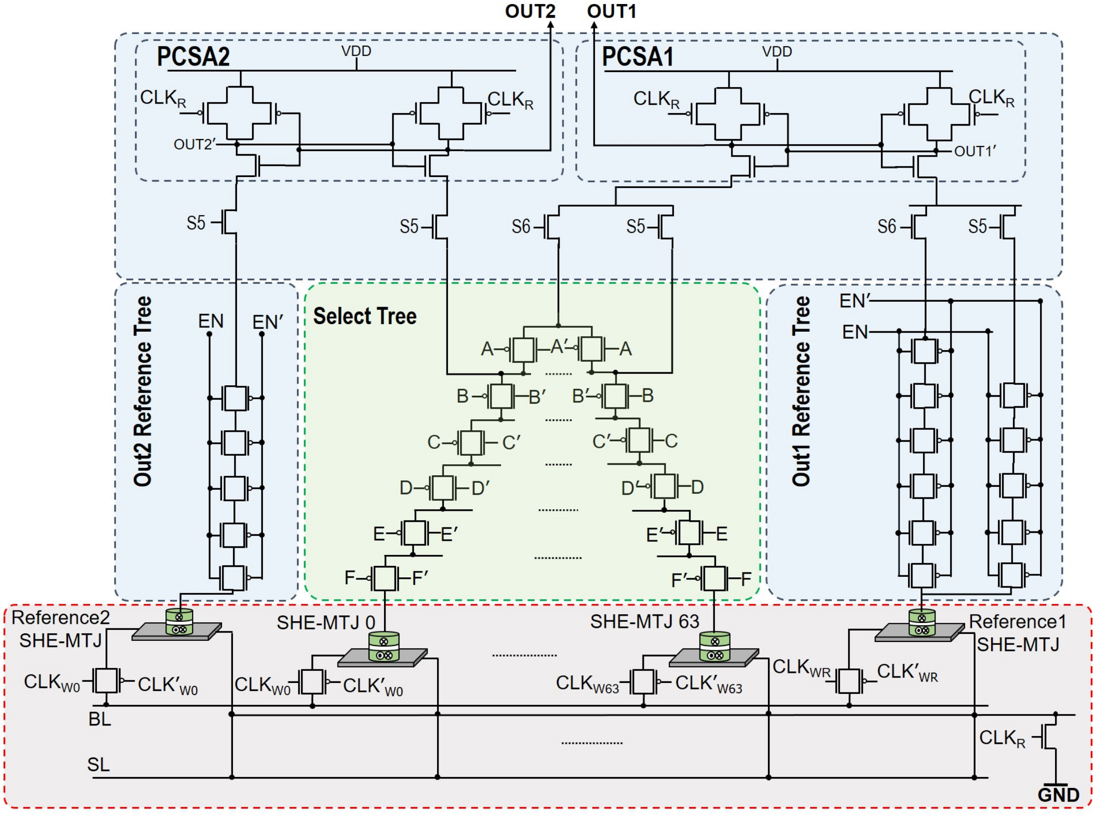


Figure 3.10: The structure of the 6-input SHE-based fracturable LUT [2].

3.2.2 SHE-MTJ LUT Simulation Results

Figure 3.11 exhibits the functionality of the proposed SHE-LUT in the read phase for a 4-input NAND logic operation. The first set of inputs applied is ABCD= 1111, which selects SHE-MTJ15 with P configuration that denotes logic 0. While, the latter input ABCD=0000 selects SHE-MTJ0 with AP configuration representing logic 1. Table 3.4 provides the SHE-LUT power and delay analysis for various input widths, as well as a comparison with a 2-terminal MTJ-based LUT previously proposed by the authors in [3]. Results show the SHE-LUT improvement in terms of power-delay product (PDP), ranging from 1.2% to 6.15%. Simulation results are obtained using SPICE circuit simulator in 90nm CMOS technology model [149].

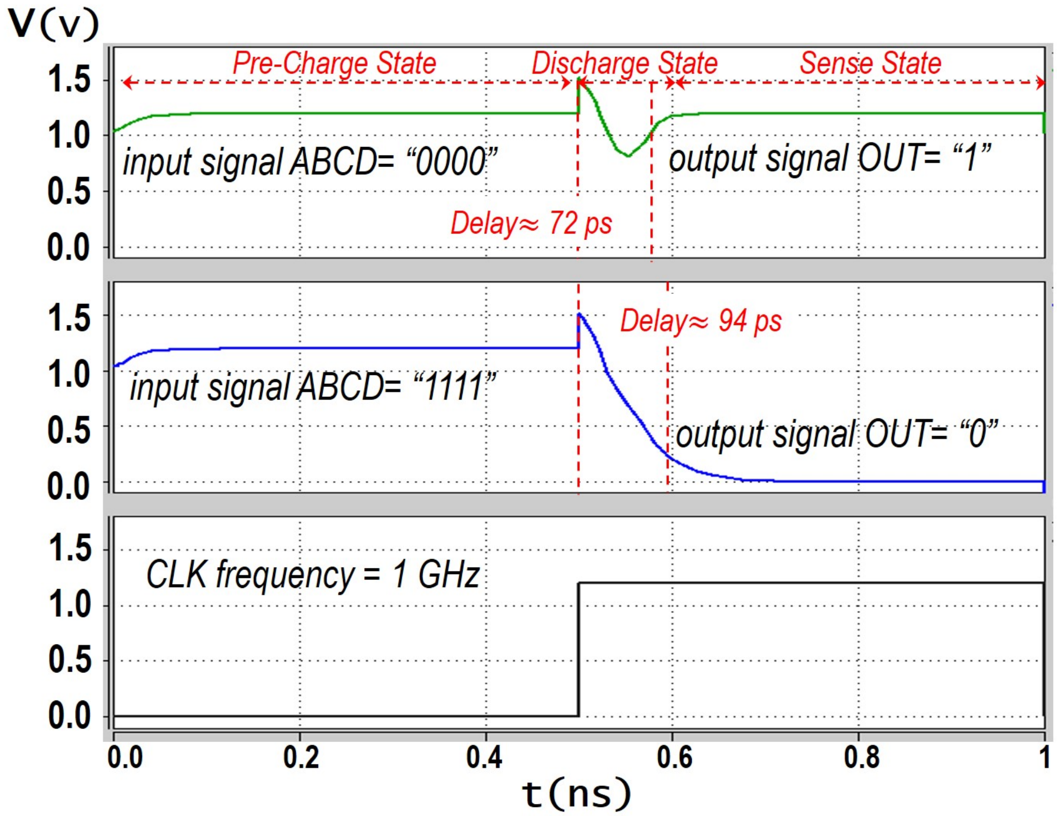


Figure 3.11: Transient response for 4-input NAND operation implemented by SHE-MTJ LUT for ABCD= “1111” (middle), and ABCD= “0000” (top) [2].

Herein, we have provided a comprehensive comparison between the read performances of our proposed 4-input SHE-LUT circuit, and previous performance-efficient 4-input MTJ-LUT designs introduced in [144, 78, 3]. Delay and power consumption for read operations are extracted for input values precipitating worst case condition for a NAND operation utilizing 1.2V nominal voltage (VDD) and 1GHz circuit clock (CLK) frequency. The obtained results are summarized in Table 3.5. SHE-MTJ LUT provides high speed and low energy read operation with improved PDP values listed in the bottom row of Table 3.5.

Table 3.4: Performance Comparison of Proposed SHE-MTJ LUT [2] versus 2-terminal MTJ-based LUT [3] for Various Input Widths [2].

Number of LUT Inputs	PDP = Power (μW) $\times$ Delay(ps)					
	LUT MTJ state = 0			LUT MTJ state = 1		
	STT-MTJ LUT [3]	SHE-MTJ LUT	PDP Improv.	STT-MTJ LUT [3]	SHE-MTJ LUT	PDP Improv.
2	3.3×67	3.17×66	5.94%	3.28×55	3.15×54	5.7%
3	3.74×79	3.6×80	2.52%	3.7×64	3.54×63	5.82%
4	4.3×94	4.01×94.6	6.15%	4.1×73	3.96×72	4.7%
5	4.54×108	4.4×110	1.2%	4.49×82	4.36×81	4.07%
6	4.92×123	4.78×124	2.05%	4.86×92	4.74×91	3.53%

Table 3.5: Performance Comparison for the Read Operation of 4-input MTJ-LUTs [2].

Features	[144]	[78]	[3]	**SHE-MTJ LUT [2]**
# of MTJs	32	36	17	**17**
# of MOSs	154	74	112	**109**
Delay (ps)	88	81	94	**94.6**
Power (μW)	13.4	7.58	4.3	**4.01**
PDP ($\mu W timesps$)	1179.2	613.98	404.2	**379.34**
PDP Improvement	**67.8%**	**38.2%**	**6.15%**	–

We have also examined the reconfiguration operation of SHE-MTJ LUT and conventional 2-terminal MTJ-based LUTs, which involves write operation to switch the state of the MTJs. The STT and SHE switching behaviors are modeled using the relations provided in Chapter 2. Table 4.9 provides a comparison between the reconfiguration operation of a 4-input SHE-LUT and a conventional MTJ-LUT. A 4-input MTJ-LUT includes sixteen MTJs having their magnetization directions aligned in a single reconfiguration operation. As listed in Table 3.6, the proposed SHE-LUT provided at least 20% PDP improvement compared to 2-terminal MTJ LUTs.

Table 3.6: Performance comparison for the Reconfiguration Operation of 4-input MTJ-LUTs Involving 16 MTJs [2].

Features		STT-MTJ LUT	**SHE-MTJ LUT**
Delay (ns)	P to AP	52.8	**31.68**
	Ap to P	53.92	**31.36**
Power (mW)	P to AP	1.16	**1.44**
	AP to P	0.89	**1.45**
PDP ($ns \times mW$)	P to AP	3.83	**2.85**
	AP to P	3.3	**2.84**
Average PDP ($ns \times mW$)		3.565	**2.845**
Average PDP Improvement		**20.1%**	–

To investigate the effect of MTJ scaling on the performance of the MTJ-based LUTs, a comprehensive comparison between a 4-input SHE-MTJ LUT and a 4-input STT-MTJ LUT is provided herein. Figure 3.12 (a) shows the obtained results for the read operation of LUTs including the PDP values for sensing both P and AP states. The performance of the SHE-MTJ LUT and STT-MTJ LUT are comparable for the read operation, while there is a significant difference for the reconfiguration operation, as shown in Figure 3.12 (b). The obtained results exhibit the superiority of the SHE-LUT in term of PDP for different MTJ dimensions. Moreover, LUTs with smaller MTJs have lower PDP values for both read and reconfiguration operations. In read operation, this is mainly achieved due to the increase in the resistance of MTJs by decreasing its dimensions. Since the supply voltage is fixed, higher resistance of MTJs results in lower read current, and consequently lower power consumption. For write operation, although a decrease in the produced write current leads to lower power consumption, it can also results in higher switching delay. This decrease in the switching speed is compensated by the significant decrease in the required switching critical currents. Eventually, smaller MTJ dimensions results in higher switching speed, as well as lower switching power, as shown in Figure 3.12 (b).

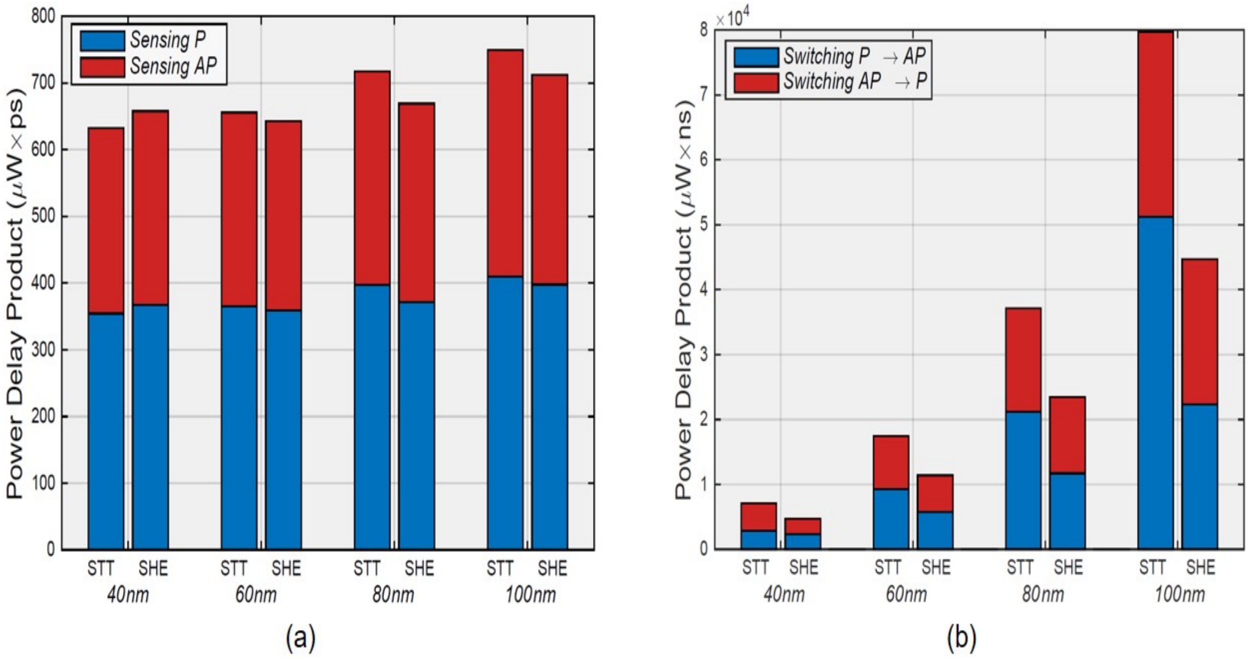


Figure 3.12: The effect of MTJ scaling on SHE-MTJ LUT and STT-MTJ LUT performances. (a) Read operation, and (b) reconfiguration operation [2].

3.2.3 SHE-MTJ based fracturable LUT Simulation Results

As mentioned, 6-input SHE-MTJ based fracturable LUT circuit can operate in two different modes: 5-input and 6-input. In 5-input mode, the fracturable LUT can simultaneously implement two different five-input Boolean functions, as long as they share common inputs. Figure 3.13 shows the functionality of the proposed 6-input SHE-LUT while operating in 5-input mode. The truth table of a 5-input NAND logic operation is stored in the least significant 32 bits of the LUT circuit, while the most significant 32-bits implement a 5-input AND Boolean function. The applied input is ABCDEF= X11111, which selects SHE-MTJ63 and SHE-MTJ31 with AP and P configurations, respectively. Table 3.7 lists the power consumption and delay of the 6-input fracturable LUT read operation for different operating modes in worst case condition. The reconfiguration operation in fracturable LUT is similar to that of the regular LUTs.

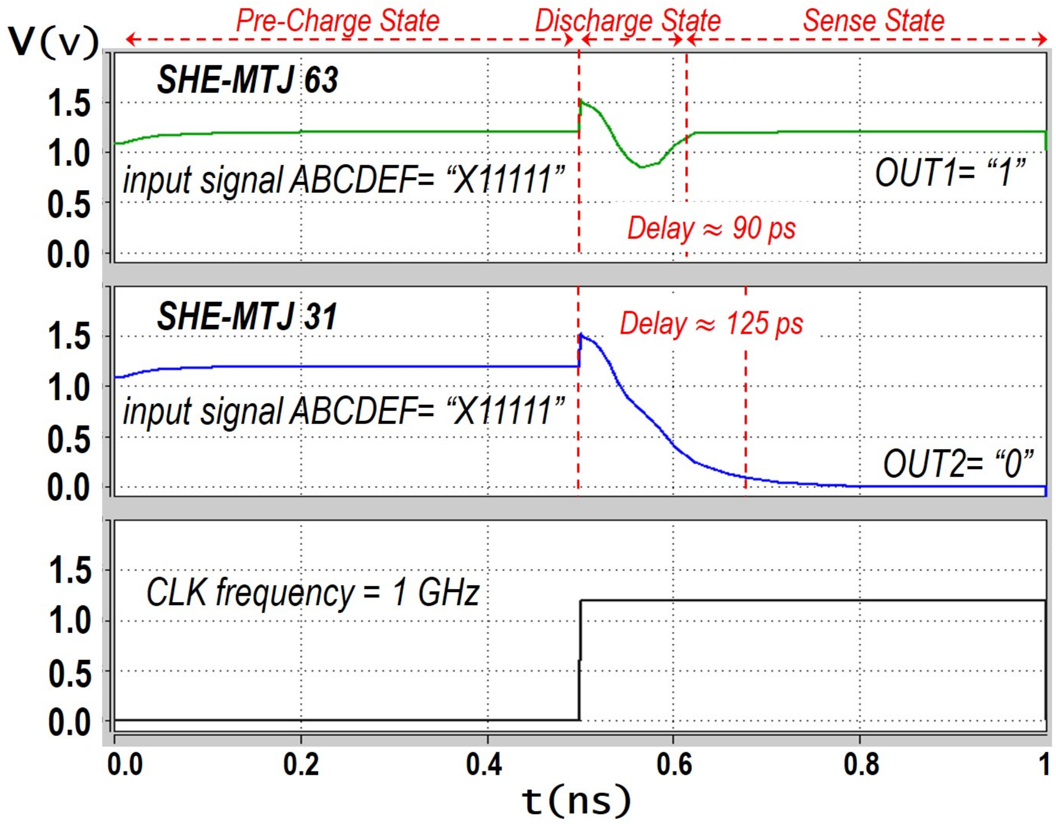


Figure 3.13: Transient response for fracturable 6-input SHE-LUT. 5-input NAND operation for ABCDEF= X11111 (middle), and 5-input AND operation for ABCDEF= X11111 (top) [2].

Table 3.7: Performance comparison for the read operation of the 6-input SHE-MTJ based fracturable LUT [2].

Features		Delay (ps)	Power (μW)
5-input	Regular	110	4.4
	Fracturable	**125**	**9.55**
6-input	Regular	124	4.78
	Fracturable	**152**	**5.85**

3.3 Advances in Clocking Schemes for MTJ-based LUTs

Clocking limitations may have a significant effect on the performance of an MTJ-LUT. The obtained results extracted in previous section are related to isolated read and reconfiguration operations. Since each of the LUT operations have different clocking limitations, more comprehensive clocking and signaling mechanisms are sought to control these operations in an MTJ-LUT. In this section, we have investigated two potential clocking schemes supporting both read and reconfiguration operations for SHE-LUT. First, we consider the use of a single clock signal for both read and write operations. Figure 3.14 shows the control signals, e.g. READ/Write enable signals, as well as the 250MHz clock signal which is utilized for both read and write operations. The clock frequency has been designed to ensure the complete switching of a single SHE-MTJ device. As it was illustrated in previous section, the required time for switching the state of a SHE-MTJ cell is significantly greater than the time needed for read operation. Thus, if a single clock signal is utilized for both operations, its period must be long enough to ensure a complete reconfiguration operation. Although using a single clock reduces the complexity of the design, it incurs an excessive delay in the read operation, which is the predominant operation by several orders of magnitude in reconfigurable fabrics. The clock frequency can be increased by amplifying the write current to reduce the MTJ switching delay. Enlarging the transistors' widths in the write circuit increases the amplitude of the produced write current. Table 3.8 lists the maximum possible operating clock frequencies for SHE-LUT, based on the different dimensions chosen for the transistors in the write circuit. As listed in Table 3.8, increasing the operating clock frequency is achieved at the expense of higher power consumption for reconfiguration operations.

Table 3.8: SHE-MTJ LUT operating clock frequencies based on different dimensions for transistors used in the write circuit [2].

Features		Width/90nm Ratio		

		NMOS= 1X
		PMOS= 2X
			NMOS= 2X
			PMOS= 4X
				NMOS= 4X
				PMOS= 8X

Current Amplitude (μA)	P to AP	150.7	265.4	415.8
	AP to P	151	272	443.8
Switching Delay (ns)	P to AP	1.98	1.05	0.65
	AP to P	1.96	1.02	0.6
Maximum CLK Frequency (MHz)		250	450	750
Reconfiguration Power (μW)	P to AP	90.43	159.2	249.5
	AP to P	90.68	163.2	266.3

Figure 3.15 shows the second approach investigated herein for controlling the functionality of a SHE-LUT circuit, in which two distinct clock signals are utilized for read and write operations. The use of separate clock signal for read operations avoids the excessive delay existed in the previous clocking method caused by being restricted to the write clock frequency. In this approach, the clock frequency for read operation is limited to the sensing delay of the read circuit. For instance, the sensing delay for a 4-input SHE-LUT is approximately equal to 100 picoseconds. Therefore, the clock frequency for read operation can be designed up to 5GHz assuming the 50% duty cycle. Figure 3.15 shows the control signals required for a 4-input SHE-LUT design, in which the clock frequencies for reconfiguration and read operations are equal to 250 MHz and 1GHz, respectively. This significant increase has been made possible by using a differentiated read clock rate that can substantially boost FPGA throughput, due to the prevalence of LUT read operations while performing logic functions.

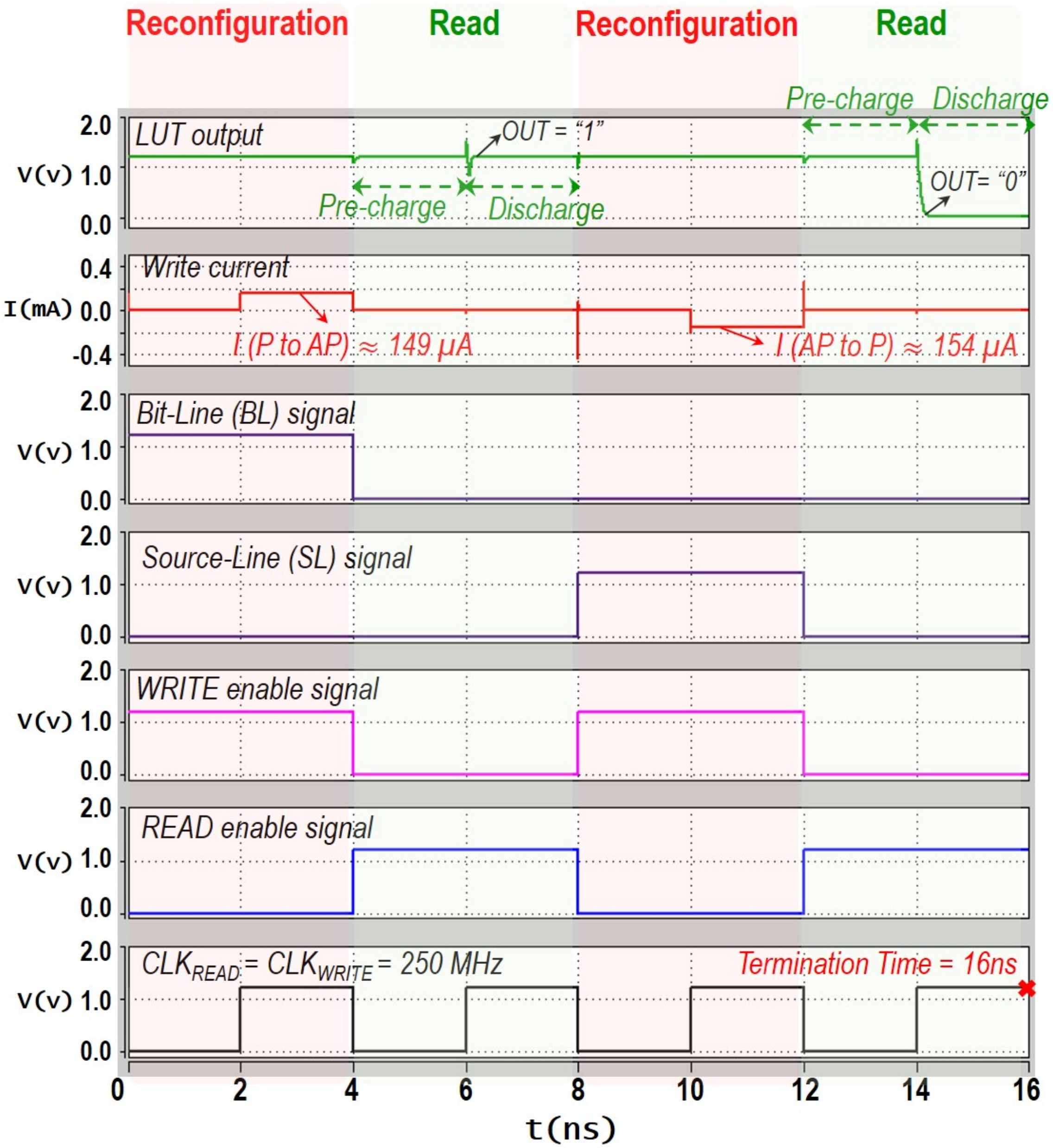


Figure 3.14: SHE-MTJ based LUT functionality using a single clock for both read and reconfiguration operations requiring 16ns termination time [2].

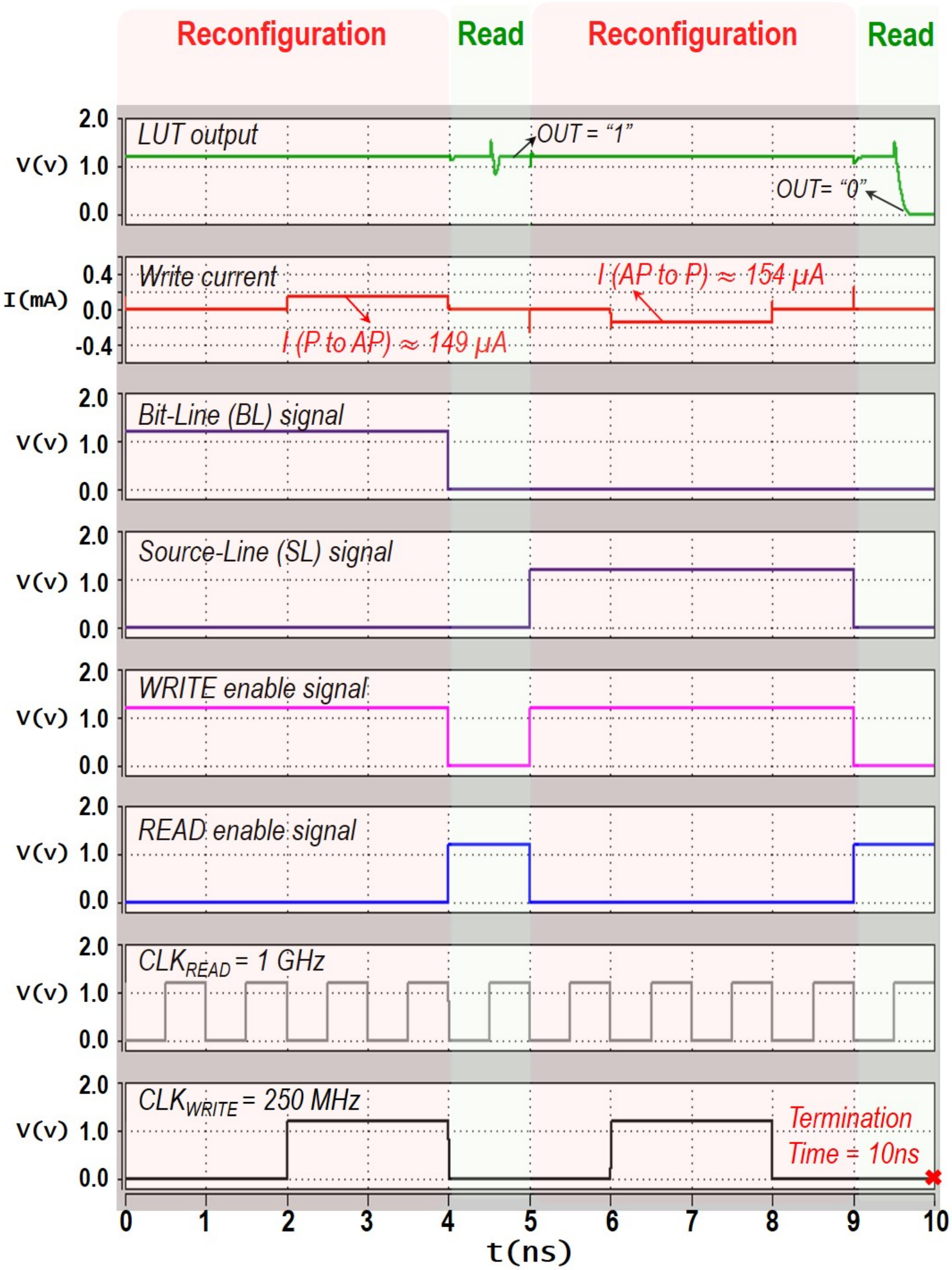


Figure 3.15: SHE-MTJ based LUT functionality using distinct clock signals for read and reconfiguration operations achieving 10ns termination time [2].

3.4 Radiation-hardened Spin Hall Effect (SHE)-Magnetic Random Access Memory (MRAM) based LUT

Radiation-induced soft errors in nanometer-scale electronic circuits are of increasing concern in mission-critical space-based [150], high altitude [151], and terrestrial applications [152]. For instance, space missions which take place in a harsh environment in terms of cosmic radiation particles, temperature, and electromagnetic disturbances have grappled with radiation-induced upsets for many decades of continued device technology scaling [153]. As device dimensions are reduced, the critical charge required to induce a logic state upset causing a soft error has decreased, which is due to a number of compounding physical phenomena. These include the cumulative impact of aggressive voltage scaling which reduces the voltage headroom available to mask errors, and the continued miniaturization of deeply-scaled CMOS-based computing technology [154]. Applications which are especially susceptible due to factors of the mission criticality, the number and density of sensitive devices, and the environmental exposures they endure are autonomous systems utilizing high capacity fine-grained configurable components, such as Field Programming Gate Arrays (FPGAs) [31].

SRAM-based FPGAs, like all CMOS-based fabrics, are susceptible to radiation-induced transient soft faults such as single event upsets (SEUs), which primarily affect SRAM-based storage cells [87]. Thus, research into the design of suitable placements with improved soft error immunity and energy profiles is urgently sought using a number of feasible physical devices including RRAM [80] and MRAM [2, 75]. This trend has been motivated by aggressive CMOS technology scaling in digital circuits has resulted in significant increase in transient fault rates, as well as timing violations due to process variation (PV) that consequently reduces the performance and reliability of the emerging very large scale integrated (VLSI) circuits. For instance, the probability of single upsets, and more realistically, multiple upsets, is projected to increase several fold at sea-level for sub-10nm technology nodes [90, 91]. By the extensions to sub-10nm regimes, error resiliency has become a major challenge for microelectronics industry, particularly mission critical systems, e.g. space and terrestrial applications. The ability of FPGAs to correctly execute the complicated tasks in harsh environments significantly relies on their fault-handling and radiation hardening techniques, such as the design approach proposed in this section.

Leveraging magnetic random access memories (MRAMs) as storage elements within LUT circuits has the potential to significantly increases their radiation immunity due to the radiation hardness characteristic of spin-based devices. However, the access and sensing circuitry for MRAM still requires transistors, and thus is still susceptible to radiation-induced faults. Therefore, circuit-level innovations are sought to achieve immunity to radiation-induced transient faults such as SEUs and double node upsets (DNUs). In recent years, various radiation hardening techniques are investigated to develop SEU-tolerant MRAM-based LUTs [56, 88]. In particular, in [89] authors have proposed a single-event double-node upset tolerant MRAM-based LUT, which provides multiple upset resiliency at the cost of increased read energy and area consumption with baseline efficacy. In this work, we develop a nonvolatile MRAM-based LUT using SHE-MTJ devices, which can tolerate DNUs with improved area, delay, and power consumption.

3.4.1 Fundamentals and modeling of radiation effect on hybrid CMOS/spin based circuits

Among the natural sources of α, β, and γ radioactivity only alpha particles are able to incur transient errors in hybrid CMOS/spin-based circuits due to their high energy [155]. Alpha particles are able to deposit a charge along their track when striking a sensitive node of a circuit. The charge will be transported into the device and collected in the sensitive region [156]. A transient fault is generated if the injected charge (Q_{inj}) exceeds the critical charge (Q_C) of the sensitive node. The

Q_C can be realized by a capacitance and a conduction component as shown below:

$$Q_C = C_N V_{DD} + I_D T_F \tag{3.2}$$

where C_N is the equivalent capacitance of the struck node, VDD is the power supply, I_D is the maximum drain conduction current and T_F is the flipping time of the cell. The computation of T_F requires a 3D device simulation, therefore to simplify the circuit simulation the conduction component of the 3.2 is normally ignored that leads to an insignificant under-estimation [157, 158].

Various approaches are proposed to model the radiation-induced transient fault such as Freeman [159] or diffusion collection [160] models. Herein, we have utilized a double exponential current source to model the radiation effect, which is the most commonly used approach in the literature [161]. The current sources are connected to do sensitive nodes of the circuit, which inject current to the nodes when radiation particles are supposedly strike. The injected current pulse is given by 3.3, in which τ_f and τ_r are falling time and rising time of the exponentials which are typically 150ps and 50ps, respectively. Moreover, Q_{inj} values range from -200fC to 200fC which relies on the particle energy as well as its linear energy transfer [162, 163]. The sign of the Q_{inj} depends on the type of the struck MOS transistor, in particular a strike in the drain of an NMOS transistor incurs a negative spike, and vice versa. Figure 3.16 depicts the injected current pulses for various Q_{inj} values, which are generated by hypothetical particles striking at t=0.

$$I(t) = \frac{Q_{inj}}{\tau_f - \tau_r}(e^{-t/\tau_f} - e^{-t/\tau_r}) \tag{3.3}$$

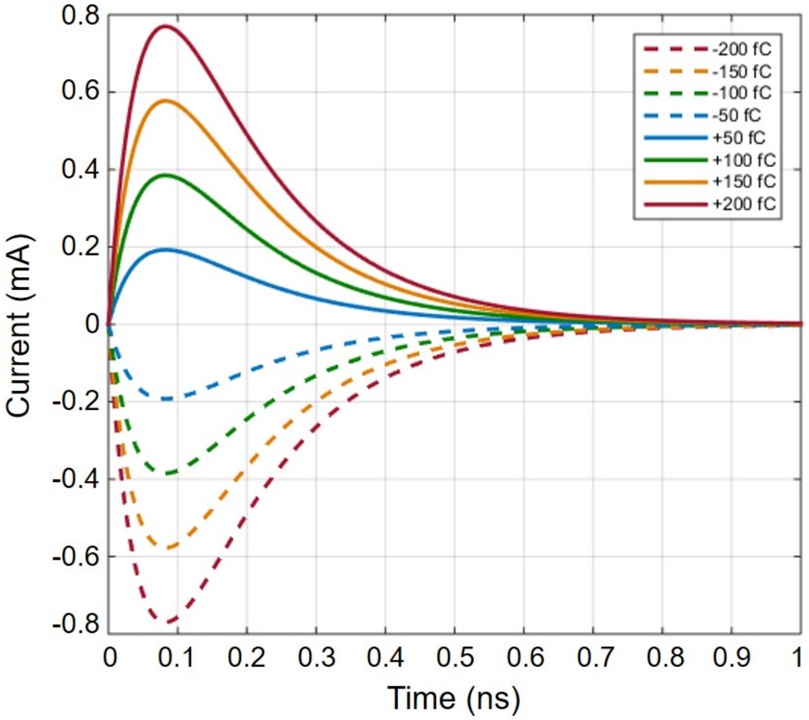
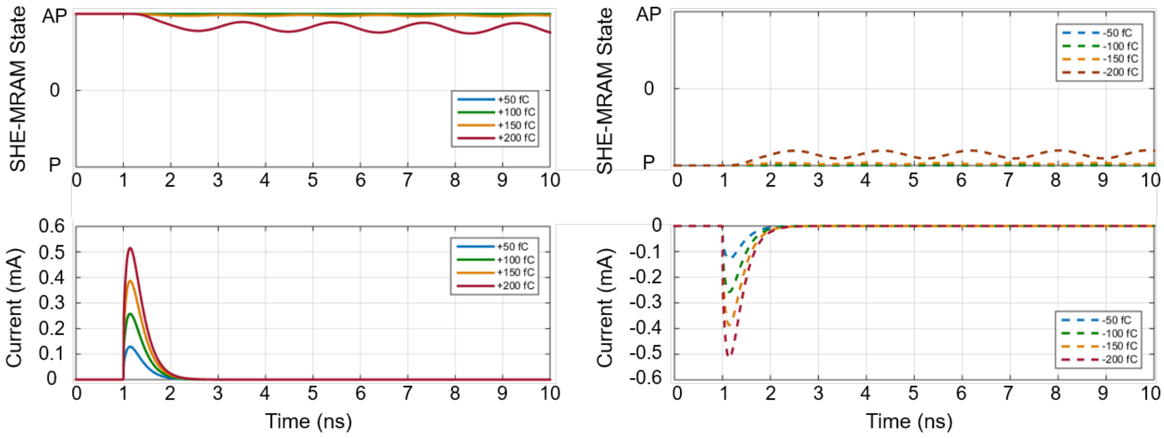


Figure 3.16: Transient current pulses induced by the particles striking at t=0 with the Q_{inj} values ranging from -200 fC to +200 fC [4].

3.5 Design and analysis of the proposed radiation-hardened MRAM-based LUT

Contrary to conventional SRAM cells, SHE-MRAM devices are characterized by their radiation hardness, since in MRAM cells the spin direction of electrons are leveraged to store data instead of the electron charges. The electric charges induced by the alpha particles striking the MRAM devices do not influence the spin direction of the electrons. However, the CMOS-based circuitry in hybrid CMOS/Spin circuits is still susceptible to radiation-induced transients. As investigated in [164], the radiation-sensitive nodes of a CMOS-based circuit are the surroundings of the reverse-biased drain junction of a transistor biased in the OFF state. Therefore, although the SHE-MRAM devices are immune to radiation during stand-by mode, their write circuit could be influenced by the striking particles. This leads to injecting a current to the write terminals of the SHE-MRAM, which normally cannot change their magnetic state due to the short duration of the injected current pulses. To exhibit the transient behavior of the SHE-MRAM devices in presence of the radiation-induced current pulses we have utilized the SHE-MRAM model developed by Camsari et al. in [12]. Figure 3.17 shows the response of the SHE-MRAM devices to the injected current pulses.

As shown in Figure 3.17, radiation does not have a significant effect on the LUT write operation. Thus, in this work we have focused on the effect of radiation on the LUT read operation. During the pre-charge operation of the PCSA, its transistors are biased in the ON state and will not be impacted by the radiation particles. While, in the discharge phase, OUT and OUT' nodes are the surroundings of reverse-biased junctions of NMOS or PMOS transistors that are biased in OFF states. Hence, OUT and OUT' are the sensitive nodes during the read operation of the LUT circuit.



(a) (b)

Figure 3.17: Transient response of the SHE-MRAM devices to the current pulses induced by the particles striking at $t = 1ns$. (a) Switching from AP to P state with the Q_{inj} values ranging from zero to $+200fC$, (b) switching from P to AP state with the Q_{inj} values ranging from $-200fC$ to zero. None of the injected current pulses can completely switch the state of the SHE-MRAM, since they have relatively short duration that is normally less than the switching duration required for completely changing the magnetic direction of the SHE-MRAM free layer [4].

Figure 3.18 shows the behavior of the LUT circuit depicted in figure 3.6 in presence of SEUs. Although the MRAM cells are resilient toward SEUs, the conventional MRAM-based LUT structure is still susceptible to the charges injected by the radiation particles.

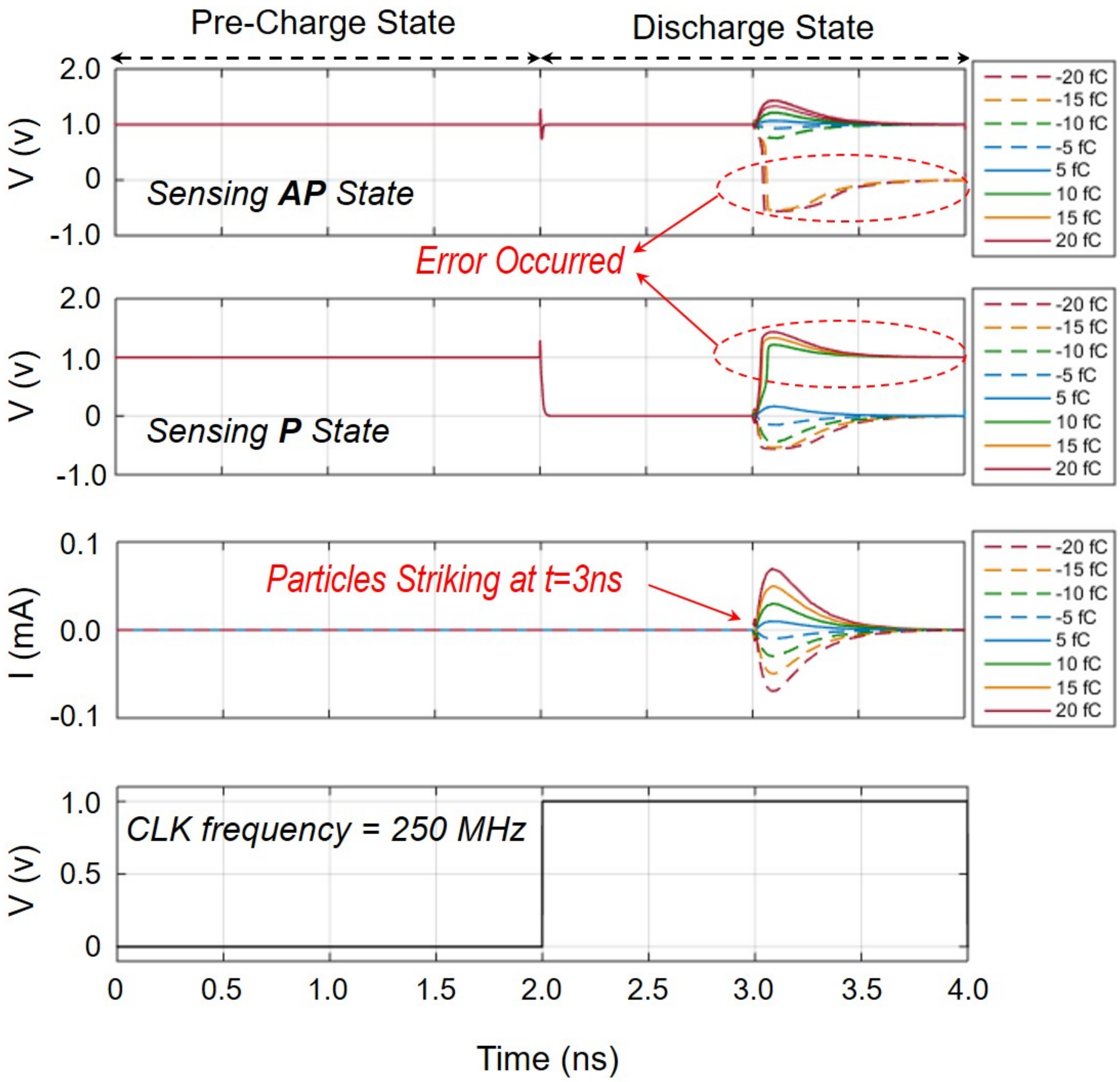


Figure 3.18: Transient response of the SHE-MRAM based LUT circuit to the current pulses induced by the particles striking at the discharge phase of the PCSA with the Q_{inj} values ranging from $-20fC$ to $+20fC$. As depicted, the ability of the circuit to recover from the SEU relies on the amount of the injected charge, as well as the critical charge of the circuit (Q_C). If the Q_{inj} exceeds the Q_C the sensed data cannot be recovered and error occurs [4].

Herein we build upon the previous radiation hardening techniques [56, 88, 89, 55] to develop a protected SHE-MRAM based LUT which can tolerate multiple node upsets. Our proposed approach is based on two hardening techniques: (1) leveraging feedback transistors to discharge the electric charges injected to the sensitive nodes through struck particles, and (2) increasing the critical charge (QC) of the sensitive nodes by increasing their equivalent capacitances while balancing tradeoffs of a corresponding increase in switching delay.

The structure of our proposed radiation-hardened 2-input SHE-MRAM LUT circuit is shown in Figure 3.19, in which the write circuitry is not shown for simplicity. The hardening circuitry includes two TGs, and four NMOS transistors which are responsible for discharging the electric charge induced by the radiation particles striking the OUT and OUT' nodes. However, the utilization of this feedback transistors introduces two new sensitive nodes to the LUT circuit, i.e. n1 and n2, as shown in Figure 6. Herein, the radiation-tolerance of n1 and n2 nodes are increased by enlarging the Q_C through increasing the equivalent capacitances of the nodes, which are linearly proportional to the width of the transistors connected to each node. The behavior of the proposed design in presence of SEUs and DNUs is shown in Figure 3.20 and Figure 8, respectively. The simulation is performed in the condition that the LUT storage cell is in P state, therefore the OUT and OUT' logic values are "0" and "1", respectively, and sensitive nodes are OUT, OUT', and n2.

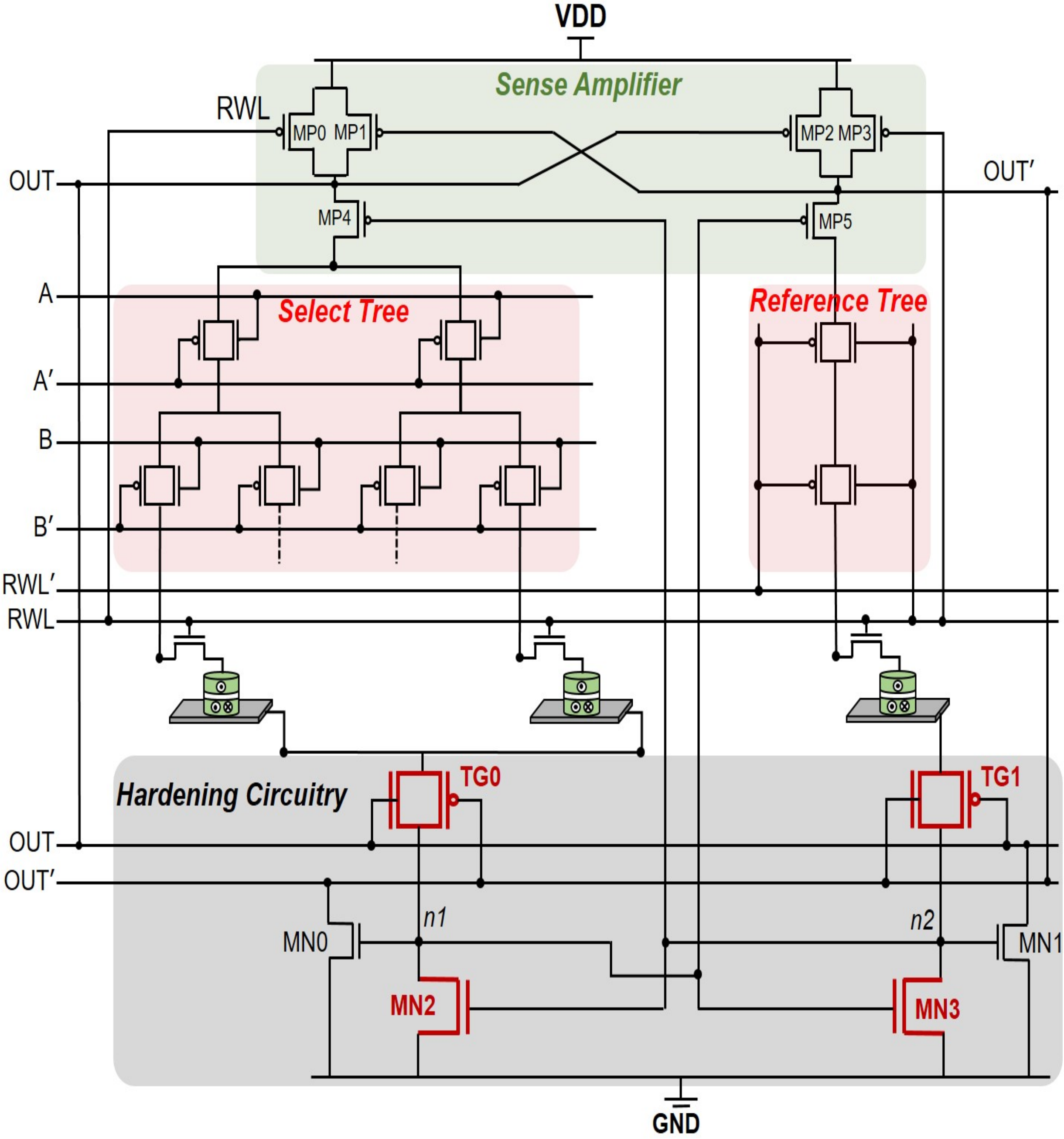
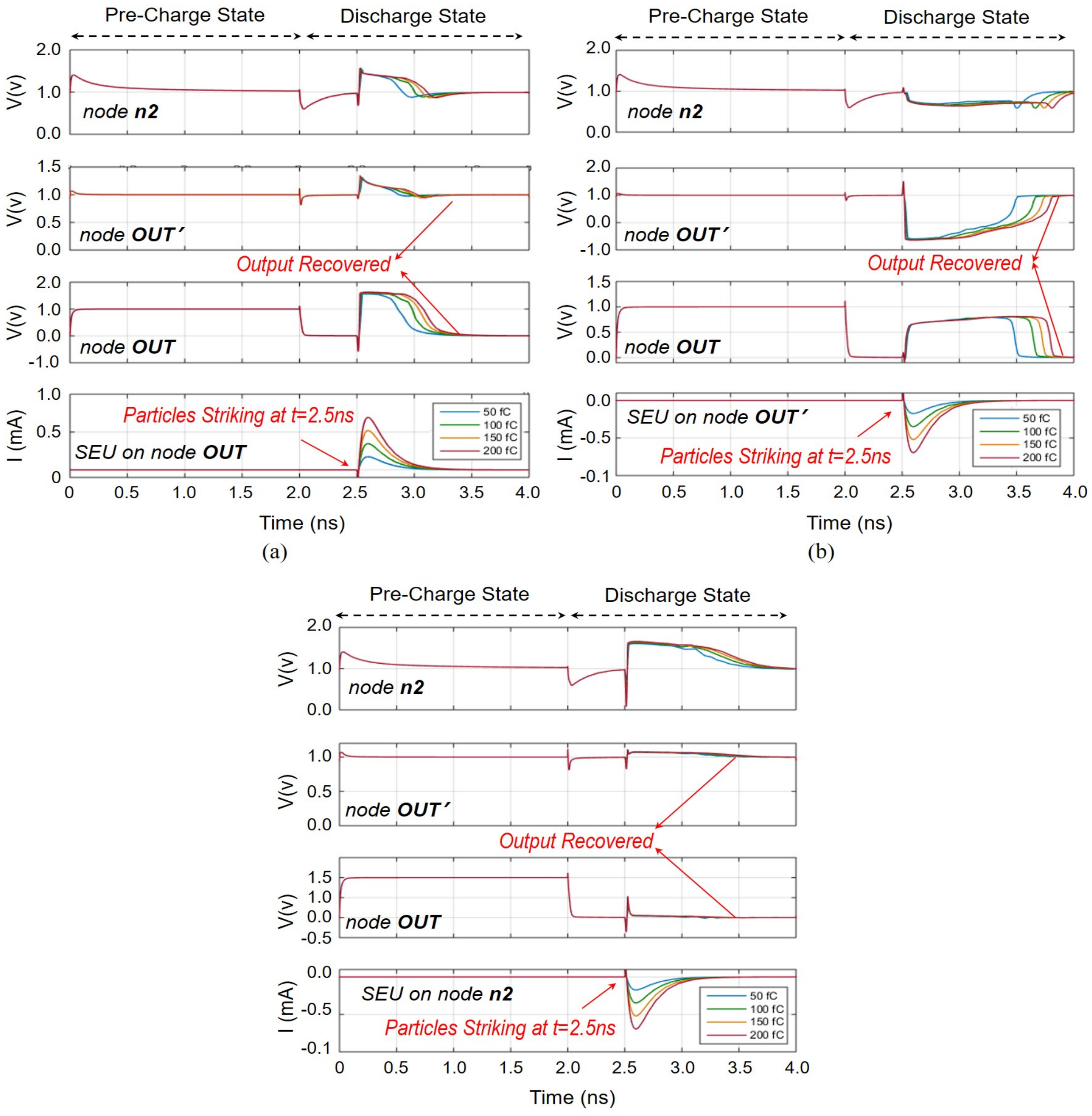


Figure 3.19: The structure of the proposed radiation-hardened 2-input SHE-MRAM based LUT circuit [4].

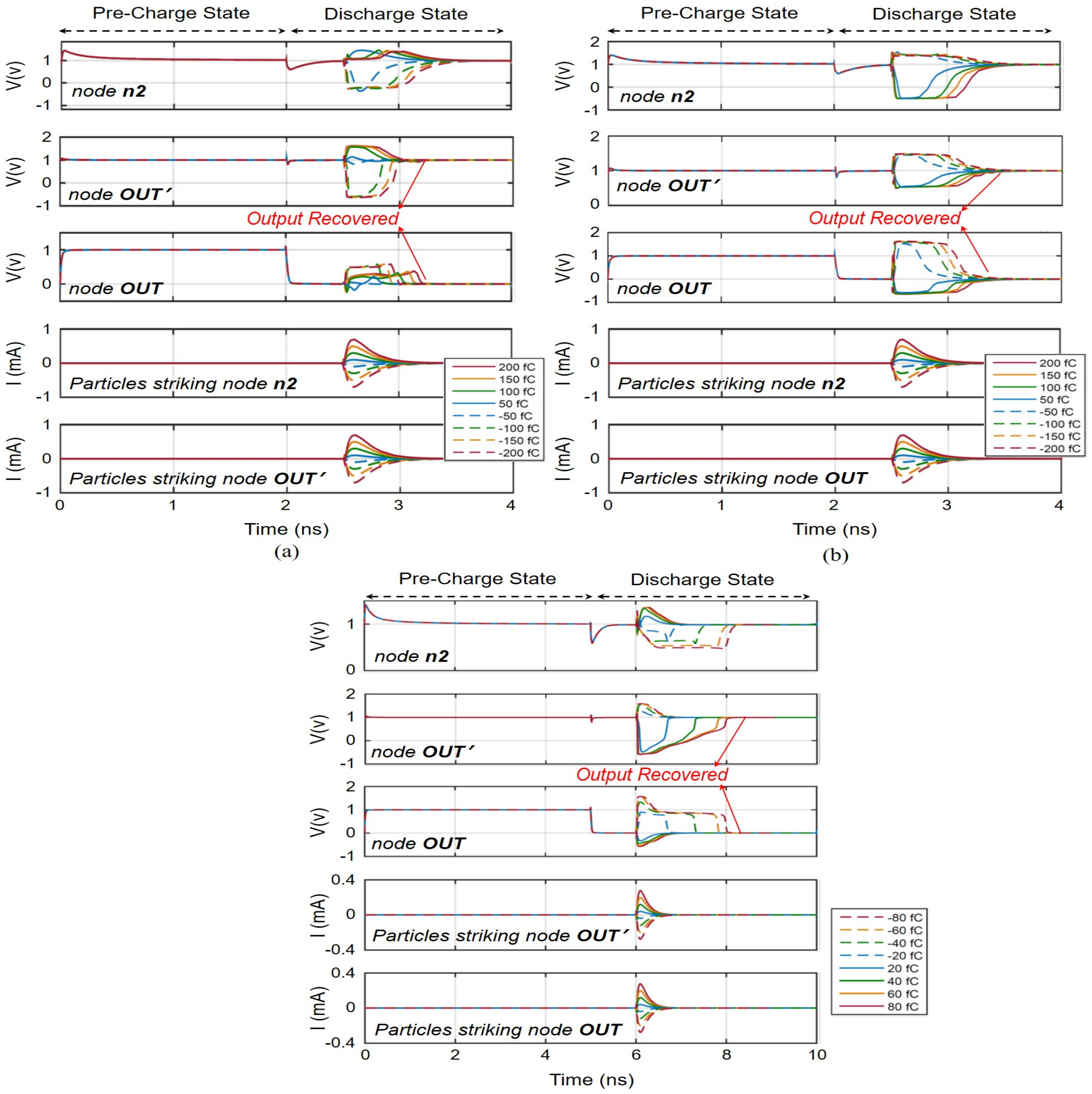


(a)

(b)

(c)

Figure 3.20: The transient response of the proposed radiation hardened 2-input SHE-MRAM LUT circuit to injected SEUs. (a) SEU on node OUT changes the voltage level of the node to VDD, however since the $n2$ node is still near VDD, thus the MN1 transistor remains ON and the injected charge will be discharged through MN1 and the output will be recovered. (b) SEU on node OUT’ changes the voltage of the node to zero, however since OUT node is still near zero, thus the MP2 transistor remains ON and the OUT’ node will be charged to VDD through MP2 and the output will be restored. (c) SEU on node $n2$ temporarily changes its voltage to zero, however it will not affect the OUT and OUT’ nodes, and TG1 and TG0 remain ON, thus the $n2$ node will be recharged to VDD through TG1 and reference tree [4].



(a)

(b)

(c)

Figure 3.21: The transient response of the proposed radiation hardened 2-input SHE-MRAM LUT circuit under injection of DNUs. (a) DNU on nodes $n2$ and OUT': the node $n2$ can tolerate the injected charge due to the increase in its Q_C, and since the OUT node remains near zero the OUT' node will be charged to VDD through MP2 and the output will be recovered. (b) DNU on nodes $n2$ and OUT: the radiation tolerance of node $n2$ is increased and will return to VDD, thus the MN1 transistor will become ON and the injected charge will be discharged through MN1 and the output will be recovered. (c) DNU on nodes OUT and OUT': it will not significantly impact node $n2$, therefore MN1 will remain ON and the injected charge at OUT node will be discharged through MN1 leading to the OUT' being recharged through MP2. However, since the charge capacity of OUT and OUT' are not increased in the LUT circuit, they can tolerate the maximum charge of $80fC$ that is smaller than the Q_C of the node $n2$ [4].

Table 3.9: Comparison of the proposed radiation-hardened SHE-MRAM LUT with the previously proposed MRAM-based LUTs. The results are obtained for LUT circuits implementing a two input NAND operation when A= 1 and B=1 inputs are applied [4].

Features	[76]	[56]	[89]	Proposed Herein
# of MTJS	12	8	8	5
# of MOSs	30	63	42	31
Delay (ps)	21.18	43.65	51.1	32.97
Power (μW)	0.21	1.08	0.6	0.57
PDP ($ps \times \mu W$)	4.45	47.14	30.66	18.79
Minimum TMR Required (%)	100	700	400	100
SEU Immune	No	Yes	Yes	Yes
DNU Immune	No	No	Yes	Yes

A comprehensive comparison of the different SHE-MRAM LUT circuits implemented and examined in this work is listed in Table 3.9. Herein, to provide a fair comparison, all the LUT circuits are simulated by SPICE circuit simulator using the SHE-MRAM model introduced in Chapter 2 along with the 45nm CMOS library with 1.0V nominal voltage. Moreover, we have utilized TGs to implement both select trees and write circuits in all of the investigated LUT designs. The results obtained are listed in Table 3.9. They indicate that the proposed SHE-MRAM LUT circuit can achieve DNU immunity with more than 38% and 60% improvement in power-delay product (PDP) as well as 26% and 50% device count improvement compared to the previous energy-efficient radiation-hardened LUT designs proposed in [89] and [56], respectively. The radiation-hardening ability is realized at the cost of increased PDP values compared to the unprotected MRAM-LUT design proposed in [76]. Finally, the sixth row of the Table 3.9 shows the minimum TMR required for MRAM cells in the LUT circuits to ensure their correct operation. As listed in the table, our proposed radiation-hardened LUT can properly operate with the TMR values similar to that of the unprotected LUT circuit. While, the previous radiation-hardened LUT designs require larger TMR values imposing more complex fabrication process [165].

Table 3.10: Parameters used for a Monte Carlo simulation in SPICE to perform the PV analysis [4].

Device	Parameter	Mean	Standard Deviation
NMOS	V_{TH}	0.34V	10%
PMOS	V_{TH}	-0.23V	10%
MTJ	t_{OX}	0.95nm	5%
	Area	60nm $\times$ 30nm $\times$ $\pi/4$	15%

3.6 Process Variation analysis of the proposed radiation-hardened MRAM-based LUT

To increase the radiation-tolerance of the LUT circuit, a number of transistors have been added in the sensing path. This can increase the error rate of the read operation caused by device mismatches due to process variation (PV). Therefore, in this section the effect of PV on various protected and unprotected LUT circuits is assessed. The impact of PV on hybrid CMOS/spin-based circuits results from a combination of systematic variations which are mostly caused by deposition and lithography aberrations, and random variations induced by random doping deviations [166, 167]. Table 3.10 lists the parameters utilized in this work for analyzing the PV.

Herein, we have fitted the experimental data extracted in [168] to an exponential curve to obtain the effect of oxide thickness (t_{OX}) variation on TMR values, as shown in Figure 3.22. The relation between the t_{OX} and TMR can be expressed by Equation 3.4, in which $C1$, $C2$, and $C3$ are fitting parameters.

$$TMR = C1 - \frac{C2}{C3}(1 - e^{-3t_{OX}}) \tag{3.4}$$

To examine the behavior of LUT circuits in presence of these sources of PV, we have leveraged a Monte Carlo simulation in SPICE, and the results are obtained for 10,000 simulation points. The results obtained exhbit that the radiation hardening is achieved at the cost of increased susceptibility to process variation, which is caused by the transistors inserted within the sensing path.

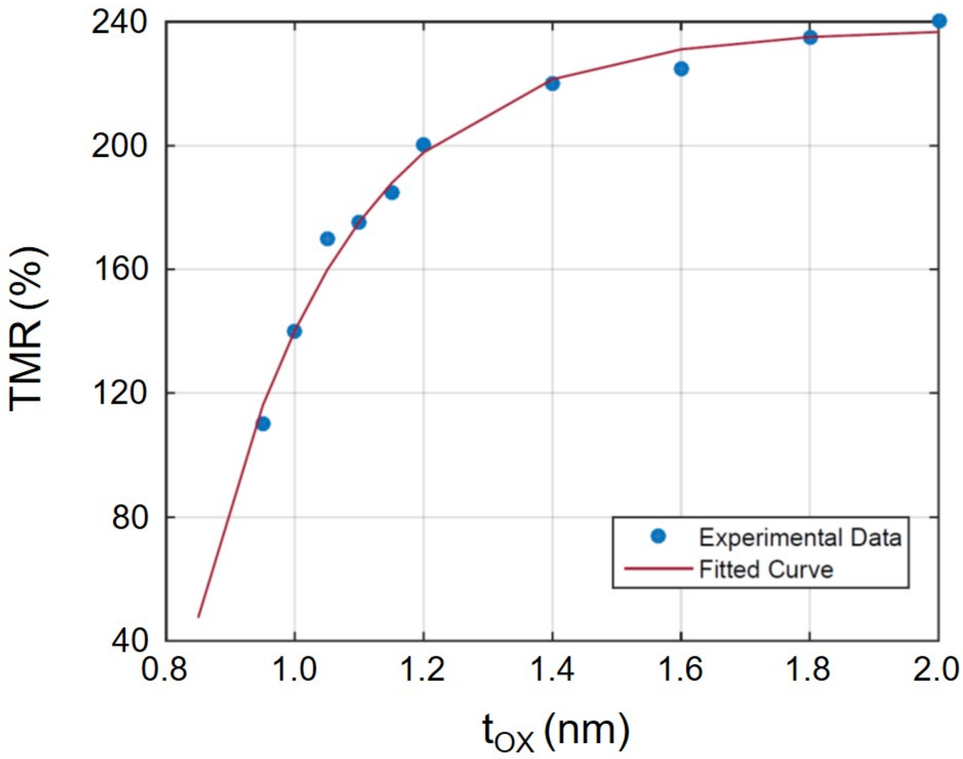


Figure 3.22: TMR ratio plotted as a function of MgO layer thickness [4].

CHAPTER 4: HSC-FPGA: A HYBRID SPIN/CHARGE FPGA LEVERAGING THE COOPERATING STRENGTHS OF CMOS AND MTJ DEVICES

Field programmable gate arrays (FPGAs) are renown for their flexibility to support circuit synthesis that is specific to the application at-hand using a palette of heterogeneous fine-grained logic elements [31, 169, 170]. Since the first FPGAs, various granularities of general-purpose configurable logic blocks and dedicated function-specific computational units have been added to reconfigurable fabrics. Over the last decade, reprogrammable fabrics have further embraced highly-dedicated special-purpose co-processing units to handle complex floating-point computations [1]. At the opposite end of the spectrum, FPGA fabrics can embrace increased heterogeneity along transformative dimensions by utilizing emerging logic and memory devices to leverage technology-specific benefits.

Recent research efforts have begun to explore the feasibility of spin-based devices such as magnetic tunnel junctions (MTJs) as an alternative for static random access memory (SRAM) cells in FPGAs [75, 171, 77]. Herein, we have used a device-to-architecture design approach to develop a Hybrid Spin/Charge based FPGA (HSC-FPGA), which leverages the cooperating strengths of CMOS devices for their rapid switching capabilities and MTJ devices for their non-volatility and near-zero standby power characteristics. The HSC-FPGA fabric includes hybrid spin/CMOS based configurable logic blocks (CLBs) using SRAM-based and magnetic random access memory (MRAM)-based look-up table (LUT) circuits to implement combinational and sequential logic, respectively.

The proposed HSC-FPGA provides a practical and feasible solution for exploiting spintronic devices within an FPGA architecture without requiring significant modifications to the interconnect structure and place/route/programming paradigms. In this chapter, we have provided thorough circuit-level and fabric-level simulations and analyses to exhibit the advantages of the proposed HSC-FPGA. Moreover, device-level optimizations are provided, which can address some of the challenges of the conventional hybrid spin/CMOS based circuits. In contrast to previous academic works of post-CMOS LUTs, we have considered challenges of the fabrication process of MTJs to investigate the feasible approaches to integrating spintronic and CMOS devices from practical viewpoints.

4.1 Structure of the HSC-FPGA

In this section, we will focus on the top-down hierarchical structure of the HSC-FPGA, which is designed to have the highest compatibility with the routing structure, programming paradigms, and synthesis tools of commercial FPGAs, while fully-leveraging the strengths of technology heterogeneity. Figure 4.1 shows the structure of the HSC-FGA, which consists of configurable logic blocks (CLBs), input-output blocks (IOBs), block RAMs, programmable switch matrices (SMs), and delay-locked loops (DLLs) for clock distribution. The logic functions are stored in the CLBs through conventional configuration bitstreams.

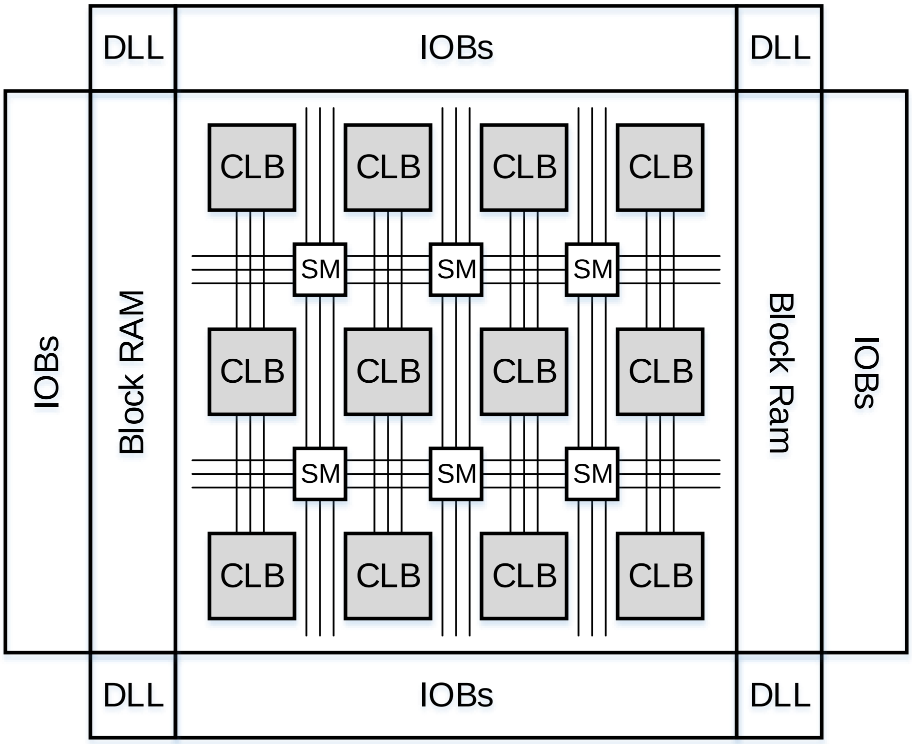


Figure 4.1: The structure of HSC-FPGA that is identical to commercial FPGAs.

4.1.1 CLB Structure

CLBs are the primary building blocks in FPGAs to implement both sequential and combinational logic circuits. Herein, we seek to identify practical methods for using the unique characteristics of the spintronic devices in the CLB structure to improve its performance without sacrificing needed functionalities provided by contemporary CMOS-based FPGAs. The CLBs being proposed will be required to provide the following logic circuits to ensure functional equivalence with modern FPGAs: (1) six-input look-up table (LUT) circuit, (2) dual five-input LUTs, (3) distributed memory, (4) shift register, and (5) dedicated carry logic for arithmetic operations.

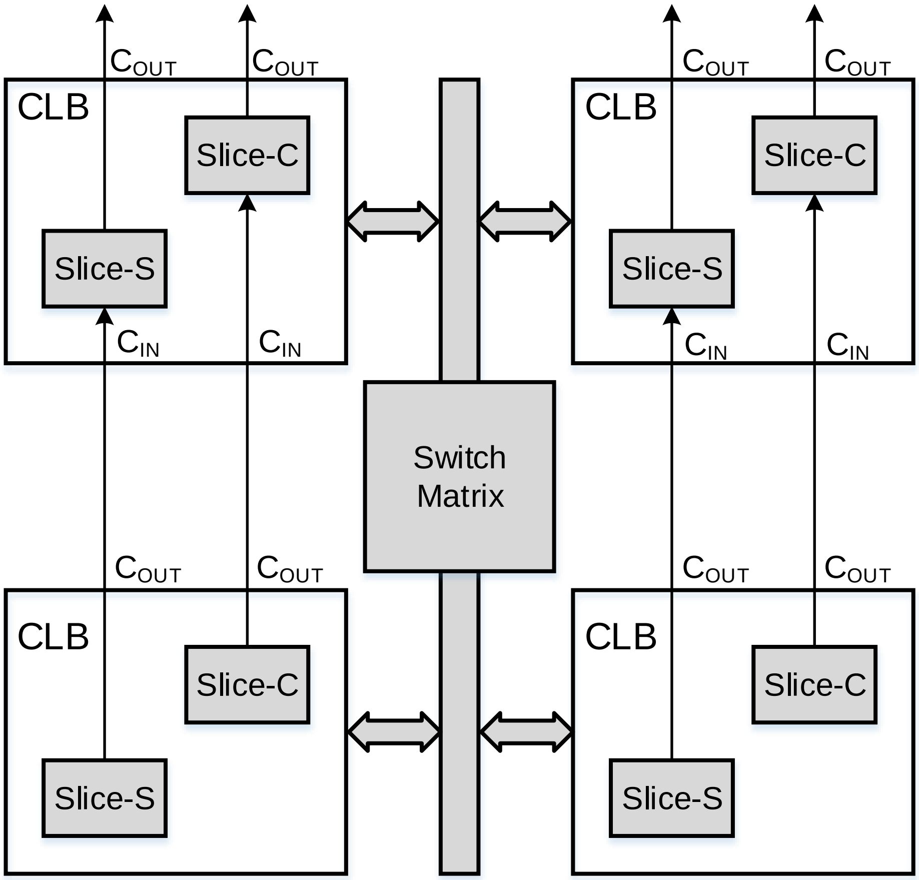


Figure 4.2: The structure of the CLB with two slices. The interconnection between CLBs is provided by switch matrices.

Spintronic devices such as MTJs can be leveraged as storage elements in the LUT circuits as an alternative for SRAM cells. However, in order to sense the state of the MTJs, a sense amplifier is required to be pre-charged and discharged using a clock signal for each read operation. Therefore, although MTJ-based LUT circuits can provide significant standby and read power-reductions, they are not directly suitable to implement combinational logic. Thus, as shown in Figure 4.2, we have proposed a CLB architecture containing two slices called Slice-S and Slice-C, which are utilized to implement sequential and combinational logic paths, respectively. Slice-C consists of SRAM-based LUT circuits that can also operate as shift registers and distributed RAM elements, while Slice-S includes spin-based LUTs that are paired with latches and flip-flops to implement sequential logic. Dedicated carry logic block are allocated to both of the slices for arithmetic operations. The simplified diagram of the proposed Slice-C and Slice-S structures are shown in Figure 4.3 and Figure 4.4, respectively.

4.1.2 Look-Up Table (LUT) circuits

An *m*-input LUT circuit is a $2^m \times 1$ memory block, in which the truth table of an *m*-input Boolean logic function is stored. The inputs of the Boolean function address a specific memory cell in the LUT, where their corresponding output is stored. Herein, we propose a hybrid-technology CLB structure, in which SRAM-based and MRAM-based LUTs are utilized to implement combinational and sequential logic circuits, respectively. In commercial FPGAs, each LUT circuit can implement any six-input Boolean function, as well as two five-input Boolean functions with shared inputs. Therefore, our MRAM-based LUT circuit should also have similar characteristics to ensure functional equivalence with SRAM-LUTs.

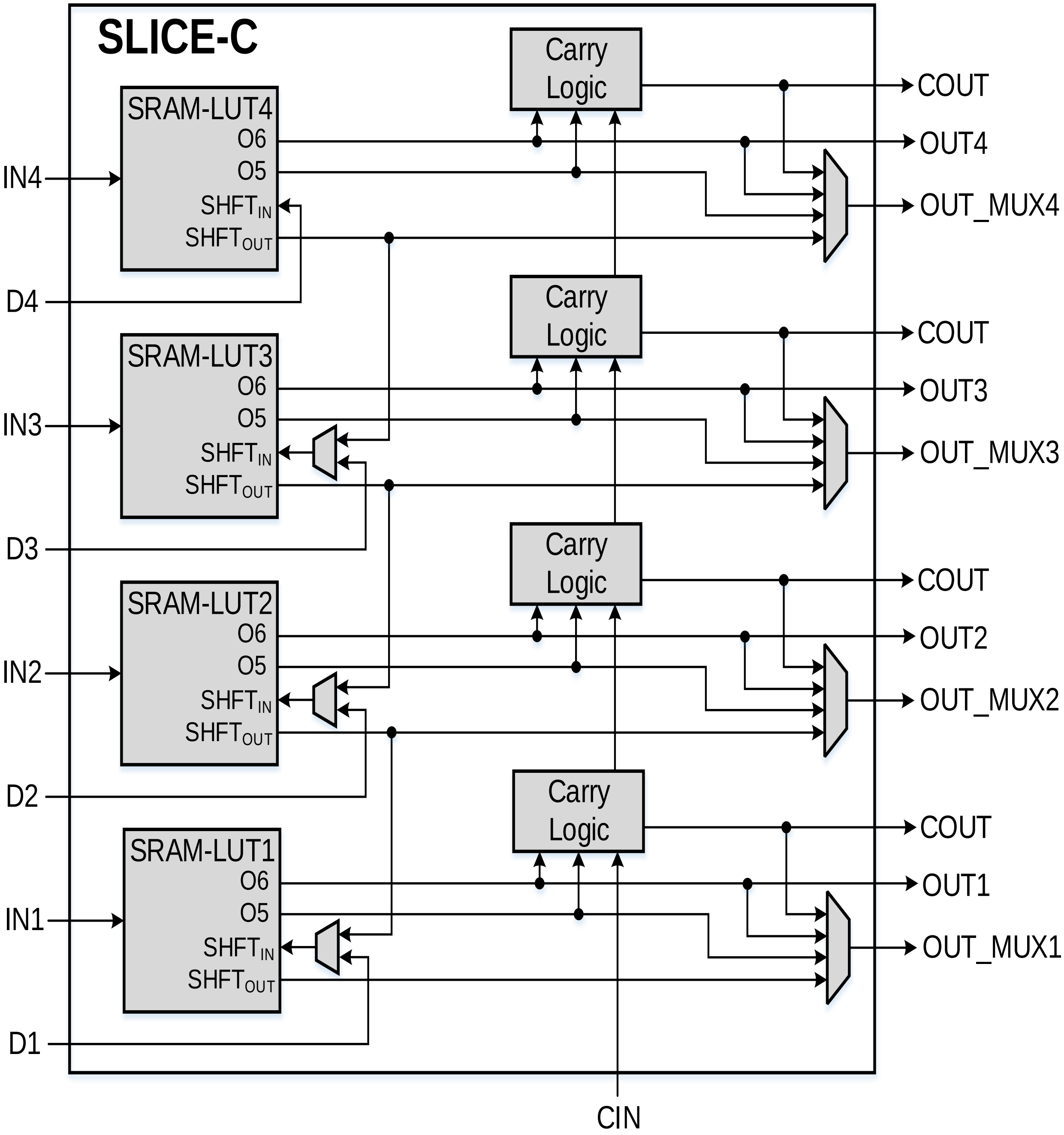


Figure 4.3: The structure of the Slice-C, which uses SRAM-based LUTs to implement combinational logic. SRAM-LUTs can also be configured to operate as shift registers.

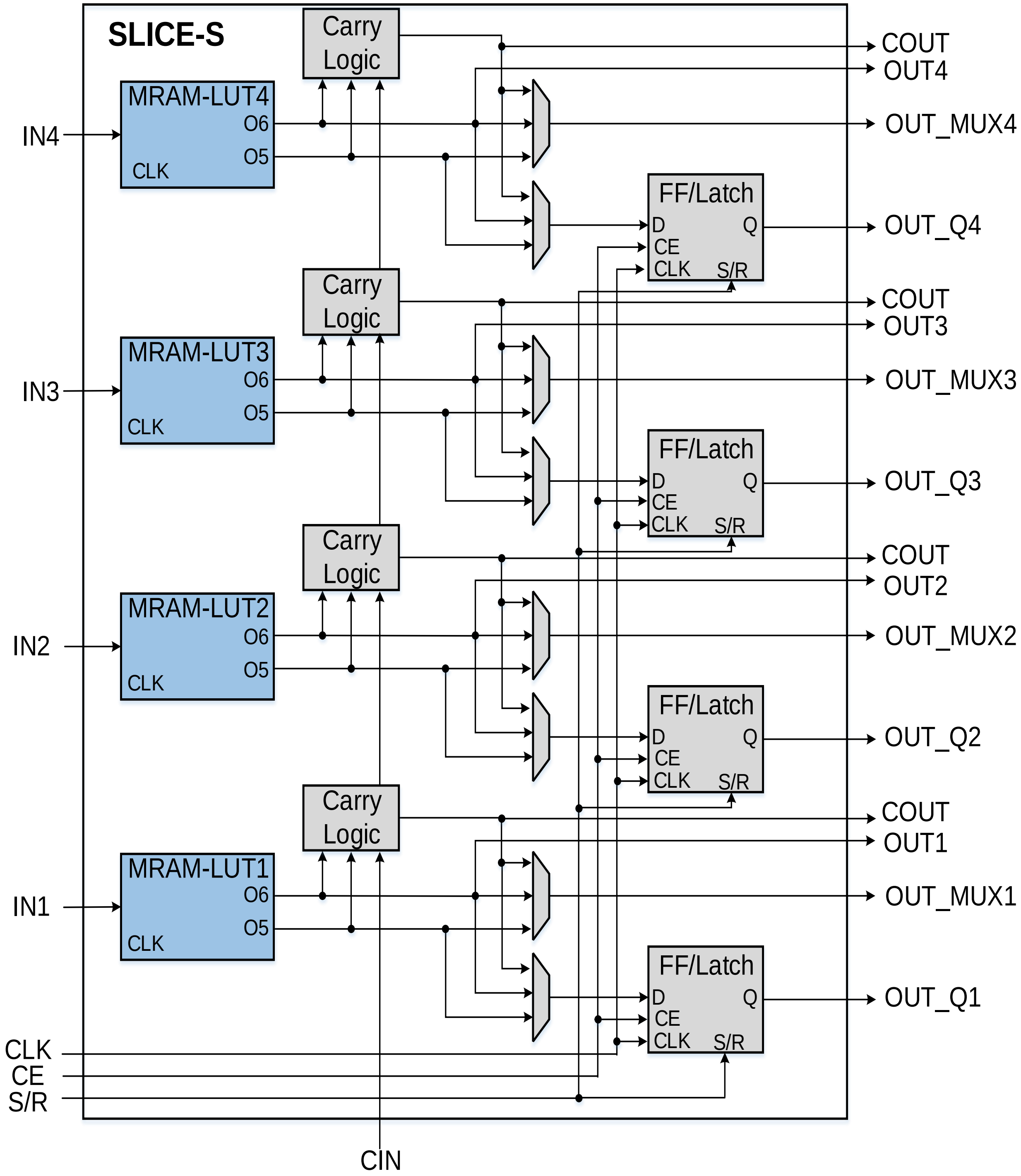


Figure 4.4: The structure of the Slice-S, which utilizes spin-based LUTs that are paired with latching and flip-flop circuits to implement sequential logic.

4.1.2.1 MRAM-based LUT Circuit

Figure 4.5 shows the structure of the proposed six-input MRAM-LUT, which includes MRAM-based storage cells, a decoding multiplexer (MUX), a mode selector, and two pre-charge sense amplifiers (PCSAs) to implement any six-input Boolean functions or two five-input Boolean functions with common shared inputs. The logic configuration is stored in MRAM cells as different resistive levels. High resistance represents logic "1" and vice versa. The resistance of the MRAM cells rely on their magnetization orientation, which can be tuned via the spin transfer torque (STT) effect that is described in next the section. The red-colored transmission gates (TGs) shown in Figure 4.5 are used as write circuits to produce the STT required for switching the state of the MRAMs. Each MRAM cell in the LUT circuit can be read through the TG-based decoding multiplexer using the corresponding input address provided by A5-A0 input signals. TGs are characterized by their full-swing switching behavior providing an energy-efficient write operation for MRAM cells [108], as well as a process variation resilient read operation for LUT circuits [3, 143].

The PCSA circuit reads the state of the MRAM cells by comparing its resistance by a reference cell, which its resistance is designed between the high and low resistance values of MRAM cells. In particular, the output terminals of the PCSA (O6-O6' or O5-O5') are charged to VDD when the clock signal (CLK) is low. When the CLK and read enable (RE) signals are high, the pre-charged nodes begin discharging and the node that has a lower resistance path to the ground discharges faster and its voltage becomes zero connecting the other node to VDD through PMOS transistors. Thus, if the MRAM cell is in low resistance state, the output node of the PCSA connected to this MRAM cell through the MUX (i.e. O6 or O5) will discharge faster and its voltage will becomes zero and vice versa. The reference tree is placed in the MRAM-LUT circuit to compensate for the resistance of the MUX. The readers are referred to Chapter 3 for additional information regarding PCSA operation. The M5 and M6 signals are used to select the 5-input or 6-input modes.

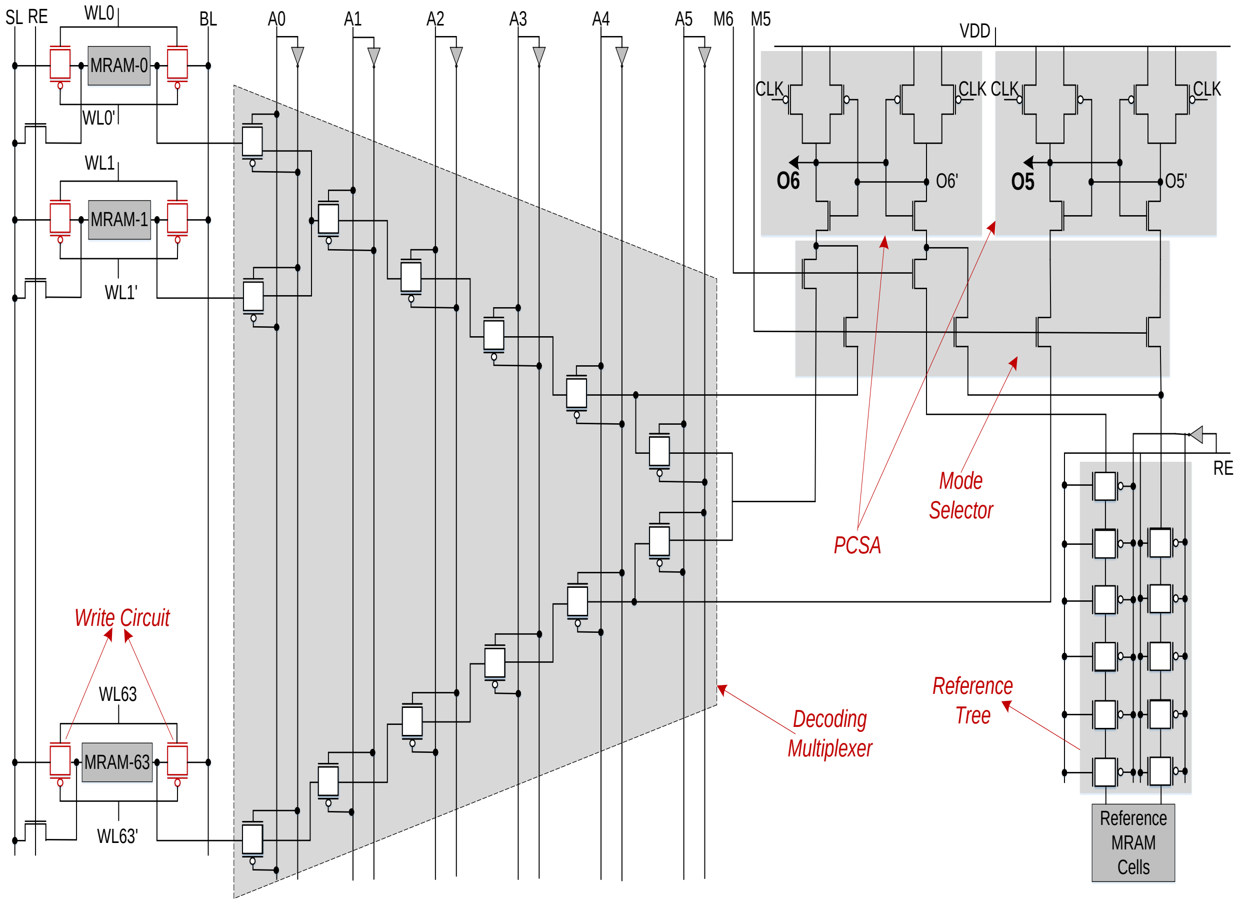


Figure 4.5: The MRAM-LUT structure consisting of MRAM cells as storage elements, decoding multiplexer, and two PCSAs.

4.1.2.2 *SRAM-based LUT Circuit and Shift Register*

Figure 4.6 shows the structure of our developed SRAM-LUT circuit. In the normal operation mode, the data stored in each SRAM cell can be accessed through the decoding MUX according to the A5-A0 input signals. Inverters are utilized before and after the MUX circuit to amplify the output of the SRAM cells. The SRAM-LUTs in Slice-C are designed such that they can be configured to operate as shift registers as well. In the shift-register mode, the shift-enable signal (SHFTE) is ON and the A5-A0 inputs are in high impedance (Hi-Z) states. Thus, data can be transferred from one SRAM cell to another by means of the pass transistors allocated for this purpose. The developed SRAM-LUT circuit is utilized to implement combinational logic circuits, while it can also be configured to operate as a 64-bit shift register. The $SHFT_{IN}$ and $SHFT_{OUT}$ lines can cascade different SRAM-LUTs in Slice-C to construct a larger shift register, as shown in Figure 4.3. The MRAM-LUTs can also be designed to operate as shift registers [172], however their energy consumption will be significantly higher than SRAM-based shift registers due to their high write energy, which will be investigated in the next section.

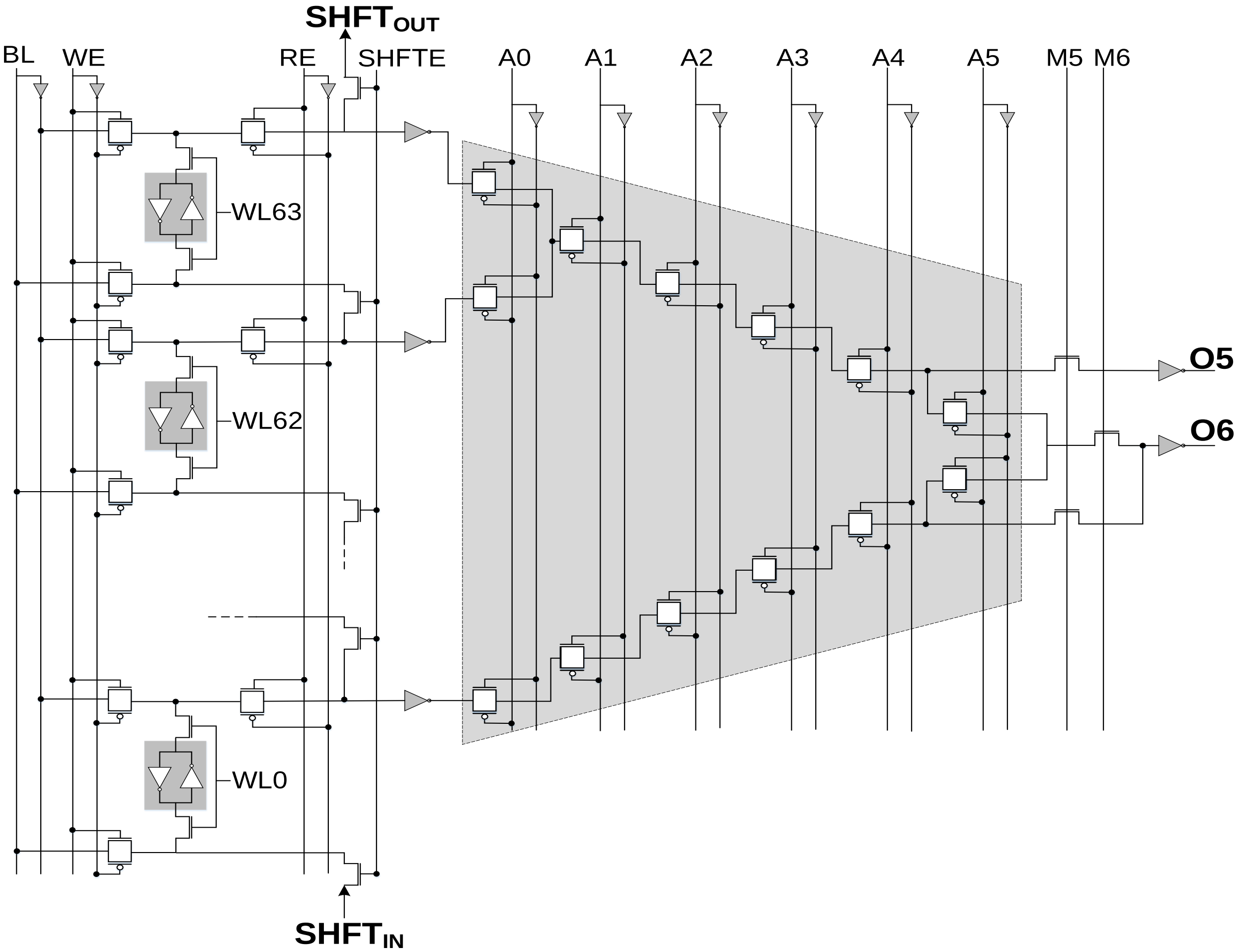


Figure 4.6: The structure of the SRAM-LUT consisting of SRAM cells as storage elements, decoding multiplexer, and the circuitry required for the shift operation.

4.2 HSC-FPGA Simulations

4.2.1 Circuit-Level Simulation

SPICE circuit simulation tool is utilized to verify the functionality of our proposed LUT circuits using 45nm CMOS technology and 1.1V nominal voltage. We have used the MTJ SPICE model proposed in [10] to implement our MRAM-LUT circuit using the parameters listed in Table 4.1.

4.2.1.1 MRAM-LUT circuit

Figure 4.7 (a) and (b) exhibit the SPICE simulation results for the MRAM-LUT circuit implementing a six-input NAND operation for A_5-A_0 = "111111" and A_5-A_0 = "000000" input signals, respectively. In order to write logic "0" (logic "1") in the MRAM-63 (MRAM-0) storage cell of the LUT, the word-line (WL), bit-line (BL), and source-line (SL) signals are required to be in high (high), high (low), and low (high) states, respectively. This results in a positive (negative) write current generated by the TG-based write circuit, which can change the MTJ state from AP (P) to P (AP) in less than $2ns$. To produce the sufficient switching current, the write circuit transistors have been enlarged four-fold.

During the read operation, when the RE signal is low, the PCSA circuit remains in the pre-charge state. Upon RE signal becoming high, the PCSA begins the discharge phase and senses the state of the MRAM cell in less than $50ps$ and $30ps$ for logics "0" an "1", respectively. Thus, the read operation can be performed by a maximum clock frequency of 10 GHz. Herein, we are using a CLK signal with 1GHz frequency, therefore the read delay is determined by the clock period and not the delay of the PCSA circuit.

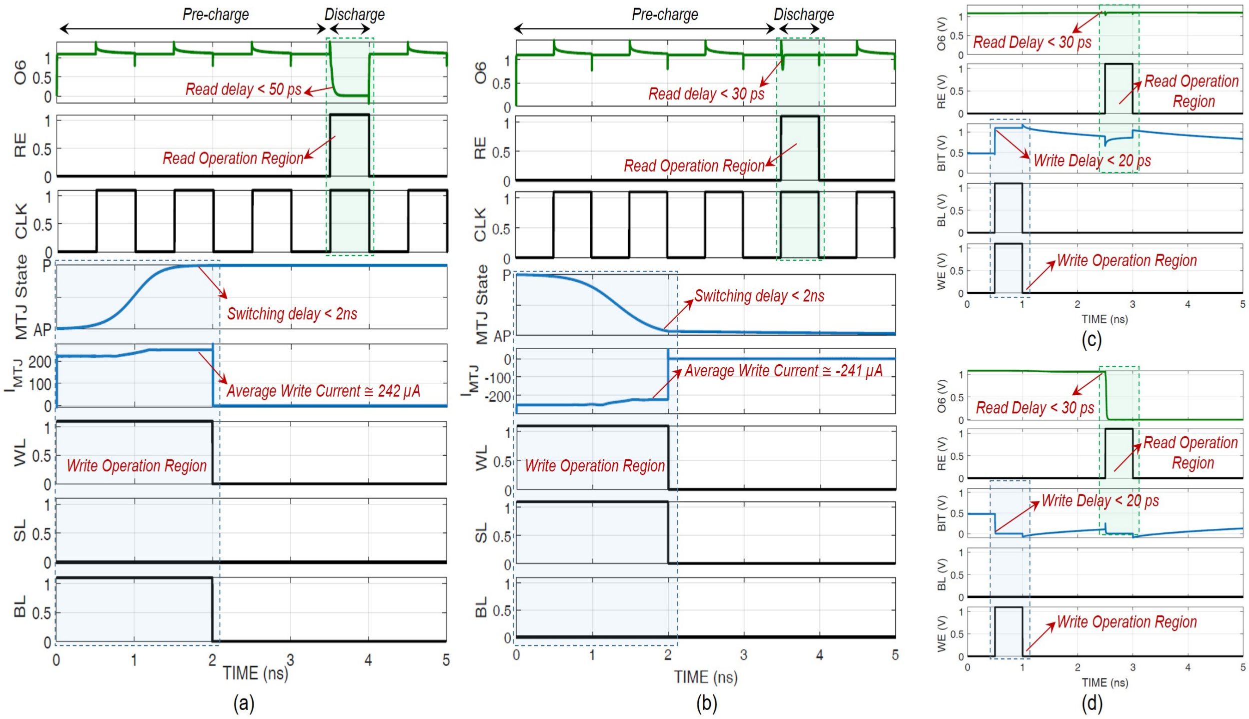


Figure 4.7: Transient response of MRAM-LUT implementing six-input NAND operation for (a) $A_5 - A_0$= “111111” and (b) $A_5 - A_0$= “000000”. SRAM-LUT implementing six-input NAND operation for (c) $A_5 - A_0$= “000000” and (d) $A_5 - A_0$= “111111”.

Table 4.1: Parameters of STT-MTJ device [10, 11].

Parameters	**Description**	**Value**
$Area$	MTJ surface	$65nm \times 65nm \times \pi/4$
t_f	Free Layer thickness	1.3 nm
RA	MTJ resistance-area product	$5\ \Omega.\mu m^2$
T	Temperature	358 K
α	Damping coefficient	0.007
P	Polarization	0.52
V_0	Fitting parameter	0.65
α_{sp}	Material-dependent constant	2e-5

4.2.1.2 SRAM-LUT circuit

Figure 4.7 (c) and (d) show the SPICE simulation results for the SRAM-LUT circuit implementing a six-input NAND operation for A_5-A_0 = “000000” and A_5-A_0 = “111111” input signals, respectively. During the write operation, the WE signal is high and the logic state stored in the SRAM cell is defined by the BL signal. If BL is high, then logic “1” will be stored in the corresponding SRAM cell and vice versa. During the read operation, RE signal is activated and the stored data in SRAM cells (bit) is propagated to the output through the decoding MUX and inverters. The read and write operations in SRAM-LUTs can be completed in less than $30ps$ and $20ps$, respectively.

4.2.1.3 Performance comparison

There are three types of power consumption profiles in the LUT circuits. During the configuration operation LUTs consume write power, which occurs infrequently. LUTs within the active logic path consume read power, while the remaining LUT circuits consume standby power that is a significant cause of power dissipation in commercial SRAM-based FPGAs. Table 6.3 lists the power and delay measurements for developed MRAM-LUT and SRAM-LUT circuits. The simulation results exhibit more than 40% and 83% reduction in average read and standby power consumption, respectively, for MRAM-LUT circuit compared to the SRAM-LUT. Both of the LUT circuits provide a high speed read operation, therefore the read delay is limited by the clock/signaling limitations rather than the device and circuit characteristics.

Table 4.2: Comparison between SRAM-LUT and MRAM-LUT.

		Power (μW)			Delay	
		Read	Write	Standby	Read	Write
SRAM-LUT	Logic "0"	2.58	28.4	1.5	30 ps	20 ps
	Logic "1"	7.55	27.7	1.85	30 ps	20 ps
	Average	5.06	25.08	1.67	30 ps	20 ps
MRAM-LUT	Logic "0"	2.85	260.9	0.28	50 ps	2 ns
	Logic "1"	3.14	265.7	0.27	30 ps	2 ns
	Average	2.99	263.3	0.28	40 ps	2 ns

Table 4.3 provides a comparison between the SRAM-LUT and MRAM-LUT circuits in terms of device count and average energy consumption. As listed, the MRAM-LUT realizes at least 40% energy reduction during the read operation. However, This improvement is achieved at the cost of higher energy consumption during the configuration operation, which occurs rarely compared to the read operation. As mentioned, the write circuit transistors in the MRAM-LUT circuit are required to be enlarged four-fold to generate the sufficient switching current for the STT-MRAM cells. Therefore, we have multiplied the write circuit transistors by four to realize a rough area estimation. The MTJs can be vertically fabricated on top of the MOS transistors, thus incurring negligible area overhead.

Table 4.3: Area and Energy Consumption comparison between SRAM-LUT and MRAM-LUT.

Features		SRAM-LUT	MRAM-LUT
Device Count	Storage Cells	384 MOS	64 MTJ
	Write/Control	384 MOS	256×4+64 MOS [1]
	Read	261 MOS	165 MOS + 4MTJ
	Total	1029 MOS	1253 MOS + 68 MTJ
Average Energy Consumption	Read	2.53 fJ	1.5 fJ
	Write	14 fJ	526.6 fJ

[1] The write transistors are four-fold larger than minimum feature size.

4.2.2 Fabric-Level Analysis

Herein, we have used Xilinx ISE Design Suite solely to obtain the resource utilization for various ISCAS-89, ITC-99, and MCNC benchmark circuits to provide a fabric-level comparison between our proposed HSC-FPGA and conventional SRAM-based FPGAs. The resource utilization summary for the implemented benchmark circuits is provided in Table 4.4. In conventional FPGAs, SRAM-based LUTs are paired with flip-flops (FFs) to form a LUT-FF circuit implementing the sequential logic circuits. However, in the proposed HSC-FPGA architecture, LUT-FF pairs are constructed by combining an MRAM-LUT circuit with a CMOS-based slave latch. The MRAM-LUT circuit can intrinsically perform the master latch behavior, while realizing the basic LUT circuit operation. Therefore, the fully-used LUT-FFs listed in Table 4.4 are implemented by MRAM-LUTs in Slice-S, while the remainder of the Slice LUTs exist in the Slice-C of the HSC-FPGA fabric.

Figure 6.6 exhibit a power consumption comparison between the HSC-FPGA and conventional SRAM-based FPGA for various benchmark circuits. As shown, the HSC-FPGA can achieve standby power reductions ranging from 7% to 66%, as well as 15% average read power reduction for various ISCAS-89 and ITC-99 benchmarks. However, the performance of the HSC-FPGA is equivalent to SRAM-based FPGA while implementing fully-combinational circuits such as *bigkey* and *sbc* benchmark circuits examined herein, since combinational logic is only implemented by SRAM-LUTs in the HSC-FPGA fabric. Finally, Figure 4.9 shows the normalized power-delay product (PDP) values of HSC-FPGA for the read operation compared to SRAM-based FPGA for the examined benchmarks. The results obtained exhibit PDP improvements ranging from 2% to 17% for various ISCAS-89 and ITC-99 benchmark circuits.

Table 4.4: Resource utilization for various benchmark circuits.

Benchmark Circuits	Slice Registers	Slice LUTs	Fully-used LUT-FFs	Bonded IOBs
ISCAS 89				
s298	14	15	11	11
s382	21	31	19	11
s510	7	32	6	28
s641	14	43	12	56
s832	6	68	5	39
s1488	12	115	9	29
s5378	152	337	147	86
s9234	130	255	112	69
s15850	128	106	84	100
s38417	1355	1971	1072	136
ITC 99				
b5	47	135	31	39
b8	21	54	18	15
b10	19	37	14	19
b12	119	206	70	13
b15	419	1814	396	108
b18	2960	11915	2699	61
b20	429	2235	361	56
b22	629	3150	611	56
MCNC				
bigkey	0	567	0	425
sbc	0	179	0	96

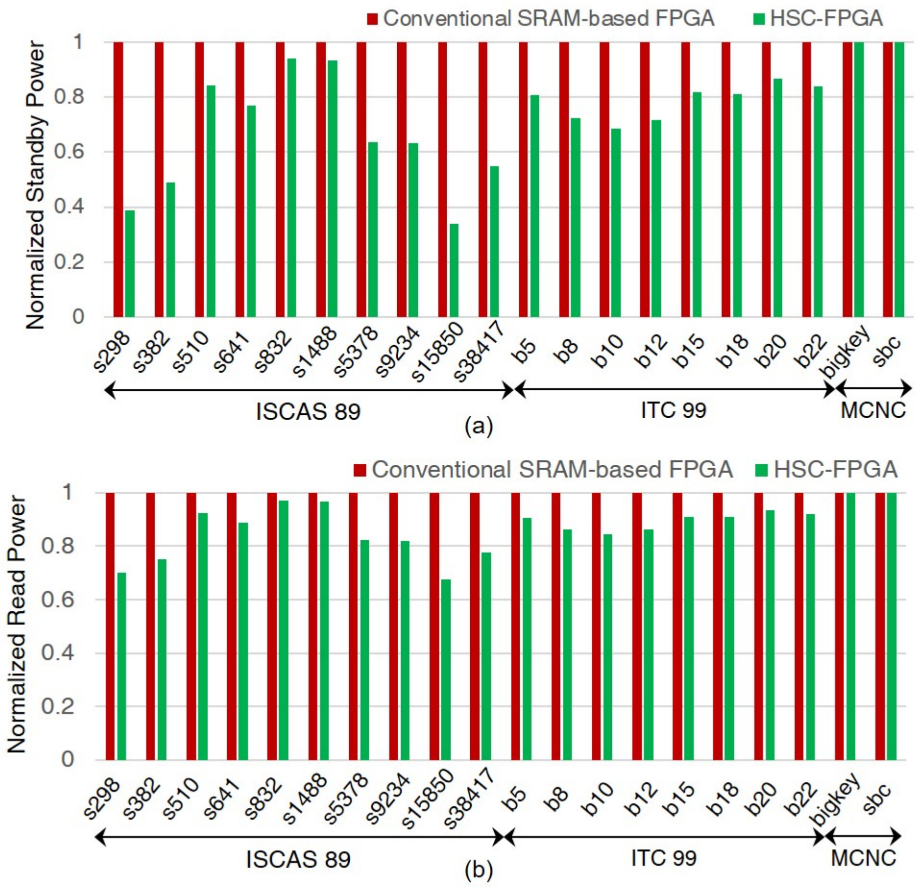


Figure 4.8: Normalized power consumption of HSC-FPGA compared to the SRAM-based FPGAs, (a) standby power, (b) read power.

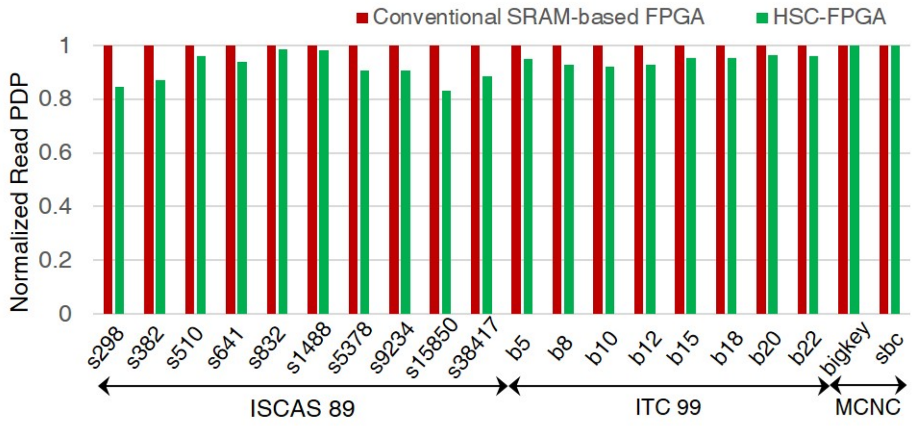


Figure 4.9: Normalized PDP values of the HSC-FPGA compared to the SRAM-based FPGA for read operation.

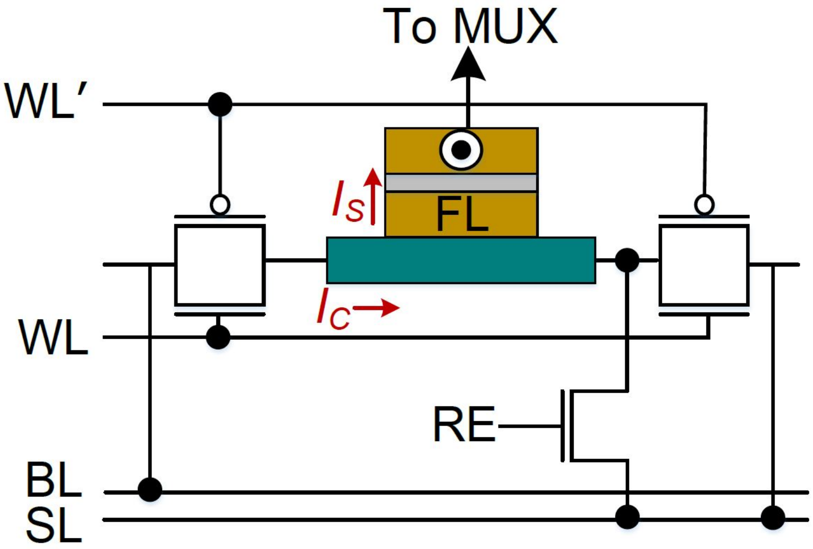


Figure 4.10: SHE-MRAM bit-cell structure.

4.3 Device Optimizations

MRAM-LUT circuits offer significant advantages in terms of read and standby power consumption. However, STT-MTJs suffer from high power and low speed switching behavior, which significantly increases the energy consumption of the MRAM-LUTs during the configuration operation. Therefore, device-level innovations are sought to improve the switching performance in MRAM cells. Herein, we have used a SPICE model of the SHE-MTJ device developed in [12] to verify the functionality and assess the performance of the SHE-MRAM based LUT circuit using the parameters listed in Table 4.5 [12, 13].

Figure 4.10 shows a SHE-MRAM bit-cell, which can be utilized in MRAM-LUT circuits as an alternative for STT-MRAMs without requiring any changes in the read circuitry. The write behavior of SHE-MTJ is depicted in Figure 4.11, realizing an equivalent switching delay compared to the STT-MRAM with significantly smaller switching current, which results in reduced write power consumption. Moreover, write circuits with minimum feature size MOS transistors are capable of generating the write current required for switching the logic state of the SHE-MRAM cells, leading to significant area savings. Comparison results provided in Table 4.6 exhibit that SHE-MRAM LUT realizes approximately 67% and 61% reductions in terms of configuration energy consumption and device count, respectively, compared to the STT-MRAM LUT circuit.

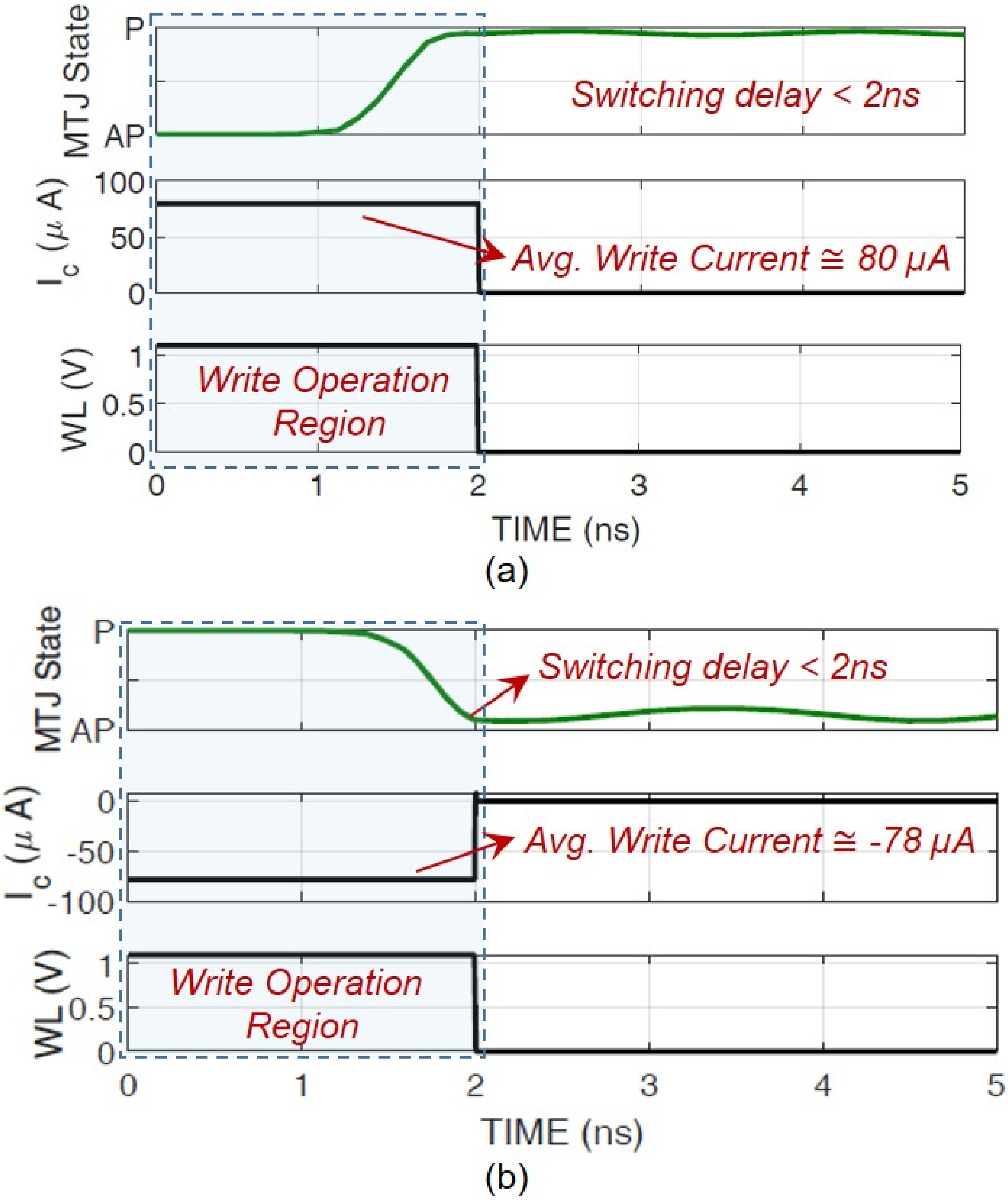


Figure 4.11: SHE-MRAM switching operation, (a) write logic “0” (AP to P), (b) write logic “1” (P to AP)

Table 4.5: Parameters of the SHE-MTJ device [12, 13].

Parameter	Description	Value
MTJ_{Area}	$l_{MTJ} \times w_{MTJ} \times \frac{\pi}{4}$	$60nm \times 30nm \times \frac{\pi}{4}$
HM_{Volume}	$l_{HM} \times w_{HM} \times t_{HM}$	$100nm \times 60nm \times 3nm$
α	Gilbert Damping factor	0.007
P	Spin Polarization	0.52
θ_{SHE}	Spin Hall Angle	0.3
ρ_{HM}	HM Resistivity	$200\mu\Omega.cm$
λ_{sf}	Spin Flip Length	$1.5nm$

Table 4.6: Area and Write Energy Consumption comparison between STT-MRAM LUT and SHE-MRAM LUT circuits.

Features		STT-MRAM LUT	SHE-MRAM LUT
Device Count	Storage Cells	64MTJ	64MTJ
	Write/Control	256×4+64 MOS [1]	256×1+64MOS [2]
	Read	165MOS+4MTJ	165MOS+4 MTJ
	Total	1253MOS+68 MTJ	485MOS+68MTJ
Average Write Energy per Cell		526.6 fJ	173.8 fJ

[1] Write circuit transistors are 4× larger than minimum feature size.

[2] Transistors with minimum feature size are used in the SHE-MRAM LUT.

Fabric-level simulation results depicted in Figure 4.12 indicate that SHE-MRAM based HSC-FPGA achieves write energy reductions ranging from 50% to 66% compared to STT-MRAM based HSC-FPGA for various ITC-99 and ISCAS-89 benchmarks. Moreover, Figure 4.13 shows average area reduction of 18% and 23% for SHE-MRAM based HSC-FPGA compared to SRAM-based FPGA and STT-MRAM based HSC-FPGA, respectively. Both improvements are obtained via technology-aware design leveraging the complementary features of each device technology.

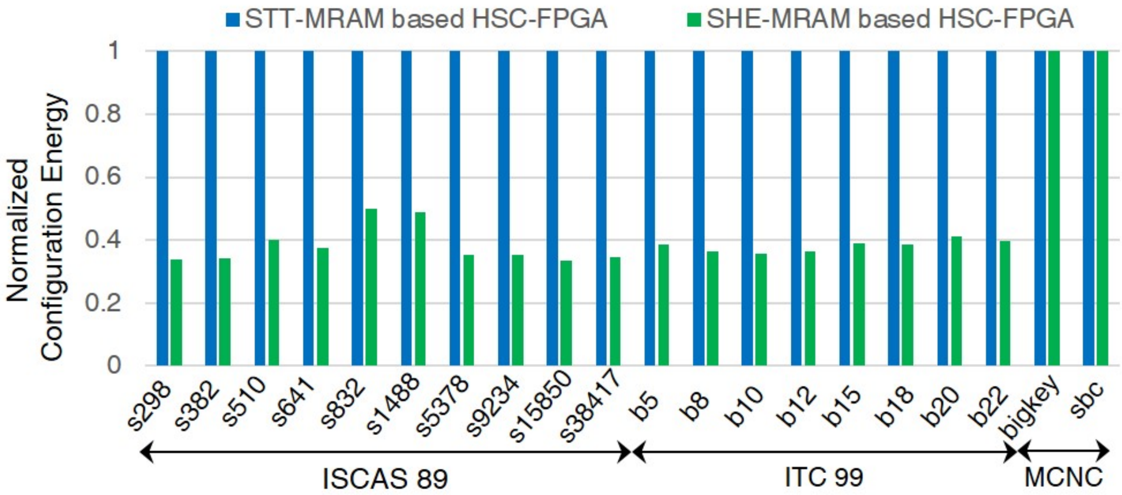


Figure 4.12: Normalized configuration energy consumption of the SHE-MRAM based HSC-FPGA compared to the STT-MRAM based HSC-FPGA.

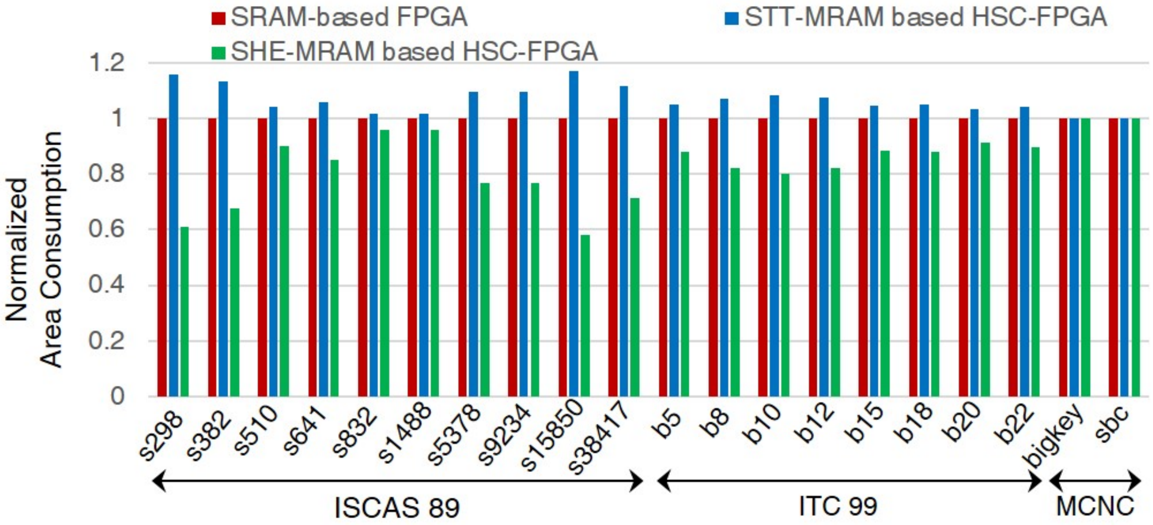


Figure 4.13: Normalized area of the STT-MRAM and SHE-MRAM based HSC-FPGAs compared to the SRAM-FPGA.

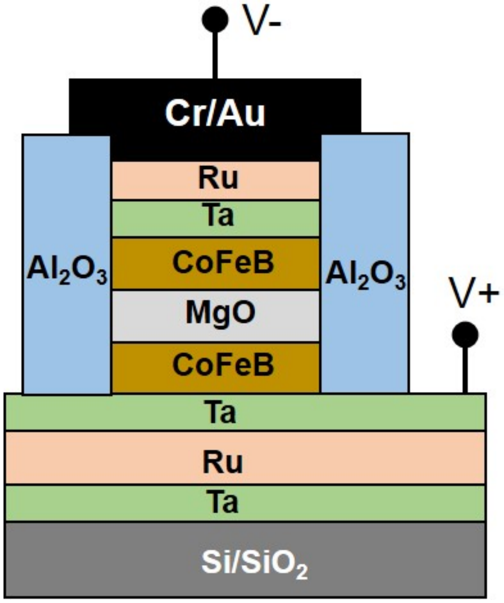


Figure 4.14: The MTJ stack structure consisting of, Ta (5)/Ru (10)/Ta (5)/CoFeB (1-1.5)/MgO (0.85-0.95)/CoFeB (1-1.5)/ Ta (5)/ Ru (5). Numbers represent the thickness in nm [5].

4.4 MTJ Fabrication Process and Challenges

In order to devise practical solutions to process variation (PV) for MRAM-based LUTs, it is helpful to consider Figure 4.14 which shows the stack structure of a perpendicular-anisotropy CoFeB-MgO based MTJ [5]. MTJ fabrication involves the following main processes: First, MTJ films are deposited on SiO2/Si substrate by using RF magnetron sputtering. Then, electron-beam lithography is followed by an Ar-ion milling to achieve high resolution Nano-pillar patterns. Next, SiO_2 is deposited to provide a barrier between different devices. This process is followed by a chemical mechanical polishing until the top contact of Nano-pillar is opened. Finally, the contacts are opened via reactive ion etching, and metallization approach is utilized to coat metal on the top of the MTJs' electrode ends. Similar techniques can be utilized to fabricate SHE-MTJ devices with in-plane anisotropy. Readers are referred to [173, 133] for additional information regarding the fabrication of MTJ devices.

Table 4.7: Parameters used for the PV analysis [14, 15, 16].

Device	Parameter	Mean	Std. Deviation
NMOS	V_{TH}	0.34 V	1-10 %
PMOS	V_{TH}	-0.23 V	1-10%
	t_f	1.3 nm	5%
MTJ	l_{MTJ}	65 nm	10%
	w_{MTJ}	65nm	10%

4.4.1 Process Variation Analysis

Device mismatches can be caused due to the PV in different steps of MTJ fabrication. In particular, the RF sputtering process can induce variations in the thickness of the films (σt), while the lithography and etching processes primarily result in variations in the width (σw_{MTJ}) and length (σl_{MTJ}) of the MTJs. On the other hand, CMOS transistors can also suffer from random variations in their threshold voltage (σV_{TH}) induced by the random doping deviations. Table 4.7 lists the parameters used herein to examine the effect of PV on the proposed MRAM-LUT circuit.

To examine the functionality of our developed SRAM-LUT and MRAM-LUT circuits in presence of PV, we have utilized Monte Carlo simulation in SPICE with 1,000 simulation points, as shown in Figure 4.15. The obtained results exhibit an unacceptable average error rate of $\sim 44\%$ for MRAM-LUT, while SRAM-LUT circuit can maintain its correct operation in presence of significant PV. These results were obtained by applying PV to all of the CMOS and MTJ devices existing in different parts of the MRAM-LUT structure, including the storage elements, write circuit, decoding multiplexer and sense amplifiers. In an effort to find the most PV-susceptible part of the MRAM-LUT circuit, we have applied random variations to all of the MOS and MTJ devices in the circuit except for the sense amplifier transistors and measured the error rate. The results exhibited a reduced error rate 4% showing the significant impact of the variations in SA circuit on the accuracy of the MRAM-LUT.

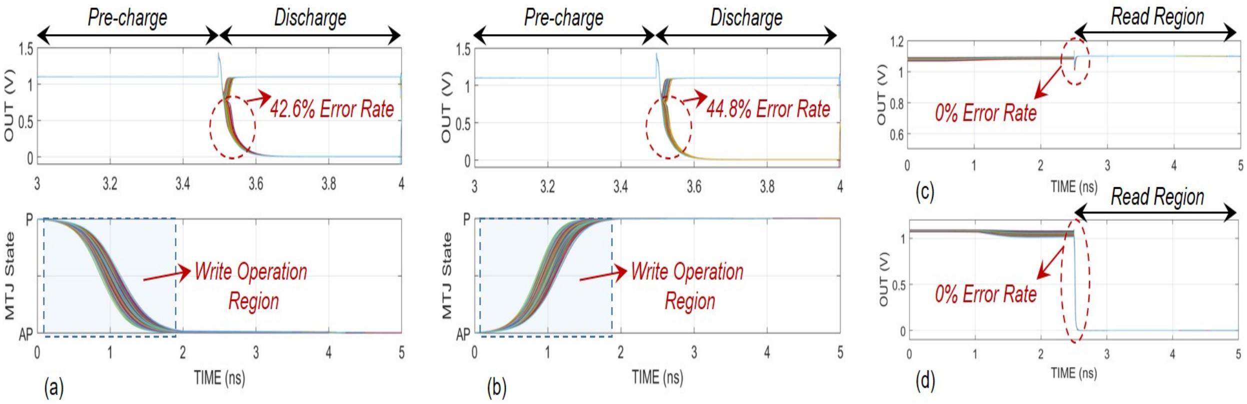


Figure 4.15: Effect of PV on the functionality of the developed LUT circuits. (a) MRAM-LUT reading logic "1", (b) MRAM-LUT reading logic "0", (c) SRAM-LUT reading logic "1", (d) MRAM-LUT reading logic "0".

There are various approaches to increase the tolerance of the hybrid spin/CMOS based circuits to PV, including increasing the size of the transistors in the sensing path or increasing the TMR ratio in the MTJ devices [43]. While these approaches mostly rely on the device-level innovations, we leverage a circuit-level method herein, which is based on the modular redundancy (MR) technique [56], to improve the resiliency of the MRAM-LUT circuit. Figure 4.16 shows the structure of the PV-tolerant MRAM-LUT circuit including three SAs and a voter circuit, which determines the output of the MRAM-LUT according to the majority of the SAs' outputs. Thus, the MR-based MRAM-LUT circuit is capable of returning the correct output even if one of the SAs malfunctions due to process variation.

Table 4.8 provides a comparison between the MR-based MRAM-LUT and regular MRAM-LUT in terms of error rate, and read power consumption. The results exhibit that the MR-based MRAM-LUT realizes an error rate of 12% at the cost of 24% and 6% read power and area overheads, respectively, compared to MRAM-LUT. Error rates of less than 0.1% can be achieved by further device-level innovations [174]. Figures 4.17 and 4.18 show the fabric-level comparisons between the HSC-FPGAs with MR-based MRAM-LUTs, HSC-FPGA with regular MRAM-LUTs, and conventional SRAM-based FPGAs in terms of average read power consumption and area, respectively. The results obtained show that the PV-resilience in the MR-MRAM based HSC-FPGA fabrics is achieved at the cost of approximately 6% and 1.5% overheads in terms of read power and area consumption, respectively, compared to HSC-FPGA with regular MRAM-LUTs. However, the MR-MRAM based HSC-FPGA still achieves more than 9% and 17% read power and area reduction compared to SRAM-based FPGA, respectively, while maintaining the significant advantages in standby power reductions, as well as non-volatility feature which enables fine-grained power-gating within the HSC-FPGA fabric.

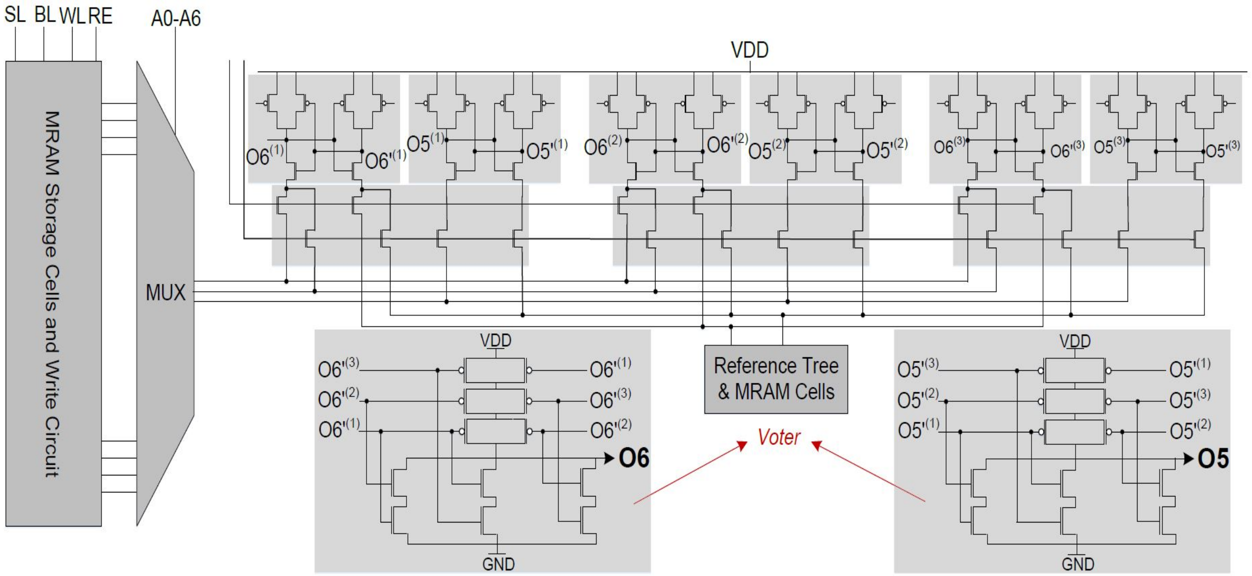


Figure 4.16: The structure of the MR-based MRAM-LUT circuit consisting of three PCSAs and two voter circuits.

Table 4.8: Comparison between the regular MRAM-LUT and MR-based MRAM-LUT circuits.

Features		Logic “0”	Logic “1”
Error Rate (%)	MRAM-LUT	42.6 %	44.8 %
	MR-based MRAM-LUT	8.2 %	16.2 %
Read Power (μW)	MRAM-LUT	2.85	3.14
	MR-based MRAM-LUT	3.21	4.22

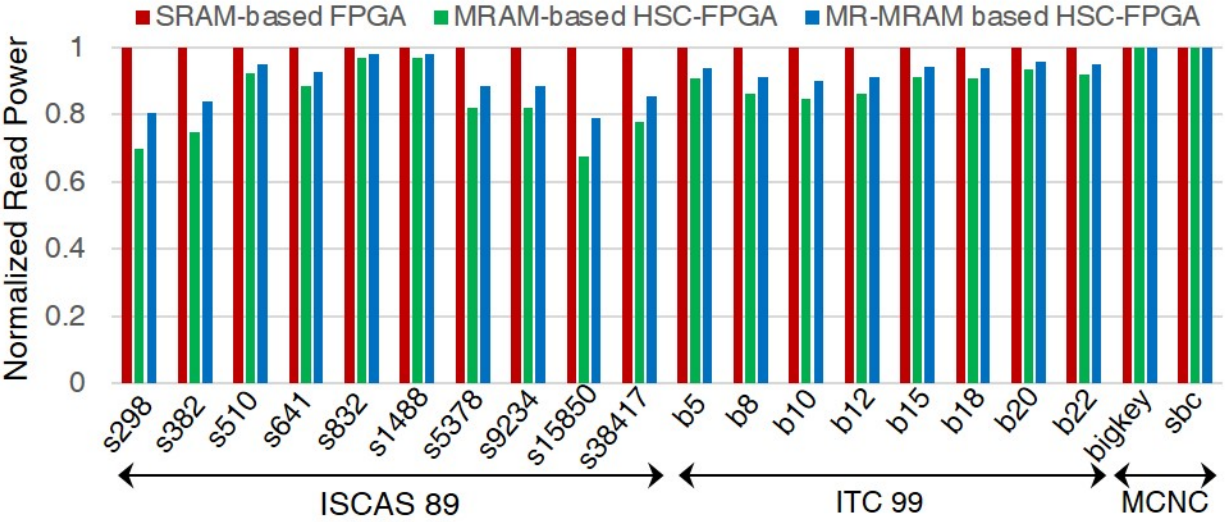


Figure 4.17: Normalized read power consumption of the regular MRAM and MR-MRAM based HSC-FPGAs compared to the SRAM-FPGA.

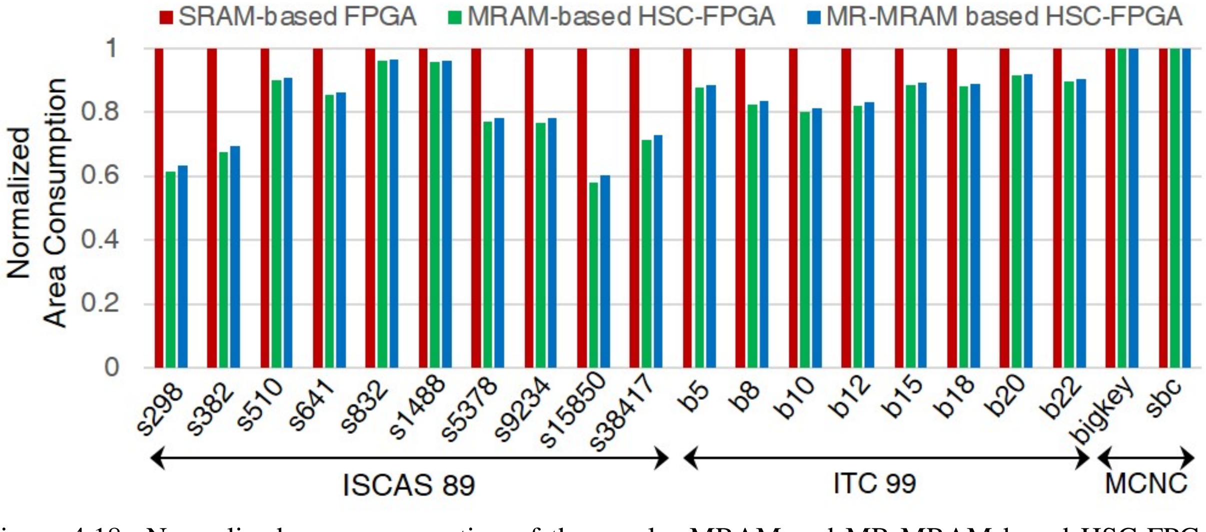


Figure 4.18: Normalized area consumption of the regular MRAM and MR-MRAM based HSC-FPGAs compared to the SRAM-FPGA.

CHAPTER 5: LOGIC PARADIGM HETEROGENEITY: LOW-ENERGY DEEP BELIEF NETWORKS USING INTRINSIC SIGMOIDAL SPINTRONIC-BASED PROBABILISTIC NEURONS[1]

An orthogonal dimension of fabric heterogeneity is also non-determinism enabled by either low-voltage CMOS or probabilistic emerging devices. It can be realized using probabilistic devices within a reconfigurable network to blend deterministic and probabilistic computational models. Herein, we will leverage the probabilistic spin logic "p-bit" device [7] as a fabric element comprising a crossbar-structured weighted array. Programmability of the resistive network interconnecting p-bit devices can be achieved by modifying the resistive states of the array's weighted connections. Thus, the programmable weighted array forms a CLB-scale macro co-processing element with bitstream programmability. This allows field programmability for a wide range of classification problems and recognition tasks to allow fluid mappings of probabilistic and deterministic computing approaches. In particular, we will focus on Deep Belief Network (DBN), which can be programmed in the field using recurrent layers of co-processing elements to form an $n \times m_1 \times m_2 \times ... \times m_i$ weighted array as a configurable hardware circuit with an n-input layer followed by $i \geq 1$ hidden layers.

The interrelated fields of machine learning (ML), and artificial neural networks (ANN) have grown significantly in previous decades due to the availability of powerful computing systems to train and simulate large scale ANNs within reasonable time-scales, as well as the abundance of data available to train such networks in recent years. The resulting research has realized a bevy of ANN architectures that have performed incredible feats including a wide range of classification problems, and various recognition tasks such as image classification and natural language processing

[1] © 2018 ACM. This chapter is reprinted, with permission, from [8].

in the form of dialog management systems [175, 176, 177].

Most ML techniques in-use today rely on supervised learning, where the systems are trained on patterns with a known desired output, or label. However, intelligent biological systems exhibit unsupervised learning whereby statistically correlated input modalities are associated within an internal model used for probabilistic inference and decision making [92]. One interesting class of unsupervised learning approaches that has been extensively researched is the Restricted Boltzmann machine (RBM) [93]. RBMs can be hierarchically organized to realize deep belief networks (DBNs) that have demonstrated unsupervised learning abilities, such as natural language understanding [94]. Most RBM and DBN research has focused on software implementations, which provides flexibility, but requires significant execution time and energy due to large matrix multiplications that are relatively inefficient when implemented on standard Von-Neumann architectures due to the memory-processor bandwidth bottleneck when compared to hardware-based in-memory computing approaches [95]. Thus, research into hardware-based RBM designs has recently sought to alleviate these constraints.

Previous hardware-based RBM implementations have aimed to overcome software limitations by utilizing FPGAs [96, 97] and stochastic CMOS [98]. In recent years, emerging technologies such as resistive RAM (RRAM) [99, 100] and phase change memory (PCM) [101] are proposed to be leveraged within the DBN architecture as weighted connections interconnecting building blocks in RBMs. While most of the previous hybrid Memristor/CMOS designs focus on improving the synapse behaviors, the work presented herein overcomes many of the preceding challenges by utilizing a novel spintronic p-bit device that leverages intrinsic thermal noise within low energy barrier nanomagnets to provide a natural building block for RBMs within a compact and low-energy package. The contribution of this work goes beyond using low-energy barrier magnetic tunnel junctions (MTJs), as has been previously introduced for a neuron in spiking neuromorphic systems [102, 103]. This is the first effort to use MTJs with near-zero energy barriers as neurons within an RBM implementation. Additionally, various parameters of a hybrid CMOS/spin weight array structure are investigated for metrics of power dissipation, and error rate using the MNIST digit recognition benchmarks.

5.1 Fundamentals of Restricted Boltzmann Machines

Restricted Boltzmann machines (RBMs) are a class of recurrent stochastic neural networks, in which each state of the network, *k*, has an energy determined by the connection weights between nodes and the node bias as described by (1), where s_i^k is the state of node *i* in *k*, b_i is the bias, or intrinsic excitability of node *i*, and w_{ij} is the connection weight between nodes *i* and *j* [178].

$$E(k) = -\sum_i s_i^k b_i - \sum_{i<j} s_i^k s_j^k w_{ij} \tag{5.1}$$

Each node in a RBM has a probability to be in state one according to (2), where σ is the sigmoid function. RBMs, when given sufficient time, reach a Boltzmann distribution where the probability of the system being in state ***v*** is found by (3), where ***u*** could be any possible state of the system. Thus, the system is most likely to be found in states that have the lowest associated energy.

$$P(s_i = 1) = \sigma(b_i + \sum_j w_{ij} s_j) \tag{5.2}$$

$$P(v) = \frac{e^{-E(v)}}{\sum_u e^{-E(u)}} \tag{5.3}$$

Restricted Boltzmann machines (RBMs) are constrained to two fully-connected non-recurrent layers called the *visible layer* and the *hidden layer*. RBMs can be readily implemented by a crossbar architecture, as shown in Figure 5.1. The most well-known approach for training RBMs is contrastive divergence (CD), which is an approximate gradient descent procedure using Gibbs sampling [179]. CD operates in four steps as described below:

1. *Feed-forward:* the training input vector, v, is applied to the visible layer, and the hidden layer, h, is sampled.

2. *Feed-back:* The sampled hidden layer output is fed-back and the generated input is sampled, v'.

3. *Reconstruct:* v' is applied to the visible layer and the reconstructed hidden layer is sampled to obtain h'.

4. *Update:* The weights are updated according to (4), where η is the learning rate and W is the weight matrix.

$$\Delta W = \eta(vh^T - v'h'^T) \tag{5.4}$$

RBMs can be readily stacked to realize a DBN, which can be trained similar to RBMs. Training a DBN involves performing CD on the visible layer and the first hidden layer for as many steps as desired, then fixing those weights and moving up a hierarchy as follows. The first hidden layer is now viewed as a visible layer, while the second hidden layer acts as a hidden layer with respect to the CD procedure identified above. Next, another set of CD steps are performed, and then the process is repeated for each additional layer of the DBN.

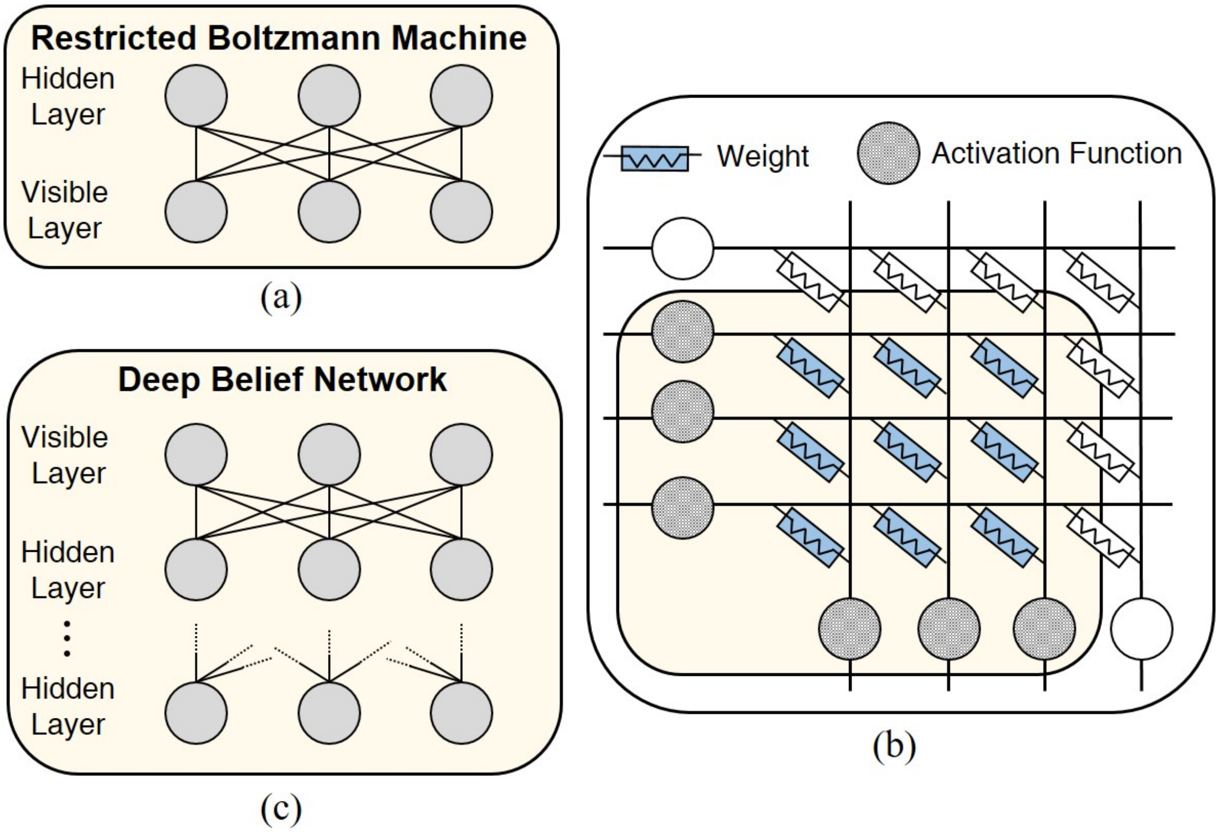


Figure 5.1: (a) An RBM structure, (b) a 3×3 RBM implemented by a 4×4 crossbar architecture, (c) a DBN structure including multiple hidden layers [6].

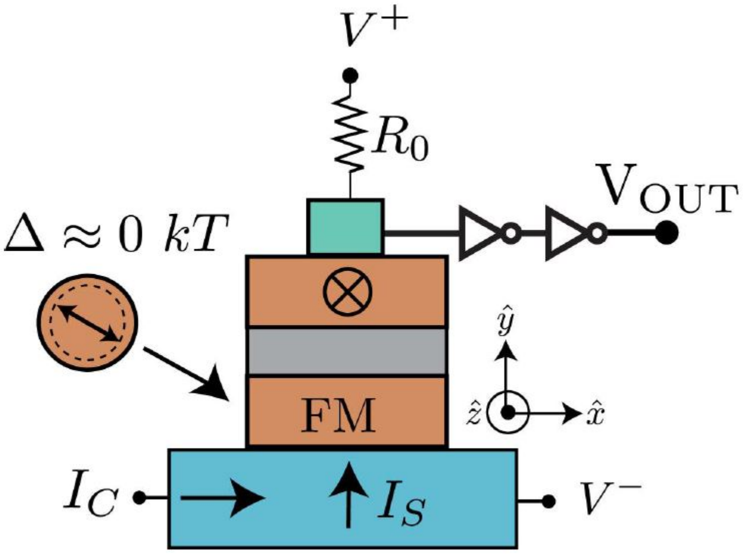


Figure 5.2: Structure of a p-bit [7].

5.2 Spin-Based Building Block For RBM

In this section, we provide a detailed description of the p-bit that provides the building block for our proposed spin-based BM architecture. Individual building blocks are interconnected by networks of memristive devices whose resistances can be programmed to provide the desired weights. For instance, in this work, we will assume that the memristive devices are implemented using the three terminal spin-orbit torque (SOT)-driven domain wall motion (DWM) device proposed in [180].

The activation function is achieved by a spintronic building block that has been used in the design of probabilistic spin logic devices (p-bits) for a wide variety of Boolean and non-Boolean problems [7, 181, 182, 183]. The basic functionality of the p-bit shown in Figure 5.2 [7] is to produce a stochastic output whose steady-state probability is modulated by an input current to generate a sigmoidal activation function. For instance, a high positive input current produces a stochastic output with a high probability of “0”, and vice versa. In the absence of any input current, the device generates either 0 or VDD outputs with roughly equal probability of 0.5, as shown in Figure 5.3.

This device consists of a 3-terminal, spin-Hall driven MTJ [133] that uses a circular, unstable nanomagnet ($\Delta \ll 40kT$), whereby its output is amplified by CMOS inverters as shown in Figure 5.2. This MTJ with an unstable free layer can be fabricated using standard technology such that the surface anisotropy to achieve perpendicular magnetic anisotropy (PMA) that is not strong enough to overcome the demagnetizing field. Thus, the magnetization stochastically rotates in the plane, due to the presence of thermal fluctuations.

The charge current that is injected to the spin-Hall layer creates a spin-current flowing into the circular FM (in the +y direction), which does not have an axis with any preferential geometry. The spin-polarization of this spin-current is in the ($\pm z$) direction, and pins the magnetization in the (+z) or (-z) direction depending on the direction of the charge current, through the spin-torque mechanism [182]. The inherent physics of the spin-current driven low-barrier nanomagnet provides a natural sigmoidal function when a long time average of magnetization is taken. Through the tunneling magnetoresistance effect, a charge current flowing through the MTJ with a stable fixed layer detects the modulated magnetization as a voltage change. To achieve this, a small read voltage V_R is applied between the $V+$ and $V-$ terminals through a reference resistance R_0, adjusted to the average conductance of the MTJ $(R_0^-1 = GP + GAP/2)$ where GP and GAP represented conductance in parallel (P) and anti-parallel (AP) states, respectively. This voltage becomes an input to the CMOS inverters that are biased at the middle point of their DC operating point, creating a stochastic output whose probability can be tuned by the input charge current.

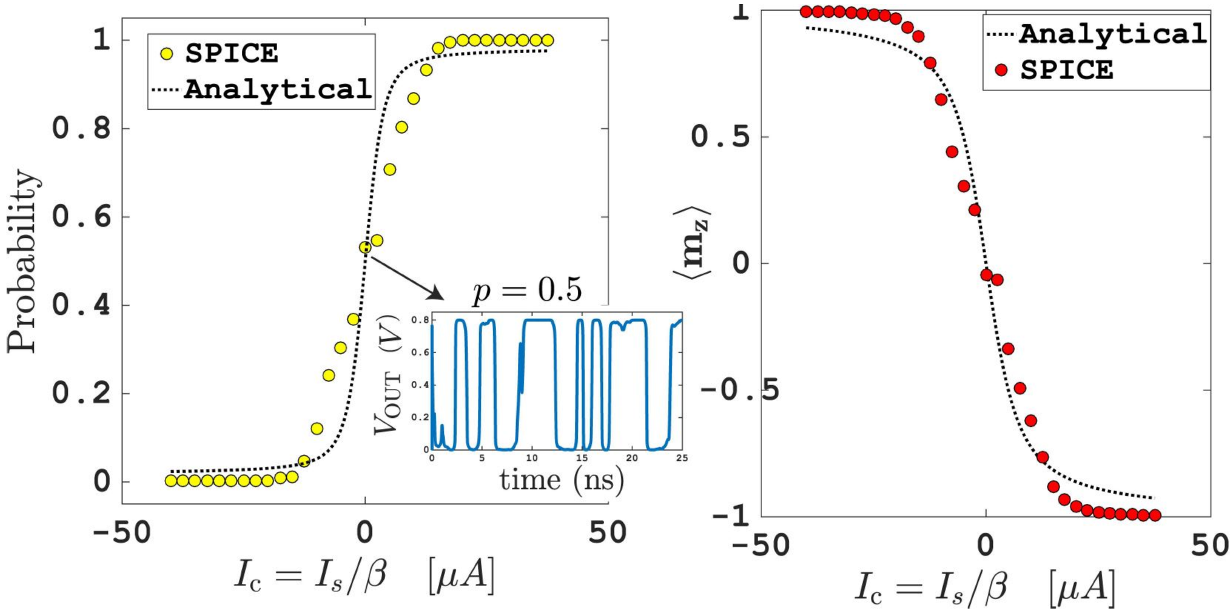


Figure 5.3: Time-averaged results over 100 ns for p-bit [8].

Each component of the device is represented by an independent spin-circuit based on experimentally-benchmarked models that have been established in [12] and simulated as a spin-circuit in a SPICE-like platform. Here, we obtain an analytical approximation to the time-averaged behavior of the output characteristics. We start by relating the charge current flowing in the spin Hall layer to the spin-current absorbed by the magnet, assuming short-circuit conditions for simplicity, i.e. 100% spin absorption by the FM:

$$I_s/I_c = \beta = \frac{L}{t}(\theta)(1 - sech(\frac{t}{\lambda})) \tag{5.5}$$

where I_s is the spin-current, I_c is the charge current, θ is the spin-Hall angle, L, t, λ are the length, thickness and spin diffusion lengths for the spin-Hall layer. The length and width of the GSHE layer are assumed to be the same as the circular nanomagnet. With a suitable choice of the L and t, the spin-current generated can be greater in magnitude than the charge current generating “gain.” For the parameters used in this work, which are listed in Table 5.1, the gain factor β is ~ 10. Next, we approximate the behavior of the magnetization as a function of an input spin-current, polarized in the ($\pm z$) direction. For a magnet with only a PMA in the $\pm z$ direction, a distribution function at steady state can be written analytically as below, as long as the spin-current is also fully in the $\pm z$ direction:

$$\rho(m_z) = \frac{1}{Z} exp(\Delta m_z^2 + 2 i_s m_z) \tag{5.6}$$

where Z is a normalization constant, m_z is the magnetization along $+z$, is the thermal barrier of the nanomagnet, and i_s is a normalization quantity for the spin-current such that $i_s = I_s/(4q/\hbar \alpha kT)$, α being the damping coefficient of the magnet, q the electron charge and $\hbar$ the reduced Planck constant. It is possible to use (4) to obtain an average magnetization $< m_z >= \int_{-1}^{+1} dm_z m_z \rho(m_z) / \int_{-1}^{+1} dm_z \rho(m_z)$. Assuming $\Delta \ll kT$, $< m_z >$ can be evaluated to give the Langevin function, $< m_z >= L(i_s)$ where $L(x) = \frac{1}{x} - coth\frac{1}{x}$, which is an exact description for the average magnetization in the presence of a z-directed spin-current for a low barrier PMA magnet.

Table 5.1: Parameters for p-bit Based Activation Function [7, 8].

Parameter	Description	Value
Circular FM		
ϕ	Diameter	$100nm$
t	Thickness	$2nm$
α	Damping coefficient	0.01
MTJ		
$G0$	Conductance	$150e^{-6}S$
P	Spin Polarization	0.52
Giant Spin Hall Layer(GSHE)		
λ	Spin-diffusion length	$2.1nm$
θ	Spin Hall Angle	0.5
$Volume$	$l \times w \times t$	$100nm \times 100nm \times 3.15nm$

In the present case, however, the nanomagnet has a circular shape with a strong in-plane anisotropy and no simple analytical formula can be derived, thus We use the Langevin function with a fitting parameter that adjusts the normalization current by a factor η, so that the modified normalization constant becomes $(4q/\hbar\alpha kT)(\eta)$. This factor increases with elevating the shape anisotropy $(H_d \sim 4\pi M_s)$ and becomes exactly one when there is no shape anisotropy. Once the magnetization and charge currents are related, we can approximate the output probability of the CMOS inverters by a phenomenological equation along with fitting parameter χ as follows:

$$p = \frac{V_{OUT}}{VDD} \approx \frac{1}{2}[1 - tanh(\chi < m_z >)] \tag{5.7}$$

which allows us to relate the input charge current to the output probability, with physical parameters. Figure 5.3 shows the comparison of the full SPICE-model with respect to aforementioned equations showing good agreement with two fitting parameters η and χ, which fit the magnetization and CMOS components, respectively.

5.3 Proposed Weighted Array Design

Figure 5.4 shows the structure of the weighted array proposed herein to implement the RBM architecture including the SOT-DWM based weighted connections and biases, as well as the p-bit based activation functions. Transmission gates (TGs) are utilized in write circuits within the bit cells of the weighted connection to adjust weights by moving the DW position. As investigated in [108], TGs can provide energy-efficient and symmetric switching operation for SOT-based devices, which are desirable during the training phase. Table 6.1 lists the required signaling for controlling the training and read operations in the weighted array structure. Herein, a chain of inverters are considered to drive signal lines, in which each successive inverter is twice as large as the previous one.

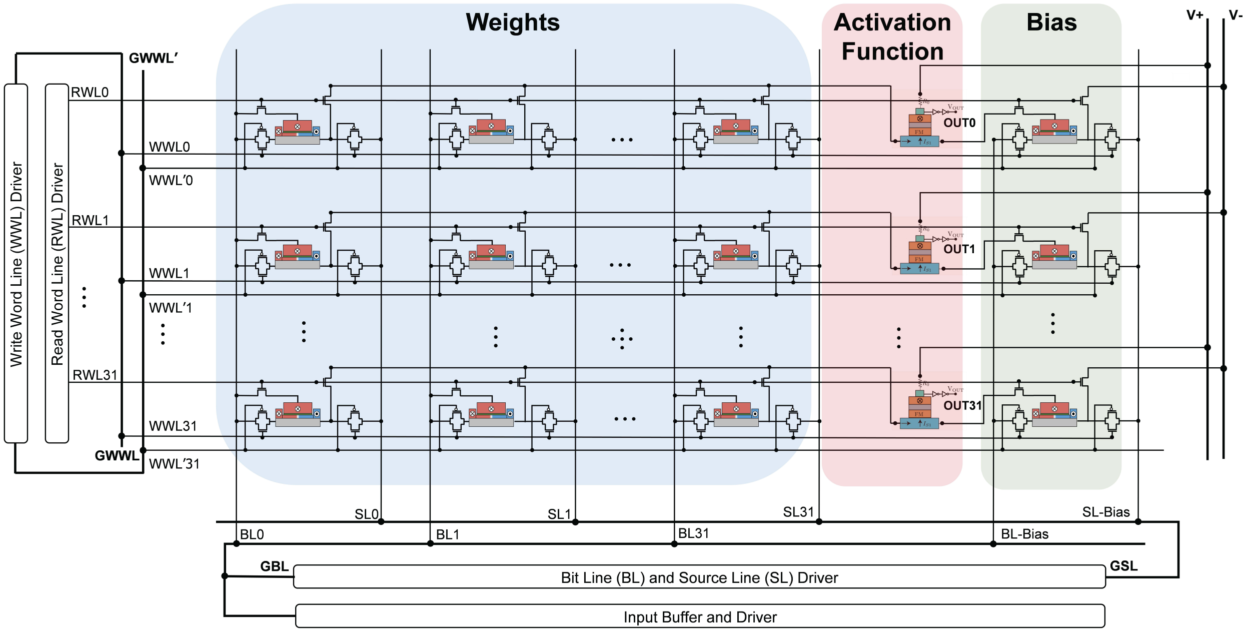


Figure 5.4: 32 × 32 hybrid CMOS/spin-based weighted array structure for RBM implementation [8].

Table 5.2: Signaling to Control The Array Operations [8].

Operation	WWL	RWL	BL	SL	V+	V-
Increase Weight	VPULSE	GND	VDD	GND	Hi-Z	Hi-Z
Decrease Weight	VPULSE	GND	GND	GND	Hi-Z	Hi-Z
Read	GND	VDD	VIN	Hi-Z	VDD	VDD/2

During the read operation, write word line (WWL) is connected to ground (GND) and the source line (SL) is in high impedance (Hi-Z) state, which disconnects the write path. The read word line (RWL) for each row is connected to VDD, which turns ON the read transistors in the weighted connection bit cell. The bit line (BL) will be connected to the input signal (VIN), which results in producing a current that affects the output probability of the p-bit device. The direction of the generated current relies on the VIN signal. In particular, since V- is supplied by a voltage source equal to VDD/2, if VIN is connected to VDD the injected current to the p-bit based activation function will have positive value, and if VIN is zero the input current will be negative. The amplitude of the generated current depends on the resistance of the weighted connection which is defined by the position of the DW in the SOT-DWM device.

Table 5.3: Relation between the input currents of activation functions and array size for $R_P = 1M\Omega$ [8].

Features	Array Size			
	8×8	16×16	32×32	64×64
Max. Positive Current (μA)	2.71	5.14	10.79	21.46
Max. Negative Current (μA)	3.57	7.14	14.23	28.28
Max. output "0" Probability	0.77	0.88	0.95	0.97
Min. output "0" Probability	0.175	0.08	0.038	0.026

During the training operation, the RWL is connected to GND, which turns OFF the read transistors and disconnects the read path. The WWL is connected to an input pulse (VPULSE) signal which activates the write path for a short period of time. The duration of the VPULSE should be designed in a manner such that it can provide the desired learning rate, η, to the training circuit. For instance, a high VPULSE duration results in a significant change in the DW position in each training iteration, which effectively reduces the number of different resistive states that can be realized by the SOT-DWM device. Resistance of the weighted connections can be adjusted by the BL and SL signals, as listed in Table 6.1. A higher resistance leads to a smaller current injected to the p-bit device. Therefore, the input signal connected to the weighted connection will have lower impact on the output probability of the p-bit device, which means the input signal exhibits a lower weight. The bias nodes can also be adjusted similar to the weighted connection.

5.4 Simulation Results And Discussion

To analyze the RBM implementation using the proposed p-bit device and the weighted array structure, we have utilized a hierarchical simulation framework including circuit-level and application-level simulations. In circuit level simulation, the behavioral models of the p-bit and SOT-DWM devices were leveraged in SPICE circuit simulations using 20nm CMOS technology with 0.9V nominal voltage to validate the functionality of the designed weighted array circuit. In application-level simulation, the results obtained from device-level and circuit-level simulations are used to implement a DBN architecture and analyze its behavior in MATLAB.

5.4.1 Circuit-level simulation

The device-level simulations shown in Figure 5.3 verified a sigmoidal relation between the input current of the p-bit based activation function and its output probability. The shape of the activation on function is one of the major factors affecting the performance of the RBM. Therefore, we have provided comprehensive analyses on the impacts of weighted connection resistance and weighted array dimensions on the input currents of the p-bit based activation functions, and the power consumption of the weighted array.

Table 5.3 lists the range of the activation function input currents for various weighted array dimensions, while the resistance of the SOT-DWM device in parallel state (R_P) is constant and equals $1M\Omega$. The experimental results provided in [184, 185] exhibit that an MTJ resistance in the $M\Omega$ range can be obtained by increasing the oxide thickness in an MTJ structure. The highest positive and negative currents can be achieved while the weighted connections are in parallel state, i.e. lowest resistance, and all of the input voltages (VIN) are equal to VDD and GND, respectively. The difference between the amplitude of positive and negative currents in a given array size with constant R_P is caused by the different pull-down and pull-up strengths in NMOS read transistors. The maximum and minimum output-level “0” probabilities are listed in Table 5.3, which can be obtained according to the measured input currents and the sigmoidal activation function shown in Figure 5.3.

Moreover, Table 5.4 illustrates the relation between the R_P values and input currents of the activation functions, and their corresponding output probabilities, for a given 32×32 weighted array. The lower R_P resistance and higher array size provides a wider range of output probabilities which can increase the RBM performance. However, this is achieved at the cost of higher area and power consumption. The trade-offs between the array size, weighted connection resistance, and average power consumption in a single read operation is shown in Figure 5.5. The lowest power consumption of 22.6 μW is realized by an 8×8 array with $R_P = 1M\Omega$. However, this array provides the narrowest range of the output probabilities, which significantly reduces the DBN's performance.

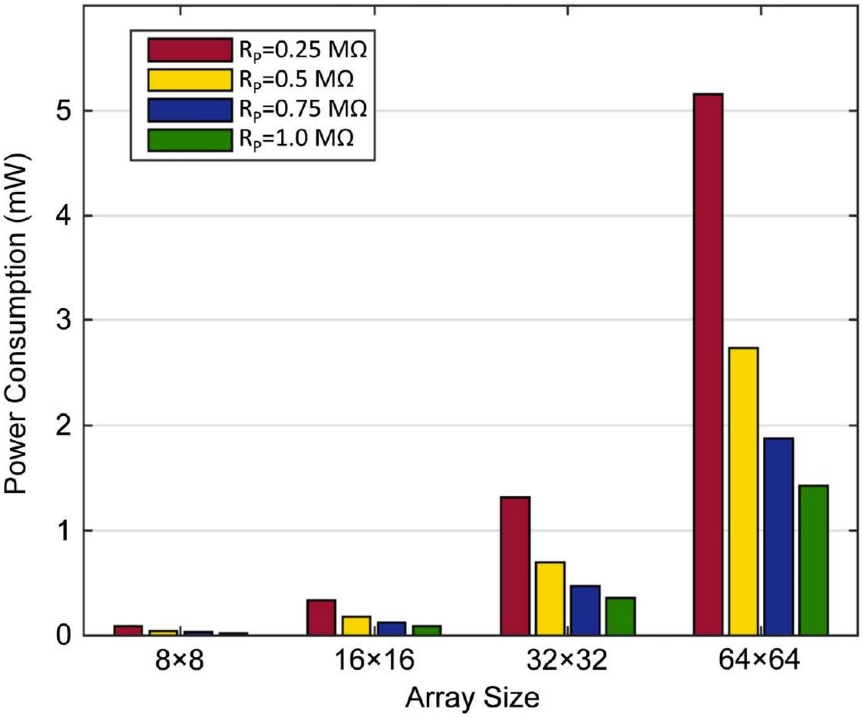


Figure 5.5: Weighted array power consumption versus the resistance of the weighted connections and array size [8].

Table 5.4: Relation between the input currents of activation functions and R_P in a 32×32 array [8].

Features	$R_P(M\Omega)$			
	0.25	0.5	0.75	1
Max. Positive Current (μA)	36.56	20.02	13.97	10.79
Max. Negative Current (μA)	54.95	28.12	18.9	14.23
Max. output “0” Probability	0.98	0.965	0.96	0.95
Min. output “0” Probability	0.01	0.026	0.032	0.038

5.4.2 Application-level simulation

In the application-level simulation, we have leveraged the obtained device and circuit behavioral models to simulate a DBN architecture for digit recognition. In particular, learning rate and the shape of the sigmoid activation function is extracted by the SOT-DWM and p-bit device-level simulations, respectively, while the circuit-level simulations defines the range of the output probabilities. To evaluate the performance of the system, we have modified a MATLAB implementation of DBN by Tanaka and Okutomi [186] and used the MNIST data set [187] including 60,000 and 10,000 sample images with 28×28 pixels for training and testing operations, respectively. We have used Error rate (ERR) metric to evaluate the performance of the DBN, as expressed by $ERR = N_F/N$, where, N is the number of input data, N_F is the number of false inference [186].

The simplest model of the DBN that can be implemented for MNIST digit recognition consists 784 nodes in visible layer to handle 28×28 pixels of the input images, and 10 nodes in hidden layer representing the output classes. Figure 5.6 shows the relation between the performance of various DBN topologies, and the number of input training samples ranging from 100 to 5,000, which is obtained using 1,000 test samples. The ERR and RMSE metrics can be improved by enlarging the DBN structure through increasing the number of hidden layers, as well as the number of nodes in each layer. This improvement is realized at the cost of larger area and power consumptions. Increasing the input training samples can improve the DBN performance as well, however it will quickly converge due to the limited weight values that can be provided by SOT-DWM based weighted connections. As shown in Figure 5.6, some random behaviors are observed for networks with smaller sizes that are trained by lower number of training samples, which will be significantly reduced by increasing the number of training samples.

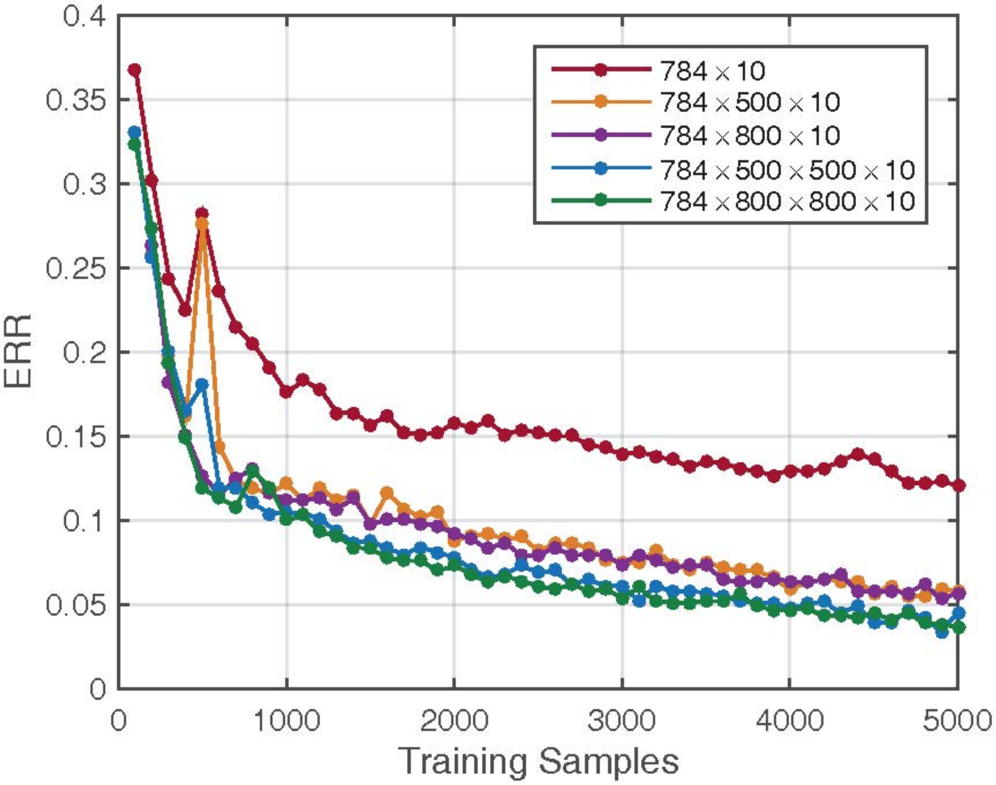


Figure 5.6: ERR for various DBN topologies [8].

The simulation results exhibit the highest error rate of 36.8% for a 784×10 DBN that is trained by 100 training samples. Meanwhile, the lowest error rate of 3.7% was achieved using a $784 \times 800 \times 800 \times 10$ DBN trained by 5,000 input training samples. This illustrates that the recognition error rate can be decreased by increasing the number of hidden layers, and training samples, which is also realized at the cost of higher area and power overheads.

5.4.3 Disucussion

Table 6.3 lists previous hardware-based RBM implementations, which have aimed to overcome software limitations by utilizing FPGAs [96, 97], stochastic CMOS [98], and hybrid memristor-CMOS designs [99, 100, 101]. FPGA implementations demonstrated RBM speedups of 25-145 over software implementations [96, 97], but had significant constraints such as only realizing a single 128×128 RBM per FPGA chip, routing congestion, and clock frequencies limited to 100MHz [97]. In [188], optimization methods are proposed to reduce memory requirements for weights and biases, however implementing each activation function still requires dedicated piecewise linear approximator, random number generator (RNG), and comparator circuits which lead to increased area and energy consumption per neuron than the p-bit based approach herein. The stochastic CMOS-based RBM implementation proposed in [98] leveraged the low-complexity of stochastic CMOS arithmetic to save area and power. However, the need for extremely long bit-stream lengths negate energy savings and lead to very long latencies. Additionally, a significant amount of Linear Feedback Shift Registers (LFSRs) were required to produce the uncorrelated input and weight bit-streams. In both the FPGA and stochastic CMOS designs, improvements were achieved by implementing parallel Boolean circuits such as multipliers and pseudo-random number generators for probabilistic behavior, which has significant area and energy overheads compared to leveraging the physical behaviors of emerging devices to perform the computation intrinsically. Bojnordi et al. [100] leveraged resistive RAM (RRAM) devices to implement efficient matrix multiplication for weighted products within Boltzmann machine applications, and demonstrated significant speedup of up to 100-fold over single-threaded cores and energy savings of over 10-fold. Similarly, Sheri et al. [99] and Eryilmaz et al. [101] utilized RRAM and PCM devices to implement matrix multiplication, while the corresponding activation function circuitry is still based on the CMOS technology, which suffers from the aforementioned area and power consumption overheads.

Table 5.5: Various DBN hardware implementations with a focus on activation function structure [8, 9].

Design	Weighted Connection	Activation Function	Energy per Neuron	Normalized area per neuron
[96]	Embedded multipliers	CMOS-based LUTs	N/A	N/A
[97]	Embedded multipliers	- 2-kB BRAM - Piecewise Linear Interpolator - Random number Generator	∼10-100 nJ	∼ 3000×
[188]	- Multiplier - Adder tree	- Piecewise Linear approximator - Random number Generator - Comparator	∼10-100 nJ	∼ 2000×
[98]	- LFSR - bit-stream - AND/OR gates	-LFSR - Bit-wise AND - tree adder - FSM-based *tanh* unit	∼10-100 nJ	∼ 90×
[99]	RRAM Memristor	Off-chip	N/A	N/A
[100]	RRAM	- 64 × 16 LUTs - Pseudo Random Number Generator - Comparator	∼1-10 nJ	∼ 1250×
[101]	PCM	Off-chip	N/A	N/A
Proposed Herein	SOT-DWM	p-bit	1-10 fJ	1×

While most of the previous hybrid Memristor/CMOS designs focus on improving the performance of weighted connections, the work presented herein overcomes many of the preceding challenges of generating sigmoidal probabilistic activation functions by utilizing a novel p-bit device that leverages intrinsic thermal noise within low energy barrier nanomagnets to provide a natural building block for RBMs within a compact and low-energy package. As listed in Table 7.3, the proposed design can achieve approximately three orders of magnitude improvement in term of energy consumption compared to the most energy-efficient designs, while realizing at least 90X device count reduction for considerable area savings. Note that these calculations do not take into account the weighted connections, since the main focus of this work is on the activation function. While SOT-DWM devices are utilized herein for the weighted connections, any other memristive devices could be utilized without loss of generality.

CHAPTER 6: SNRA: A SPINTRONIC NEUROMORPHIC RECONFIGURABLE ARRAY FOR IN-CIRCUIT TRAINING AND EVALUATION OF DEEP BELIEF NETWORKS[1]

Within the post-Moore era ahead, several design factors and fabrication constraints increasingly emphasize the requirements for in-circuit adaptation to as-built variations. These include device scaling trends towards further reductions in feature sizes [104], the narrow operational tolerances associated with the deployment of hybrid Complementary Metal Oxide Semiconductor (CMOS) and post-CMOS devices [91, 105], and the noise sensitivity limits of analog-assisted neuromorphic computing paradigms [106]. While many recent works have advanced new architectural approaches for the evaluation phase of neuromorphic computation utilizing emerging hardware devices, there have been comparatively fewer works to investigate the hardware-based realization of their training and adaptation phases that will also be required to cope with these conditions. Thus, this work develops one of the first viable approaches to address post-fabrication adaptation and retraining in-situ of resistive weighted-arrays in hardware, which are ubiquitous in post-Moore neuromorphic approaches. Namley, a tractable in-field reconfiguration-based approach is developed to leverage in-field configurability to mitigate the impact of process variation. Reconfigurable fabrics are characterized by their fabric flexibility, which allows realization of logic elements at medium and fine granularities, as well as in-field adaptability, which can be leveraged to realize variation tolerance and fault resiliency as widely-demonstrated for CMOS-based approaches such as [31, 39]. Utilizing reconfigurable computing by applying hardware and time redundancy to the digital circuits offers promising and robust techniques for addressing the above-mentioned reliability challenges. For instance, it is shown in [39] that a successful refurbishment for a circuit with 1,252 look-up tables (LUTs) can be achieved with only 10% spare resources to accommodate both

[1] © 2018 IEEE. This chapter is reprinted, with permission, from [6].

soft and hard faults.

Within the post-Moore era, reconfigurable fabrics can also be expected to continue their transition towards embracing the benefits of increased heterogeneity along several cooperating dimensions to facilitate neuromorphic computation [1]. Since the inception of the first field-programmable devices, various granularities of general-purpose reconfigurable logic blocks and dedicated function-specific computational units have been added to their structures. These have resulted in increased computational functionality compared to homogeneous architectures. In recent years, emerging technologies are proposed to be leveraged in reconfigurable fabrics to advance new transformative opportunities for exploiting technology-specific advantages. Technology heterogeneity recognizes the cooperating advantages of CMOS devices for their rapid switching capabilities, while simultaneously embracing emerging devices for their non-volatility, near-zero standby power, high integration density, and radiation-hardness. For instance, spintronic-based LUTs are proposed in [4, 189, 171] as the primary building blocks in reconfigurable fabrics realizing significant area and energy consumption savings. In this chapter, we extend the transition toward heterogeneity along various logic paradigms by proposing a heterogeneous technology fabric realizing both probabilistic and deterministic computational models. The cooperating advantages of each are leveraged to address the deficiencies of the others during the neuromorphic training and evaluation phases, respectively.

In this chapter, we propose a spintronic neuromorphic reconfigurable Array (SNRA) that uses probabilistic spin logic devices to realize deep belief network (DBN) architectures while leveraging deterministic computing paradigms to achieve in-circuit training and evaluation. Most of the previous DBN research has focused on software implementations, which provides flexibility, but requires significant execution time and energy due to large matrix multiplications that are relatively inefficient when implemented on standard Von-Neumann architectures. Previous hardware-based implementation of RBM have sought to overcome software limitations by using FPGAs [96, 97], stochastic CMOS [98], and hybrid memristor-CMOS designs [100]. Recently, Zand et al. [8] utilized a spintronic device that leverages intrinsic thermal noise within low energy barrier nanomagnets to provide a natural building block for RBMs. While most of the aforementioned designs only focus on the test operation, the work presented herein concentrates on leveraging technology heterogeneity to implement a train and evaluation circuitry for DBNs with various network topologies on our proposed SNRA fabric.

6.1 Proposed RBM Structure

A feasible hardware implementation of a 4×2 RBM structure is shown in Figure 6.1(a), in which three terminal spin Hall effect (SHE)-driven domain wall motion (DWM) device [184] is used as weights and biases, while the probabilistic spin logic devices (p-bits) are utilized to produce a probabilistic output voltage that has a sigmoid relation with the input currents of the devices, as shown in Figure 6.1(b) and Figure 6.1(c), respectively. The p-bit device consists of a SHE-driven magnetic tunnel junction (MTJ) with a circular near-zero energy barrier nanomagnet, which provides a natural sigmoidal activation function required for DBNs as studied in [7, 181, 182, 183]. Transmission gates (TGs) are used within the bit cell of the weighted connections to adjust the weights by changing the domain wall (DW) position in SHE-DWM devices, as well as controlling the RBM operation phases. TGs can provide an energy-efficient and symmetric switching behavior [108], which is specifically desired during the training operation.

Table 6.1 lists the required signaling to control the RBM's training and test operations. During the feed-forward, feed-back, and reconstruct operations, write word line (WWL) is connected to ground (GND) and the bit line (BL) and source line (SL) are both in high impedance (Hi-Z) state disconnecting the write path. The read word line (RWL) is connected to VDD, which turns ON the read TGs in the weighted connection bit cell shown in Figure 6.1(b). The voltage applied by the input neuron generates a current through TG1 and TG2, which is then injected to the output neuron and modulates the output probability of the p-bit device. The amplitude of the current depends on the resistance of the weighted connection which is defined by the position of the DW in the SHE-DWM device.

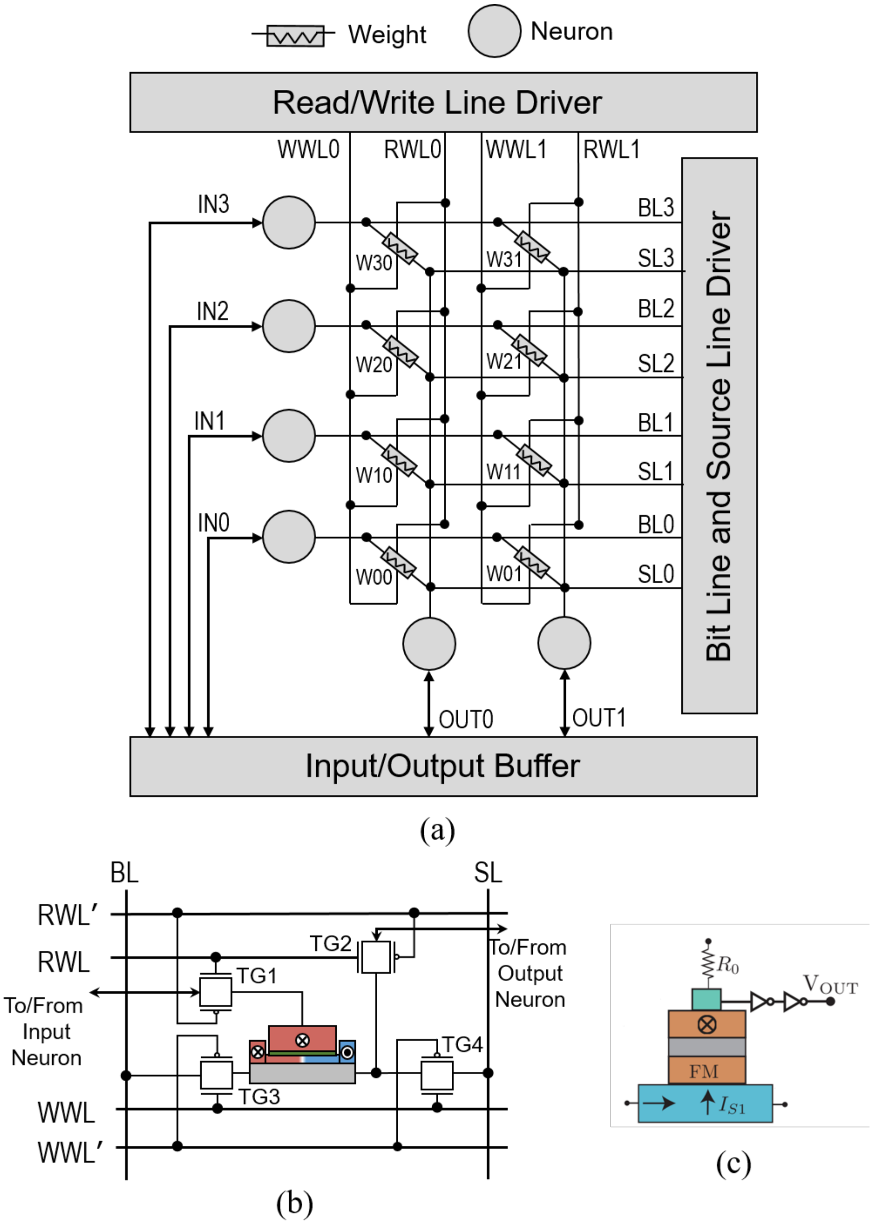


Figure 6.1: (a) A 4×2 RBM hardware implementation [6], (b) SHE-DWM based weighted connections, and (c) p-bit based probabilistic neuron [7].

Table 6.1: Required signaling to control the RBM operation phases [6].

Operation Phase		WWL	RWL	BL	SL
Feed-Forward / Test Reconstruct Feed-Back		GND	VDD	Hi-Z	Hi-Z
Update	Increase Weight	VDD	GND	Vtrain	GND
	Decrease Weight			GND	Vtrain

During the update phase, the RWL is connected to GND, which turns off TG1 and TG2 and disconnects the read path. Meanwhile, the WWL is set to VDD which activate the write path. Resistance of the weighted connections can be adjusted by the BL and SL signals, as listed in Table 6.1. The amplitude of the training voltage (Vtrain) connected to BL and SL should be designed in a manner such that it can provide the desired learning rate, η, to the training circuit. For instance, a high amplitude $Vtrain$ results in a significant change in the DW position in each training iteration, which effectively reduces the number of different resistive states that can be realized by the SHE-DWM device. On the other hand, a higher SHE-DWM resistance leads to a smaller current injected to the p-bit device. Thus, the input signal connected to the weighted connection with higher resistance will have lower impact on the output probability of the p-bit device, representing a lower weight for the corresponding connection between the input and output neurons.

6.2 Proposed Hardware Implementation of Contrastive Divergence Algorithm

To implement the contrastive divergence (CD) algorithm required for training the weights in an RBM structure, we have designed a four-state finite state machine (FSM) as shown in Figure 6.2. The proposed FSM is in the *feed-forward* state during the test operation. When the training begins, the input of the visible layer and the corresponding output of the hidden layer will be stored in the ***v*** and ***h*** registers, respectively. The size of the ***v*** and ***h*** registers depend on the number of neurons in the visible and hidden layers. For instance, in the sample 4×2 RBM shown in Figure 6.1 the size of the ***v*** and ***h*** registers are 4-bits and 2-bits, respectively. In the *feed-back* state, the sampled hidden layer is fed-back to the RBM array and the corresponding output of the visible layer is stored in the ***v_bar*** register. Next, the stored values in ***v_bar*** are applied to the RBM to reconstruct the hidden layer, and the obtained output of the hidden layer will be stored in ***h_bar*** register. Finally in the *update* state, the data stored in ***v***, ***h***, ***v_bar***, and ***h_bar*** registers are used to provide the required BL and SL signals to adjust the weights according to Equation (5.4).

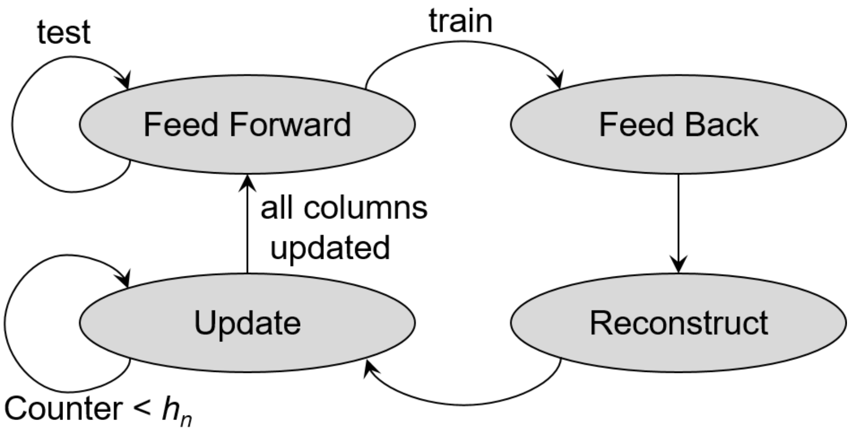
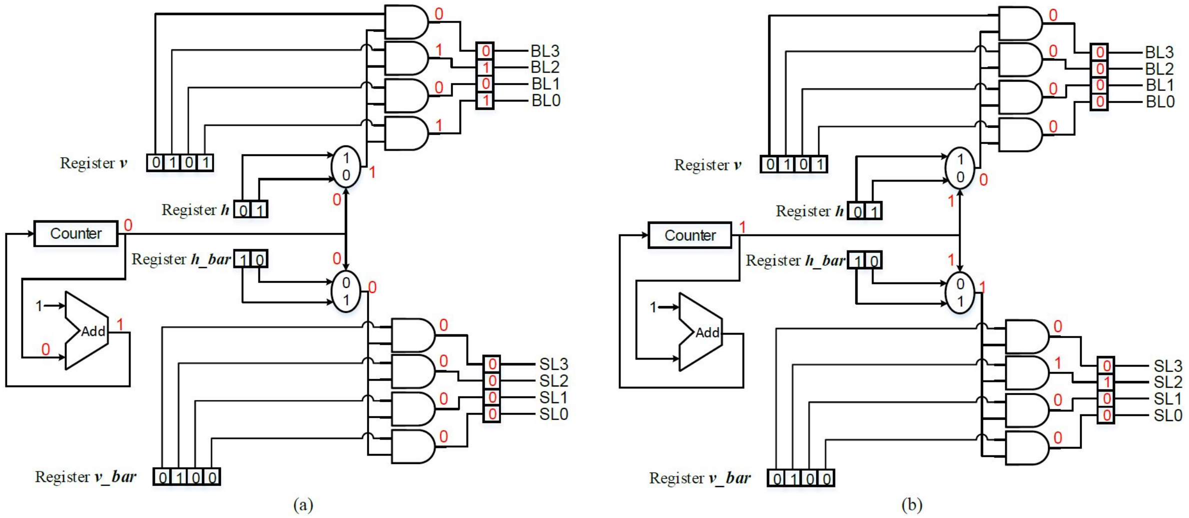


Figure 6.2: FSM designed to control the train and test operations in a DBN [6].



(a)

(b)

Figure 6.3: The hardware realization for the *update* state in the FSM developed to train a 4×2 RBM, (a) first clock cycle, and (b) second clock cycle [6].

Figure 6.3 depicts the schematic of the hardware designed for the *update* state of the FSM developed for a 4×2 RBM. In each clock cycle, The designed circuit adjusts the weights in a single column of the RBM shown in Figure 6.1. Thus, the number of clock cycles required to complete the update state depends on the number of neurons in the hidden layer of the RBM. A *counter* register is used in the design to ensure that all of the columns in the RBM are updated. The counter value starts from zero and will be incremented in each clock cycle until it reaches the h_n value, which is the total number of nodes in the hidden layer. Once the counter reaches h_n, the update state is completed and the FSM goes to the *feed-forward* state. The logical AND gates are used to implement the vh^T and $v'h'^T$ expressions required to find ΔW for the weights in each column. The output of Boolean gates implementing vh^T and $v'h'^T$ are stored in ***BL_reg*** and ***SL_reg*** registers, respectively, which provide the required signaling for adjusting the weights according to the Table 6.1.

Herein, to better understand the functionality of the hardware developed for the *update* state, we have used an example with the v, h, v', and h' matrices having the hypothetical values mentioned below:

$$v = \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad h = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad v' = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad h' = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Hence, the ΔW can be calculated using (4) as shown below:

$$\Delta W = \eta(vh^T - v'h'^T) = \eta \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & -1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \delta w_{00} & \delta w_{01} \\ \delta w_{10} & \delta w_{11} \\ \delta w_{20} & \delta w_{21} \\ \delta w_{30} & \delta w_{31} \end{bmatrix}$$

According to the obtained ΔW, w_{21} should be decreased while the w_{00} and w_{20} increases, and the remaining weight values remain unchanged. The hardware realization of the mentioned example is shown in Figure 6.3, in which the values stored in the registers are ***v***=4'b0101, ***h***=2'b01, ***v_bar***=4'b0100, and ***h_bar***=2'b10. It is worth noting that, the v_0 element in the v matrix is stored in the least significant bit of the ***v*** register, while v_3 is stored in the most significant bit. Other matrices are stored to their corresponding registers in the similar manner. In this example, RBM has two output neurons, therefore h_n is equal to two and the update operation can be completed in two clock cycles. In the first cycle shown in Figure 6.3(a), the counter is equal to zero and the first bits of ***h*** and ***h_bar*** registers are selected by the multiplexers to be used as the input of the AND gates. Therefore, the below BL and SL signals are generated,

$$BL = \begin{bmatrix} BL0 \\ BL1 \\ BL2 \\ BL3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad SL = \begin{bmatrix} SL0 \\ SL1 \\ SL2 \\ SL3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

As listed in Table 6.1, the above BL and SL signals will increase w_{00} and w_{20} weights shown in Figure 6.1, if the WWL0 and WWL1 signals are "1" and "0", respectively. Similarly, in the second clock cycle, the counter is equal to one and the second bits of ***h*** and ***h_bar*** registers are used to produce below BL and SL signals as below,

$$BL = \begin{bmatrix} BL0 \\ BL1 \\ BL2 \\ BL3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad SL = \begin{bmatrix} SL0 \\ SL1 \\ SL2 \\ SL3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

This results in a decrease in the w_{21} weight, while the other weights remain unchanged. Thus, the proposed hardware provides the desired functionality required for the *update* state according to Equation (5.4).

Herein, we have used the Verilog hardware description language (HDL) to implement our proposed four-state FSM. The ModelSim simulator is used to simulate the developed register-transfer level (RTL) Verilog codes. Figure 6.4 shows the obtained waveforms required for training a 4×2 RBM array with the hypothetical register values mentioned above. The results show that the desired BL, SL, RWL, and WWL control signals are generated in five clock cycles, which verifies the functionality of our proposed FSM.

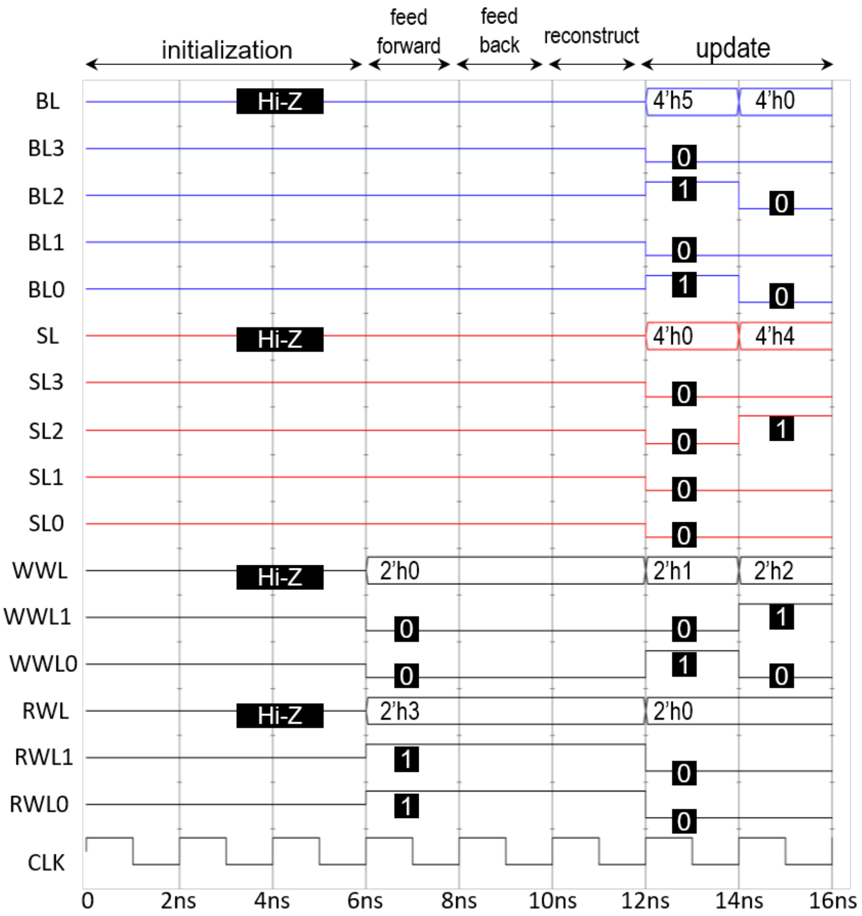
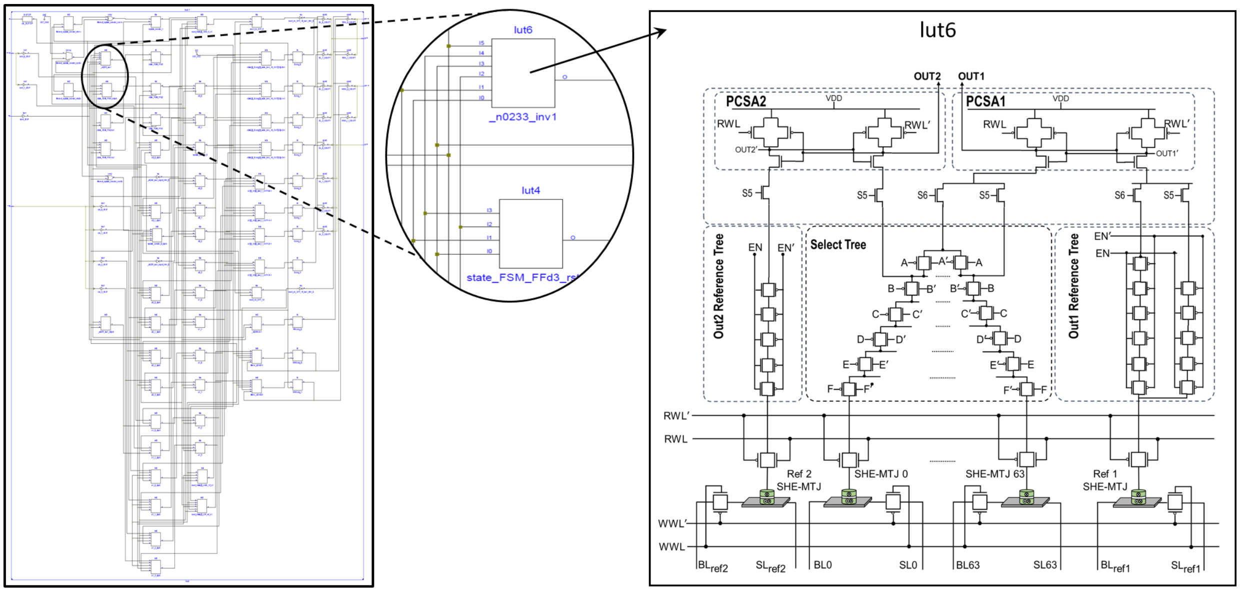


Figure 6.4: The output signals generated by the proposed FSM. The clock frequency is 500MHz, which can be modified based on the design requirements [6].



(a) (b)

Figure 6.5: (a) The schematic of the hardware designed to control the testing and training operations of a 4×2 RBM, (b) the structure of a 6-input SHE-MTJ based fracturable LUT used as the building block of the proposed SNRA architecture [6].

To obtain the hardware resources required for our proposed DBN control circuitry, we have synthesized and implemented it using Xilinx ISE Design Suite 14.7. The schematic of the hardware developed to control the testing and training operations for a 4×2 RBM is shown in Figure 6.5(a), in which 32 six-input fracturable look-up table (LUT) and Flip Flop (FF) pairs are used to implement both sequential and combinational logic. It is worth noting that out of the 32 LUT-FF pairs only three of them are utilized for the test operation, thus roughly 90% of the circuit can be power-gated during the test operation. However in conventional homogeneous technology FPGAs, volatile static random access memory (SRAM) cells are employed in LUTs to store the logic function configuration data. Therefore, by power-gating the SRAM-based LUTs the configuration data will be lost and the FPGA is required to be re-programmed. In addition to volatility, SRAM cells also suffer from high static power and low logic density [69]. Hence, alternative emerging memory technologies have been attracting considerable attention in recent years as an alternative for SRAM cells.

6.3 The proposed SNRA architecture

Herein, we propose a heterogeneous-technology spintronic neuromorphic reconfigurable array (SNRA), which can combine both deterministic and probabilistic logic paradigms. The SNRA fabric is organized into islands of probabilistic modules surrounded by Boolean configurable logic blocks (CLBs). Both the probabilistic and deterministic elements are field programmable using a configuration bit-stream based on conventional FPGA programming paradigms.

Herein, the probabilistic modules consist of RBMs, which can be connected hierarchically within the field-programmable fabric to form various topologies of DBNs. Each RBM leverages SHE-MTJs with unstable nanomagnets ($\Delta \ll 40kT$) to generate the probabilistic sigmoidal activation function of the neurons. With respect to the deterministic logic, the CLBs are comprised of LUTs which realize the training and evaluation circuitry. Non-volatile high energy barrier ($\Delta \geq 40kT$) SHE-MTJ devices are used as an alternative for SRAM cells within LUT circuits. The routing networks include routing tracks, as well as switch and connection blocks similar to that of the conventional FPGAs. The feasibility of integrating MTJs and CMOS technologies in an FPGA chip has been verified in 2015 by researchers in Tohoku University [189]. They have fabricated a non-volatile FPGA with 3,000 6-input MTJ-based LUTs under 90nm CMOS and 75nm MTJ technologies. The measurement of fabricated devices under representative applications exhibited significant improvements in terms of power consumption and area. Despite the mentioned improvements, the conventional spin transfer torque (STT)-based MTJ devices suffer from high switching energy and reliability issues. Thus, we propose using SHE-MTJ based LUT circuits with reduced switching energy and increased reliability of tunneling oxide barrier [13]. Readers are referred to [124] for additional information regarding the STT-MTJ and SHE-MTJ devices.

Figure 6.5(b) shows the structure of a six-input SHE-MTJ based fracturable LUT [2], which can implement a six-input Boolean function or two five-input Boolean functions with common inputs.

In general, LUT is a memory with 2^m cells in which the truth table of an m-input Boolean function is stored. The logic function configuration data is stored in SHE-MTJs in form of different resistive levels determined based on the magnetization configurations of ferromagnetic layer in MTJs, i.e parallel configuration results in a lower resistance standing for logic “0” and vice versa. The LUT inputs can be considered as the address according to which corresponding output of the Boolean function will be returned through the select tree. The LUT circuit shown in Figure 6.5(b) includes two pre-charge sense amplifiers (PCSAs) that are used to read the logic state of the SHE-MTJs. The PCSA compares the stored resistive value of the SHE-MTJ cells in the LUT circuit with a reference MTJ cell that its resistance is designed between the low and high resistances of the LUT’s SHE-MTJ cells. Therefore, if the resistive value of a SHE-MTJ cell in the LUT circuit is greater than the resistance of the reference cell, the output of the PCSA will be “1” and vice versa. The readers are referred to Chapter 3 for additional information regarding the functionality of a SHE-MTJ based LUT circuit.

6.4 Results and Discussions

Herein, we have modified a MATLAB implementation of DBN developed in [186] and utilized MNIST data set [187] to calculate the error rate and evaluate the performance of our DBN architecture. The simplest model of the belief network that can be used for MNIST digit recognition includes a single RBM with 784 nodes in the visible layer to handle 2828 pixels of the input images, and 10 nodes in hidden layer representing the output classes. Herein, we have examined the error rate for five different network topologies using 1,000 test samples as shown in Figure 5.6. As it is expected, increasing the number of the hidden layers, nodes, and training images improves the performance of the DBN, however these improvements are realized at the cost of higher area and power dissipation.

To compare the resource utilization between the five network topologies investigated in this work, we have used Xilinx ISE Design Suite 14.7 to implement their control circuitry based on the proposed FSM design. The obtained logic resource utilization for each of the mentioned DBN topologies is listed in Table 6.2. Since the training operation in different layers of the DBN does not happen simultaneously, the resources can be shared for training each RBM. Therefore, the amount of logic resources utilized to implement the FSM of a DBN relies on the size of the largest RBM in the network. For instance, as listed in Table 6.2, the resource utilization for training a 784×500×10 DBN is equal to that of a 784×500×500×10 DBN, since the size of the largest RBM in both networks is 784×500.

Table 6.2: FSM logic resource utilization and power dissipation for various DBN topologies [6].

Topology	Slice Registers	Slice LUTs	Fully-used LUT-FFs	Power Consumption
784×10	3185	123	51	0.32 mW
784×500×10	4655	3545	1771	14.2 mW
784×800×10	5533	2449	2421	19.3 mW
784×500×500×10	4655	3545	1771	25.3 mW
784×800×800 ×10	5617	2449	2421	34.5 mW

Table 6.3: Comparison between six-input fracturable SRAM-based LUT and SHE-MTJ based LUT [6].

Features		SRAM-LUT	SHE-MTJ LUT
Device Count	MOS	1163	565
	MTJ	-	66
Power (μW)	Read	6.28	1.1
	Write	28	188
	Static	1.6	0.21
Delay	Read	< 10 ps	< 30 ps
	Write	< 0.1 ns	< 2 ns
Energy	Read	~ 62.8 aJ	~ 33 aJ
	Write	~ 2.8 fJ	~ 376 fJ

To provide a fair power consumption comparison between the investigated DBN topologies, we have simulated an SRAM-based six-input fracturable LUT-FF pair in SPICE circuit simulator using 45nm CMOS library with 1.0V nominal voltage. The obtained static and dynamic power dissipation are listed in Table 6.3. Herein, we have only focused on the power dissipated by the LUT-FF pairs, and used the below relation to measure the power consumption for each topology:

$$P_{total} = \sum_{i} A_i P_{read} + I_i P_{standby} \tag{6.1}$$

where A_i and I_i are the number of active and idle LUT-FF pairs in RBM i of the DBN, respectively. The obtained power dissipation values for various DBN topologies are listed in the last column of Table 6.2. The provided trade-offs between the error rate and power consumption can be leveraged to design a desired DBN based on the application requirements.

To investigate the effect of *technology heterogeneity* on the performance of the proposed DBN control circuitry, we have simulated a SHE-MTJ based six-input fracturable LUT in SPICE using 45nm CMOS and 60nm MTJ technologies. The modeling approach proposed in [2, 107] is leveraged to model the behavior of SHE-MTJ devices. In particular, first, a Verilog-A model of the device is developed and used in SPICE to obtain the write current, as well as the power dissipation of the read/write operations. Next, the write current is used in a descriptive MATLAB model of a SHE-MTJ device to extract the corresponding write delay. The simulation results obtained for a SHE-MTJ based six-input fracturable LUT circuit are listed in Table 6.3.

Three types of power consumption profiles can be identified in FPGA LUTs. During the configuration phase, the LUTs must be initialized and thus written. This incurs an initial write energy consumption, which occurs infrequently thereafter. Second, upon configuration the LUTs comprising active logic paths will consume read power including a certain sub areas within high gate equivalent capacity of FPGA chips. Third, the remainder of the LUTs, which can be a large number, may be inactive and consume standby power. SRAM-based FPGA is challenged by the difficulty with power-gating LUTs which must retain the stored configuration. While, a SHE-MTJ based LUT can be readily power-gated and incur near-zero standby energy due to its non-volatility characteristic. On the other hand, replacing SRAM cells with SHE-MTJ devices results in a considerable reduction in the transistor count of the LUT circuit since each SRAM cell includes 6 MOS transistors in its structure, while SHE-MTJ devices can be fabricated on top of the MOS circuitry incurring very low area overhead. In particular, SHE-MTJ based LUT circuit achieves at least 51% reduction in MOS transistor count compared to the conventional SRAM-based LUT, as listed in Table 6.3. Transistors with minimum feature size are utilized in the SHE-MTJ based LUT circuit to control the SHE-MTJ write and read operations. Thus, the device count results can provide a fair comparison between SHE-MTJ based LUTs and conventional SRAM-based LUTs in terms of area consumption, since all of the MOS transistors used in both designs have the minimum feature size possible by the 45nm CMOS technology.

Figure 6.6 provides a comparison between the conventional SRAM-based FPGA and the proposed SNRA with a focus on the power dissipation induced by LUT-FF pairs utilized to implement the developed DBN control circuitry. The combined improvements in the read and standby modes of the proposed SNRA resulted in realizing at least 80% reduction in power consumption compared to the conventional CMOS-based reconfigurable fabrics for various DBN topologies. The results obtained for the read operation are comparable to that of the STT-MTJ based FPGA proposed by the Suzuki et al. [189]. However, the utilization of SHE-MTJ based LUTs within the SNRA architecture instead of STT-MTJs can result in at least 20% reduction in configuration energy as demonstrated by in Chapters 3 and 4.

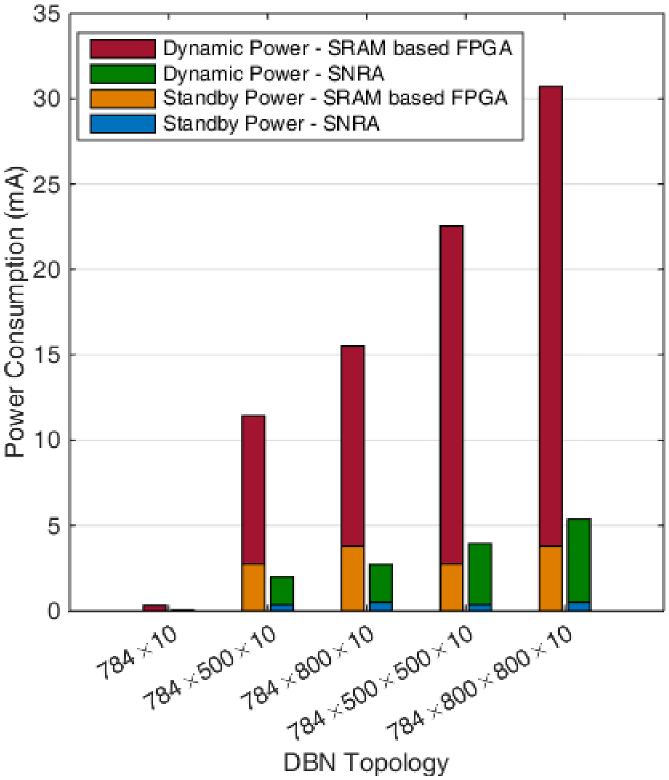


Figure 6.6: Power dissipation of developed FSM for various DBN topologies [6].

CHAPTER 7: COMPOSABLE PROBABILISTIC INFERENCE NETWORKS USING MRAM-BASED STOCHASTIC NEURONS

In Chapter 5, a current-driven low energy-barrier spintronic device has been proposed to be utilized in RBMs as the activation function [8], while similar devices have been previously proposed for spiking [190, 191] and hard axis clocked [183] neural systems. However, the current-mode operation of these devices imposes a significant power consumption to the activation functions, while requiring weighted connections with $M\Omega$ resistances. The design proposed in this chapter takes a new approach from the device-level upward to overcome the challenges mentioned above by utilizing a voltage-driven spintronic device with embedded magnetoresistive random access memory (MRAM) constructed by low energy barrier nanomagnets, which leverages intrinsic thermal noise to provide a natural and power-efficient building block for RBMs. Moreover, we propose a simulation framework for probabilistic learning networks, called PIN-Sim, which is utilized herein to realize a feasible circuit-level implementation of DBN architectures using a SPICE model of our proposed embedded MRAM-based neuron. the main contributions of this chapter are as follows:

1. A Probabilistic Inference Network Simulator (PIN-Sim) to realize a circuit-level implementation of DBN using voltage-controlled embedded MRAM-based neurons as the probabilistic sigmoidal activation functions. The PIN-Sim framework can be used for design space exploration to achieve an optimized network implementation based on the application requirements.

2. Detailed results and analyses of the effects of various circuit- and device-level tunable parameters on the accuracy and power consumption of the DBNs implemented by PIN-Sim framework.

3. Discussions regarding the effects of noise, and variations in the resistance of the weighted connections on the accuracy of our proposed probabilistic spin logic-based DBN circuits.

7.1 Embedded MRAM Based Neuron as a Building Block for RBMs

The basic building block of Boltzmann Machines is a stochastic binary neuron that produces a binary output with a given probability. This probability is modulated by the weighted input the neuron receives from the other neurons [192], as shown Figure 7.1 (a). Here, we show that a recently proposed building block that leverages the highly scaled embedded magnetoresistive random access memory (MRAM) technology, which is conventionally used as a memory device, can enable an approximate hardware representation of the binary stochastic neuron in RBM structure as shown in Figure 7.1 (b).

The functional component of an MRAM architecture is a magnetic tunnel junction (MTJ) that is a multilayer 2-terminal device that exhibits a resistance change depending on the orientation of its magnetic layers. One of these magnetic layers is designed to have a fixed magnetic orientation (fixed layer) while the magnetization of the other layer can be switched by a magnetic field or by a spin-polarized current (free layer). In the latter, a current that flows through the fixed layer can exert a “spin-transfer-torque” to switch the magnetization of the free layer allowing an electrical writing mechanism [193]. In conventional memory devices, the free layer is designed to have a large energy barrier with respect to the thermal energy (kT) so that the fixed layer can function as a non-volatile memory. In recent years the use of superparamagnetic MTJs that are not thermally stable have been experimentally and theoretically investigated in search of functional spintronic devices [194, 195, 196, 197, 198, 199, 200, 201, 8].

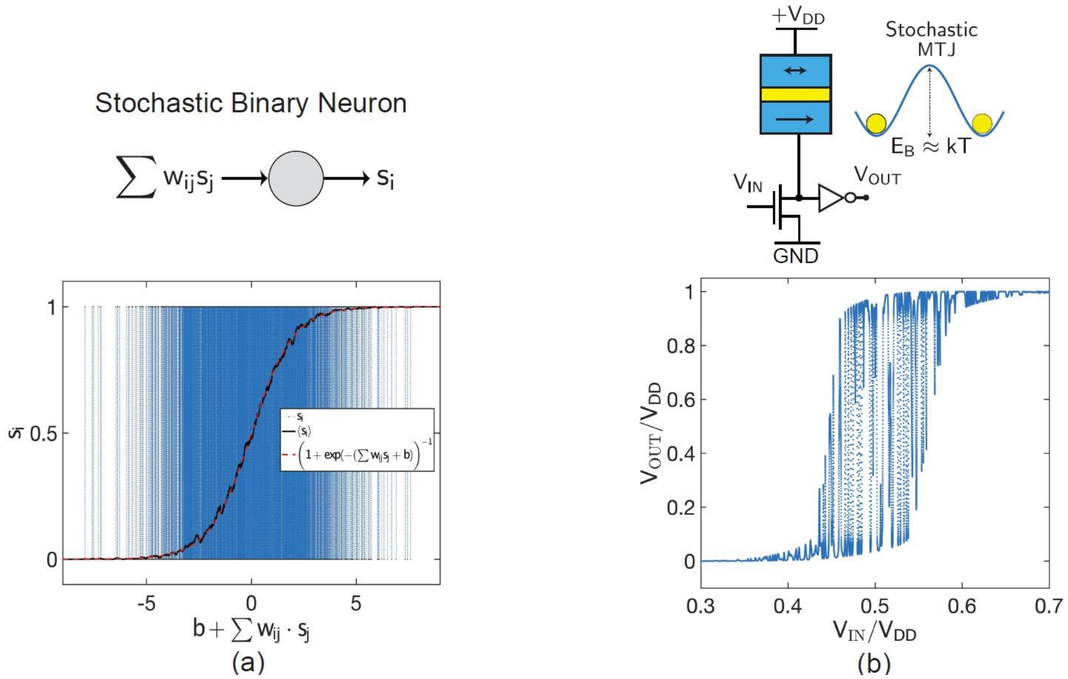


Figure 7.1: a) The building block of the proposed spin-based RBMs, the stochastic binary neuron and its ideal input output characteristics are shown. The dashed red curve indicates the mean of the output that is given by the sigmoid function, $\sigma(z) = 1/(1 + \exp(-z))$, where z is the input. The dashed blue curve is the instantaneous output while the input is being swept. The running average of the output, as indicated by the black curve, shows a mean that is equal to the sigmoid function. b) A hardware representation of the stochastic binary neuron in terms of an Embedded Magnetic Tunnel Junction architecture is shown. The free layer of a conventional Embedded MTJ has an energy barrier E_B of 40-60 kT and thus is non-volatile. Reducing the energy barrier of the free layer results in a resistive behavior that is fluctuating between a low (R_P parallel orientation) and a high (R_{AP} anti-parallel) resistance. The gate voltage of the transistor (V_{IN}) controls the resistance of the transistor to regulate the output voltage to approximate the behavior of a stochastic binary neuron in hardware [9].

Herein, we use a recently proposed design that makes minimal modifications to the 1 Transistor / 1 MTJ architecture of the commercially available embedded MRAM technology [17]. The first modification is to replace the stable free layer with a low-barrier nanomagnet ($E_B \ll 40kT$) that can be achieved by either reducing the total number of spins in the nanomagnet (by reducing M_sVol., where M_s is the saturation magnetization and Vol. is the volume [202]) or by using circular disk magnets that have no preferential easy-axis [198]. The resistance of an MTJ with such a low-barrier nanomagnet randomly fluctuates between high (R_{AP}) and low resistance states (R_P), creating a fluctuating output voltage at the drain of the NMOS transistor (Figure 7.1b). If the transistor resistance that is controlled by the input voltage ($\mathsf{V_{IN}}$) is matched to that of the average MTJ resistance at $\mathsf{V_{IN}} = \mathsf{V_{DD}}/2$, large voltage fluctuations are obtained at the drain output. For typical $\mathsf{R_{AP}}/\mathsf{R_P}$ ratios, a CMOS inverter can amplify these fluctuations to produce a rail-to-rail stochastic output at this input value. Changing the input voltage modulates the transistor resistance, and can suppress these fluctuating outputs either by making the transistor resistance too small and shorting the output to ground, or by making the transistor resistance too high and making the output node $\mathsf{V_{DD}}$. The basic device operation can be understood by considering the MTJ conductance [17]:

$$G_{MTJ} = G_0 \left[1 + m_z \frac{TMR}{(2 + TMR)}\right] \tag{7.1}$$

where m_z is the instantaneous free layer magnetization that is fluctuating stochastically in the presence of thermal noise, G_0 is the average MTJ conductance, $(G_P + G_{AP})/2$, and TMR is the tunneling magnetoresistance ratio, that is defined as $TMR = (G_P - G_{AP})/G_{AP}$. The voltage division between the transistor and the MTJ (Figure 7.1b) produces a drain voltage that can be expressed as:

$$V_{DRAIN}/V_{DD} = \frac{(2 + TMR) + TMR\, m_z}{(2 + TMR)(1 + \alpha) + TMR\, m_z} \tag{7.2}$$

where we introduce a parameter, α, that is defined as the ratio of the transistor conductance (G_T) to the average MTJ conductance (G_0), i. e, $\alpha = G_T/G_0$. As the input voltage $\mathsf{V_{IN}}$ changes the transistor conductance G_T, the drain output behaves as a noisy inverter. It can be seen from Equation 7.2 that the noise amplitude at the drain is maximum when $\alpha \approx 1$, therefore the MTJ resistance is matched to the NMOS resistance ($\alpha = 1$) when $\mathsf{V_{IN}}/\mathsf{V_{DD}} = 0.5$ to obtain an output with large fluctuations at the symmetry point. Even though the drain voltage shows fluctuations of the order of hundreds of mV for typical TMR values, an additional inverter is used to amplify the noise to produce rail-to-rail voltages for a range of input voltages.

The full circuit behavior of the embedded MRAM based neuron is modeled by a solving the magnetization dynamics of the low barrier nanomagnet using the stochastic Landau-Lifshitz-Gilbert (LLG) equation self-consistently with the transport equations in a SPICE framework [12]. The NMOS transistor is modeled by the predictive technology models (PTM) and for simplicity a bias-independent MTJ model is used that is modeled according to Equation 7.1. The magnetization input for the MTJ conductance is instantaneously provided from the stochastic LLG equation. The stochastic LLG reads:

$$(1+\alpha^2)d\hat{m}/dt = -|\gamma|\hat{m} \times \vec{H} - \alpha|\gamma|(\hat{m}\times\hat{m}\times\vec{H}) + 1/qN(\hat{m}\times\vec{I_S}\times\hat{m}) + \Big(\alpha/qN(\hat{m} \times \vec{I_S})\Big) \quad (7.3)$$

where α is the damping coefficient of the nanomagnet, γ is the electron gyromagnetic ratio, q is the electron charge, and $\vec{I_S}$ is the spin current incident to the free layer. The spin current is polarized along the direction of the fixed layer polarization ($\hat{z}$) and its amplitude is proportional to the charge current I_c flowing through the MTJ, such that $\vec{I_S} = PI_c\hat{z}$. N is the total number of spins in the free layer (CoFeB), $N = M_s\text{Vol.}/\mu_B$, where M_s is the saturation magnetization of CoFeB and μ_B is the Bohr magneton. For the free layer, we use a monodomain circular disk magnet whose effective field $\vec{H}$ is given as $-4\pi M_s m_x\hat{x} + \vec{H}_n$, $\hat{x}$ being the out-of-plane direction of the magnet. $\vec{H}_n$ is the isotropic thermal noise field, uncorrelated in three directions: $(H_n^{x,y,z})^2 = 2\alpha kT/(|\gamma|M_s\text{Vol.})$. The transistors are based on 14nm HP-FinFET PTM.

Table 7.1: Parameters Used for Modeling and Simulation [17]

Parameters	**Value**
Saturation magnetization (CoFeB) (M_s)	$1100emu/cc$ [203]
Free Layer diameter, thickness	$22nm$, $2nm$
Polarization	0.59 [204]
TMR	110% [204]
MTJ RA-product	$9\Omega - \mu m^2$ [204]
Damping coefficient	0.01 [203]
Temperature	$26.85°C$

In this chapter, we use a circular disk magnet with $\ll kT$ energy barrier in the absence of any shape anisotropy. Such magnets have been fabricated and characterized in [205, 206]. Moreover, elliptical magnets showing GHz telegraphic oscillations have also been experimentally observed in [207]. The demonstrated parameters listed in Table 7.1 [17] are used to generate all of the results that are provided within this chapter. We also note for the chosen parameters with a circular free layer with an in-plane anisotropy that the results are not significantly influenced by the current that is flowing at the midpoint ($\mathsf{V}_{\mathsf{IN}} = \mathsf{V}_{\mathsf{DD}}/2$), and note that any pinning at higher input voltages benefits the switching operation of the device.

7.1.1 RBM Hardware Implementation

Figure 7.2 exhibits a feasible hardware implementation of an $n \times m$ RBM, in which neurons based on the concise embedded MRAM-based design described in the previous section are used to generate the required probabilistic sigmoidal activation function. The resistive crossbar arrays are utilized to realize the matrix multiplication. In this work, the weights are trained off-chip and the resistive weighted connections will be programmed accordingly. Any resistive devices such as memristors [208] or spin-orbit torque (SOT)-driven domain wall motion (DWM) devices [180] can be utilized for weighted connections without the loss of generality.

7.2 Proposed DBN structure

To implement the positive and negative weights in the ***w*** matrix, two resistive weighted arrays with the same dimensions are required [209], as shown in Figure 7.2. The outputs of the positive and negative weighted connections are linked to differential amplifiers which are implemented by op-amps as shown in Figure 7.2. The output voltage of the op-amp, i.e. $V_{out} = \frac{R_1}{R_0}(V_{in}^{+} - V_{in}^{-})$, is applied to the MRAM-based neuron as an input signal. The neuron with embedded MRAM will generate an output voltage signal, which fluctuates between VDD and GND with a probability that is modulated based on the applied input voltage. Finally, a resistor-capacitor (RC) integrator circuit is utilized to convert the probabilistic output of the neuron to an analog voltage level, which can be later converted to a digital output through digital to analog conversion. In order to verify the functionality and assess the performance of our proposed RBM implementation, we have simulated a 2×2 RBM via SPICE circuit simulation using the 14nm HP-FinFET technology library with an MRAM-based neuron used as the activation function. The results obtained validate the functionality of our proposed design as elaborated in Figure 7.3.

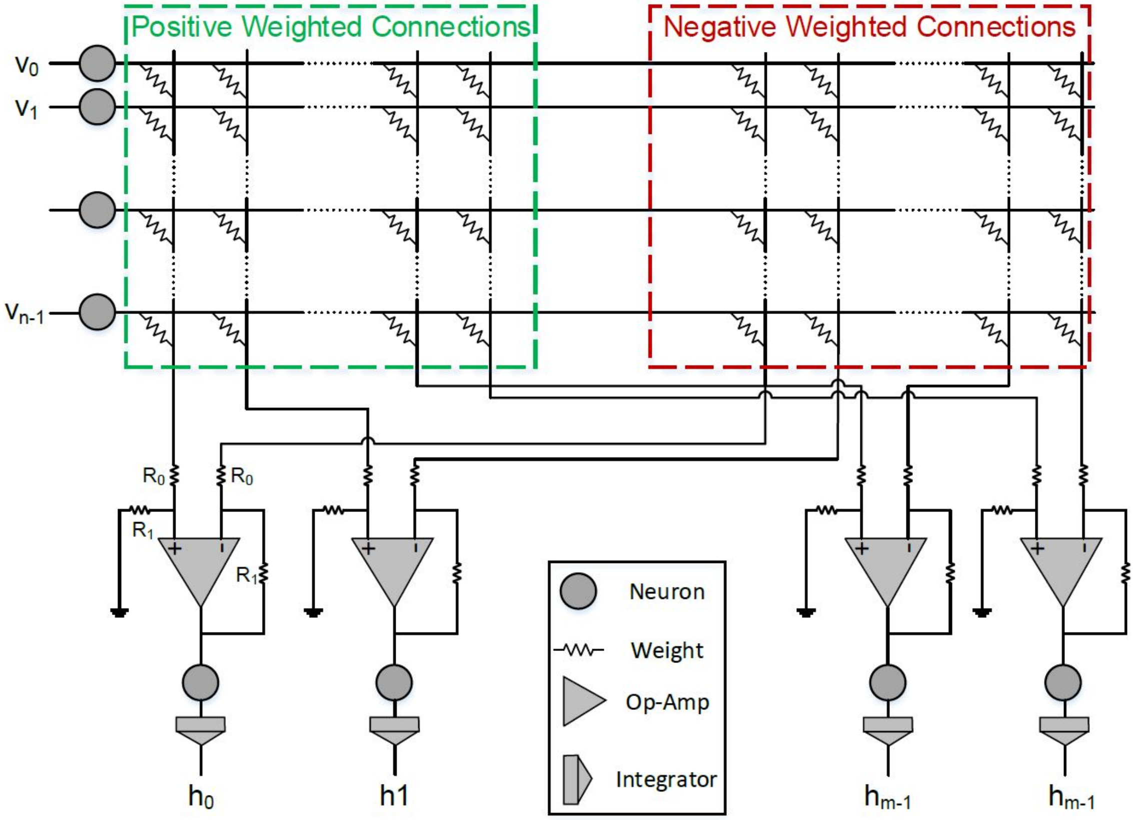


Figure 7.2: An $n \times m$ RBM hardware implementation. Two resistive arrays are leveraged along with differential amplifiers to implement both positive and negative weights. The embedded MRAM-based neurons are used to evaluate the activation functions. The fluctuating output voltage of the neurons are integrated through an RC circuit to generate the output of the proposed RBM structure [9].

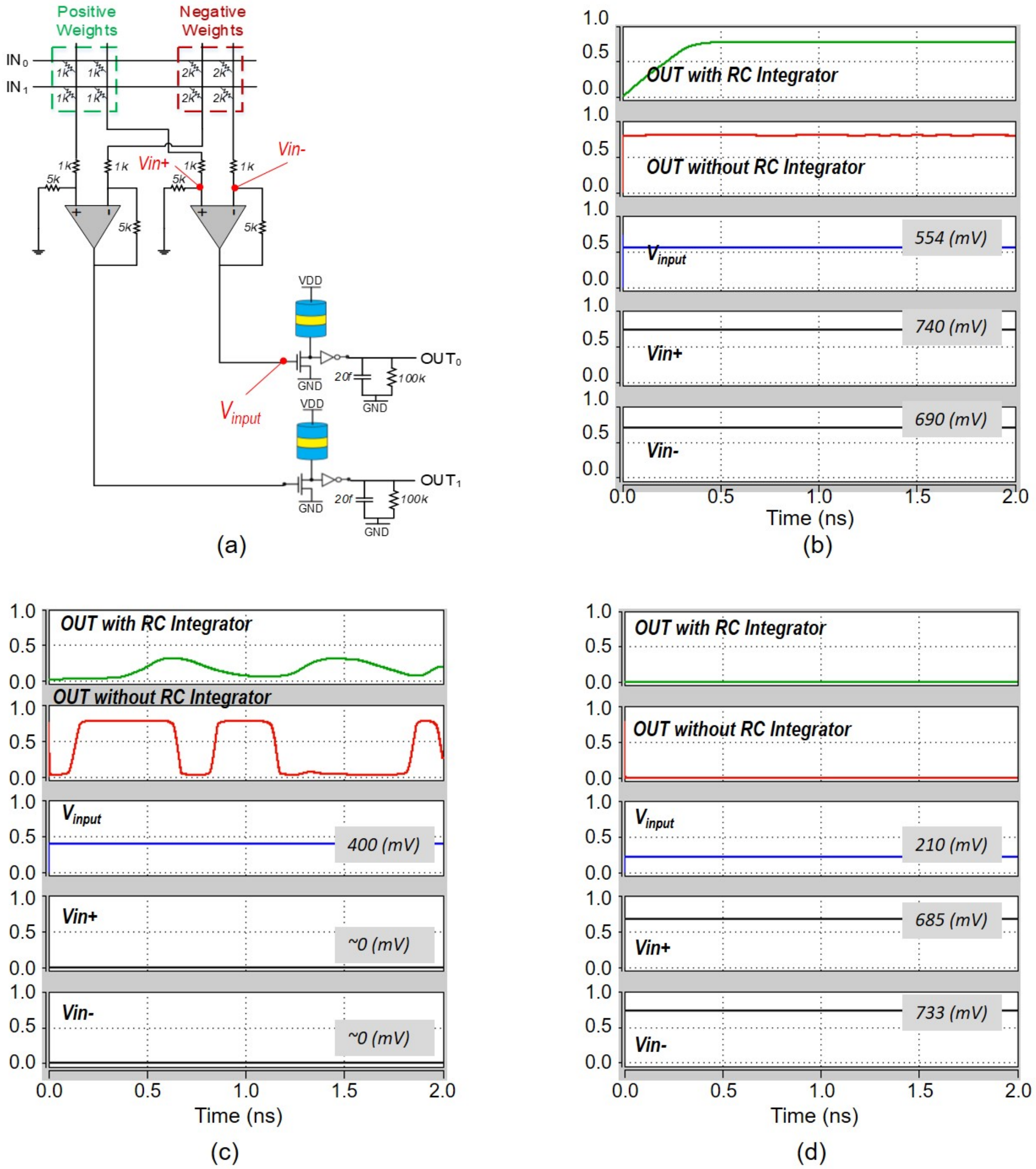


Figure 7.3: (a) a 2 × 2 RBM implementation using the embedded MRAM based neuron. The DC bias voltage of $V_{DD}/2 = 400mV$ is added to the output of the differential amplifier to set our proposed neuron at its midpoint. (b) The behavior of the implemented RBM for $IN_0 = V_{DD}$ and $IN_1 = V_{DD}$ while the positive and negative weight resistances are $1k\Omega$ and $2k\Omega$, respectively. The input voltage connected to the positive terminal of the differential amplifier is larger than the negative terminal resulting in an output voltage larger than VDD/2. The output of the differential amplifier is connected to the input of the neuron, thus the $V_{IN}/V_{DD} =\sim 0.7$ for the neuron leading to output logic “1”, as shown in Figure 7.1 (b). (c) The behavior of the RBM for $IN_0 = 0$ and $IN_1 = 0$. The inputs of the differential amplifiers are near zero, thus $V_{IN}/V_{DD} =\sim 0.5$ and the state of the neuron fluctuates between “0” and “1”. (d) The RBM behavior for $IN_0 = V_{DD}$ and $IN_1 = V_{DD}$ while the positive and negative weight resistances are $2k\Omega$ and $1k\Omega$, respectively. The $V_{IN}/V_{DD} =\sim 0.3$ resulting in the neuron being in state “0” according to Figure 7.1 (b) [9].

7.2.1 Probabilistic Inference Network Simulator (PIN-Sim)

In order to automate and scale up the design space exploration of DBNs at the circuit-level, we have developed a hierarchical simulation framework called PIN-Sim, which can be utilized to implement any probabilistic learning networks. The block diagram of the PIN-Sim framework used to implement DBNs in our work is shown in Figure 7.4, which is comprised of five primary blocks. The PIN-Sim methodology is described in Algorithm 1. First, we have modified a MATLAB implementation of DBN developed in [186] to train the network and obtain the trained weight (***W***) and bias (***B***) matrices. The extracted (***W***) and (***B***) matrices are then applied to a MATLAB module called ***mapWEIGHT***, the functionality of which is described in Algorithm 2. The *mapWEIGHT* module first converts each of the ***W*** and ***B*** matrices with positive and negative elements to two separate matrices with only positive elements as described below:

$$w^{+}_{(i,j)} = \begin{cases} w_{(i,j)}, & \text{if } w_{(i,j)} \geq 0 \\ 0, & \text{if } w_{(i,j)} < 0 \end{cases}, \qquad w^{-}_{(i,j)} = \begin{cases} 0, & \text{if } w_{(i,j)} \geq 0 \\ -w_{(i,j)}, & \text{if } w_{(i,j)} < 0 \end{cases} \tag{7.4}$$

$$b^{+}_{j} = \begin{cases} b_{j}, & \text{if } b_{j} \geq 0 \\ 0, & \text{if } b_{j} < 0 \end{cases}, \qquad b^{-}_{j} = \begin{cases} 0, & \text{if } b_{j} \geq 0 \\ -b_{j}, & \text{if } w_{j} < 0 \end{cases} \tag{7.5}$$

Next, the *mapWEIGHT* module maps the elements in W^{+}, W^{-}, B^{+}, and B^{-} matrices to their corresponding conductance values using the below equations:

$$\forall w_{(i,j)} \in (W^{+}, W^{-}) : gw_{(i,j)} = \frac{(g_{max} - g_{min}) \times (w_{(i,j)} - w_{min})}{w_{max} - w_{min}} + g_{min} \tag{7.6}$$

$$\forall b_{(i,j)} \in (B^{+}, B^{-}) : gb_{(i,j)} = \frac{(g_{max} - g_{min}) \times (b_{(i,j)} - b_{min})}{b_{max} - b_{min}} + g_{min} \tag{7.7}$$

where $\forall g_{(i,j)} \in \mathbf{G} : g_{min} \leq g_{(i,j)} \leq g_{max}$, in which $g_{min} = 1/r_{max}$ and $g_{max} = 1/r_{min}$ are minimum and maximum conductances of all weighted connections in the crossbar weighted array. Moreover, b_{max}, b_{min}, w_{max}, and w_{min} are the maximum and minimum values in all of the bias and weight matrices, respectively. Finally, Equation 7.8 is utilized to convert and quantize all of the obtained conductance values to their corresponding resistance values, which can then be utilized to implement the required resistive crossbar array.

$$\forall g_{(i,j)} \in (GW^{+}, GW^{-}, GB^{+}, GB^{-}) : r_{(i,j)} = \frac{round(Q \times 1/g_{(i,j)})}{Q} \tag{7.8}$$

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

Once the positive and negative weight and bias resistance matrices are obtained, they will be converted to *text* files and applied to a Python module called ***mapRBM.py***, shown in Figure 7.4, which produces plural crossbar weighted array circuits in SPICE automatically based on the defined network topology. Finally, a ***testDBN.py*** module is developed using Python scripts, which utilize the generated circuit of the DBN, and the model of the probabilistic neuron to perform a SPICE circuit simulation and calculate the error rate using the *test inputs* and *test labels*, which are provided for the *testDBN* module in form of *text* files.

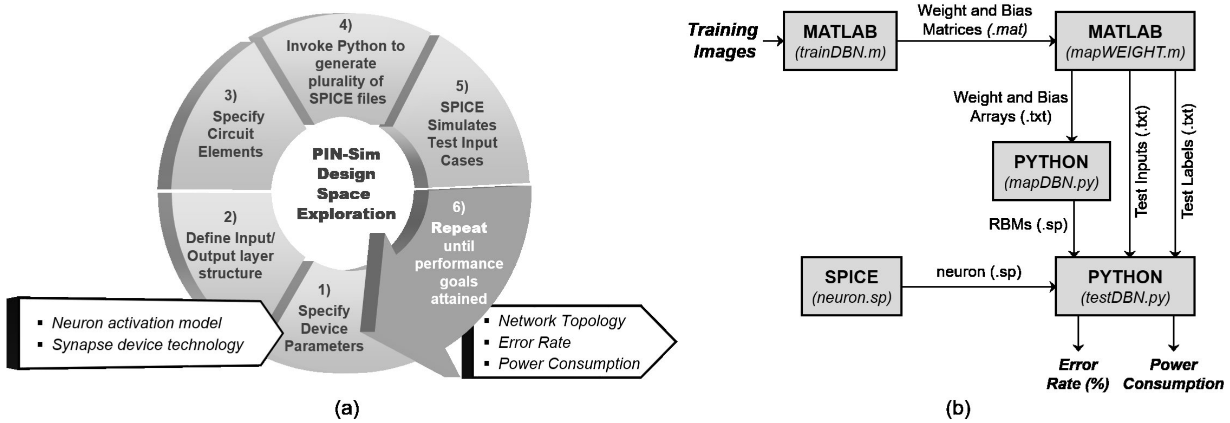

4) Invoke Python to generate plurality of SPICE files
3) Specify Circuit Elements
5) SPICE Simulates Test Input Cases
PIN-Sim Design Space Exploration
2) Define Input/ Output layer structure
6) Repeat until performance goals attained
1) Specify Device Parameters
▪ Neuron activation model
▪ Synapse device technology
▪ Network Topology
▪ Error Rate
▪ Power Consumption
(a)
Training Images
MATLAB (trainDBN.m)
Weight and Bias Matrices (.mat)
MATLAB (mapWEIGHT.m)
Weight and Bias Arrays (.txt)
Test Inputs (.txt)
Test Labels (.txt)
PYTHON (mapDBN.py)
RBMs (.sp)
SPICE (neuron.sp)
neuron (.sp)
PYTHON (testDBN.py)
Error Rate (%)
Power Consumption
(b)

Figure 7.4: (a) The PIN-Sim framework can be utilized to explore the design space to realize the optimized network implementation based on the application requirements. (b) The block diagram of the PIN-Sim framework, which consists of five main modules: *(1) trainDBN:* a MATLAB-based module used for training the DBN architecture. *(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 circuit-level implementation of the RBMs using the obtained weight and bias resistances. *(4) neuron:* A SPICE model of the MRAM-based stochastic neuron. *(5) testDBN:* the main module developed in Python that executes test evaluations to assess the error rate and power consumption using the outputs of the other modules in PIN-Sim [9].

7.3 Simulation Results and Discussion

Herein, we have leveraged a hierarchical simulation method to examine the performance of our DBN implementation. In software-level simulation, the behavioral results of the developed embedded MRAM-based neuron model are used to implement a DBN in MATLAB for MNIST pattern recognition application [187]. In the hardware-level simulation, the proposed framework is used to develop a circuit-level DBN implementation using the p-bit SPICE model and 14nm HP-FinFET PTM technology in SPICE circuit simulator with 0.8V nominal voltage.

Algorithm 1: PIN-Sim Methodology [9]

```
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: $Err = 0$
$weight.mat, bias.mat \Leftarrow$ Contrastive_Divergence Algorithm
$posWeight.txt, negWeight.txt, posBias.txt, negBias.txt \Leftarrow$
  mapWeight($(Weight.mat, Bias.mat)$)
for i= 1 : S do
    $input\_data = D_{test}(i)$ ;
    for j= 1 : M do
        $RBM(j).sp \Leftarrow$
          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 $\neq$ Label(i)) then
                $Err+ = 1$ ;
            end
        else
            $input\_data = OUTPUT$ ;
        end
    end
end
$ErrorRate = Err/S$ ;
```

Algorithm 2: mapWeight Methodology [9]

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**
    $W^{+}, W^{-} \Leftarrow weight(i)$ Matrix ;
    $B^{+}, B^{-} \Leftarrow bias(i)$ 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 $B_{pos}, B_{neg}$ ;
    $GW^{+} = \frac{(g_{max}-g_{min})\times(W^{+}-w_{min})}{w_{max}-w_{min}} + g_{min}$ , $RW^{+} = \frac{round(Q\times 1/GW^{+})}{Q}$;
    $GW^{-} = \frac{(g_{max}-g_{min})\times(W^{-}-w_{min})}{w_{max}-w_{min}} + g_{min}$ , $RW^{-} = \frac{round(Q\times 1/GW^{-})}{Q}$;
    $GB^{+} = \frac{(g_{max}-g_{min})\times(B^{+}-b_{min})}{b_{max}-b_{min}} + g_{min}$ , $RB^{+} = \frac{round(Q\times 1/GB^{+})}{Q}$;
    $GB^{-} = \frac{(g_{max}-g_{min})\times(B^{-}-b_{min})}{b_{max}-b_{min}} + g_{min}$ , $RB^{-} = \frac{round(Q\times 1/GB^{-})}{Q}$;
    $posWeight(i).txt \Leftarrow RW^{+}$ ;
    $negWeight(i).txt \Leftarrow RW^{-}$ ;
    $posBias(i).txt \Leftarrow RB^{+}$ ;
    $negBias(i).txt \Leftarrow RB^{-}$ ;
**end**
```

7.3.1 MATLAB simulation

Herein, we have modified the sigmoid activation function in a MATLAB implementation of DBN [186] by using the device-level simulation results of the proposed embedded MRAM-based neuron. To assess the performance of the implemented DBN, we have used the MNIST data set [187] including 60,000 training and 10,000 test sample images of hand-written digits, each of which having 28×28 pixels. We have used Error rate (ERR) and root-mean-square error (RMSE) metrics to evaluate the performance of the DBN, as expressed by the following equations [186]:

$$ERR = \frac{N_F}{N} \tag{7.9}$$

$$RMSE = \sqrt{\frac{1}{MN}\sum_{k=1}^{N}(y_k - F(x_k)^2)} \tag{7.10}$$

where M is the number of output classes, N is the number of input data, N_F is the number of false inference, F is the inference of the trained DBN, x_k is the k-th input data and y_k represents its corresponding target output.

As shown in Figure 7.5, the most elementary model of the DBN requires 784 nodes in visible layer for the 28×28 pixels of the input images, and 10 nodes in hidden layer for 0-9 output digits. Figure 7.6 shows the relation between the error rate and the number of training samples for seven distinct DBN topologies, which is obtained using 1,000 test samples. The results obtained by MATLAB simulation exhibit that an error rate of 28.2% for a 784×10 DBN trained by 500 training inputs can be decreased to a 2.5% error rate achieved using $784 \times 500 \times 500 \times 500 \times 10$ and $784 \times 500 \times 500 \times 10$ DBN topologies, which are trained by 10,000 input training samples. Thus, the recognition accuracy can be improved by increasing the number of hidden layers in the network, number of nodes in each layer, and number of training samples. However, these improvement can lead to higher power consumption and area overheads as investigated in the hardware-level simulations elaborated below.

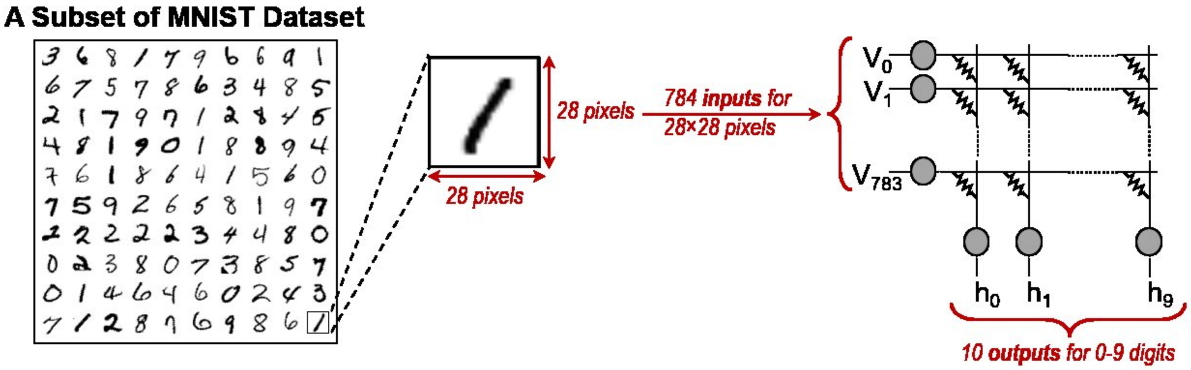
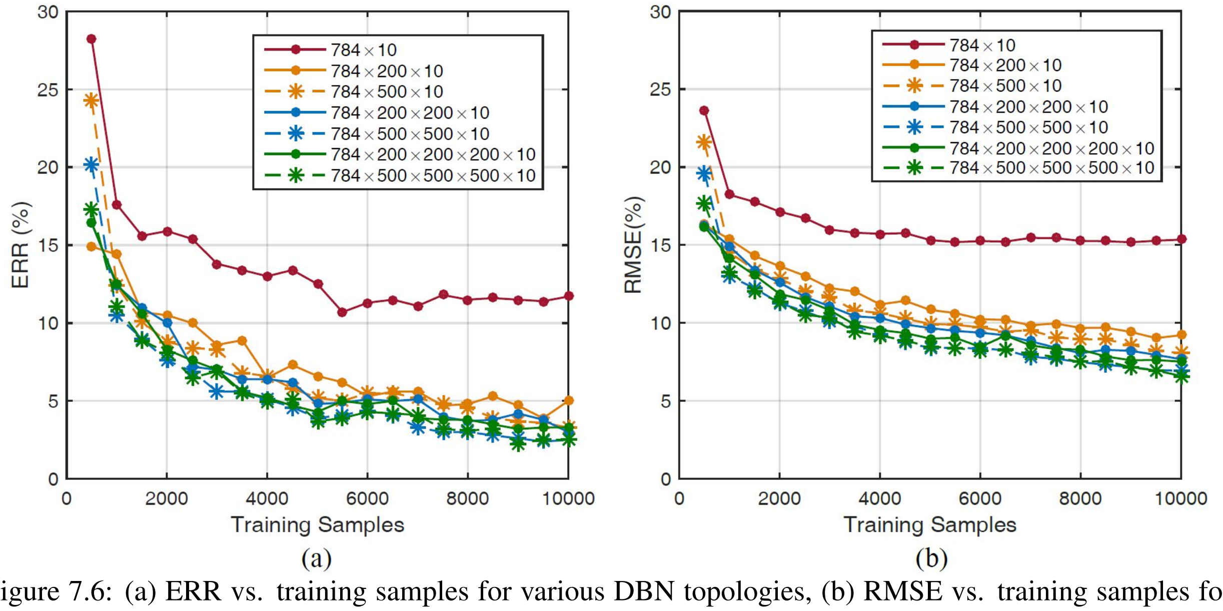

A Subset of MNIST Dataset
28 pixels
28 pixels
784 inputs for 28×28 pixels
V_0
V_1
V_{783}
h_0
h_1
h_9
10 outputs for 0-9 digits

Figure 7.5: The most elementary 784 × 10 DBN required for MNIST digit recognition application. The visible layer includes 784 nodes to handle 28 × 28 pixels of the input images, while the 10 nodes in hidden layer represent the output classes [9].



(a) (b)

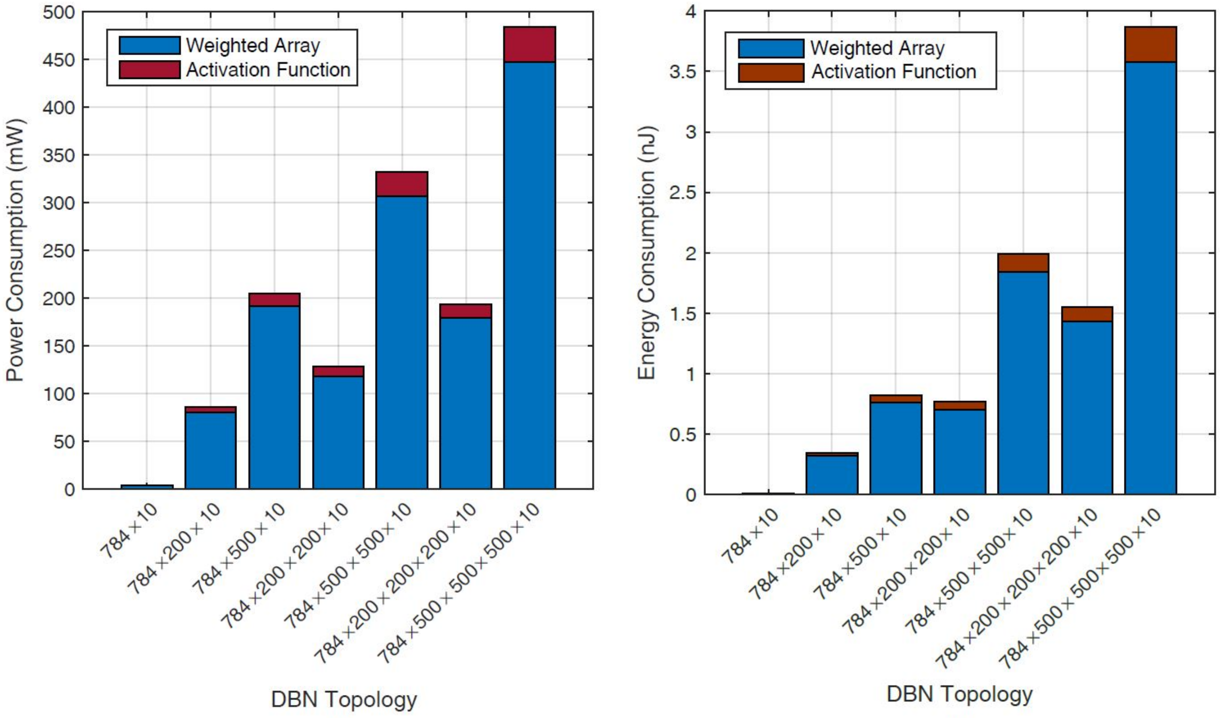
Figure 7.6: (a) ERR vs. training samples for various DBN topologies, (b) RMSE vs. training samples for various DBN topologies [9].

7.3.2 PIN-Sim simulation

In this section, we utilize our proposed PIN-Sim framework to provide a circuit-level model of DBN architecture. Next, we will provide the energy and power consumption profiles of the seven different DBN topologies investigated in the previous section to analyze the energy and accuracy trade-offs of these networks. Finally, we will focus on the effect of various important hardware-level parameters. These are vital parameters during design space exploration that influence the accuracy of DBN architectures as tradeoffs necessary to obtain efficient hardware-level implementation for pattern recognition applications.

7.3.2.1 Power and Energy Consumption Analysis

Figure 7.7(a) depicts the power consumption of various DBN topologies while evaluating a single input image. As shown, a significant amount of power is consumed in the weighted connections, while less than 10% of the total power is consumed in the neurons of an embedded MRAM-based p-bit approach. For instance, the total power consumption of a $784 \times 200 \times 10$ DBN is approximately equal to 86 mW, only 5.6 mW of which is dissipated in the activation functions. This is achieved by using the proposed power-efficient embedded MRAM-based neurons to implement the activation functions, as opposed to more elaborate floating-point circuits and pseudo-random number generators. Moreover, it is shown that the total power consumption depends primarily upon the aggregate number of neurons that are used in a network and not the number of layers. For instance, the power consumption of a $784 \times 500 \times 10$ DBN is greater than that of a $784 \times 200 \times 200 \times 10$ network, although the latter has higher number of hidden layers. However, the test operation delay is linearly proportional to the number of hidden layers which is determined by the signal propagation and computation progression. In particular, the RC integrator circuit shown in Figure 7.2 is sampled every 2 ns, leading to an operating clock frequency of 500 MHz and a delay of 2 ns for each RBM. Thus, the $784 \times 200 \times 200 \times 10$ DBN mentioned above requires three clock cycles to complete the evaluation operation, while a $784 \times 500 \times 10$ DBN can produce its output in two clock cycles. Figure 7.7(b) shows the energy consumption for various DBN topologies, which simultaneously includes the impact of number of nodes and hidden layers on power consumption and delay, respectively.



(a) (b)

Figure 7.7: Test operation: (a) Power Consumption for various DBN topologies, (b) Energy Consumption for various DBN topologies [9].

Table 7.2: PIN-Sim tunable parameters and their default values [9].

Parameters	Description	Default Value
$Topology$	Defines the number of layers and nodes	$784 \times 200 \times 10$
$TrainNum$	# of training images	3,000
R_{min}	Minimum resistance of the weighted connections	1 $k\Omega$
ΔR_W	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 $k\Omega$, 5 $k\Omega$
R_i, C_i	Resistance and capacitance of the RC integrator circuits	100 $k\Omega$, 20 fF

7.3.2.2 PIN-Sim tunable parameters and their affect on DBN performance

Table 7.2 lists the tunable parameters in the PIN-Sim framework, which can be adjusted based on the application requirements. The last column of the table shows the default values that are utilized herein for the MNIST digit recognition application. Figure 7.8 shows the output voltages of the neurons in the last hidden layer of a $784 \times 200 \times 10$ DBN utilized for MNIST pattern recognition tasks, each of which represents an output class. The probabilistic outputs of the p-bit devices are shown in Figure 7.8(a), while Figure 7.8(b) exhibits the outputs of their corresponding integrator circuits. The outputs of the integrators are sampled after 2 ns, which is equal to the time constant of the integrator circuit. The output with the highest voltage amplitude represents the class to which the input image belongs. The results obtained exhibit a correct recognition operation for a sample input digit “4” within the MNIST dataset.

Next, we will focus on the effect of some of the tunable parameters on the accuracy and power consumption of DBN architectures implemented by the proposed PIN-Sim framework. First, the effect of ΔR_W is investigated, which defines the possible resistance range of weights and biases as follows, $r_{max} = (1 + \frac{\Delta R_W}{100}) \times r_{min}$. The r_{max} and r_{min} parameters are utilized in the *map-WEIGHT* module in the PIN-Sim tool to map the trained weights and biases to their corresponding resistance values according to Equations 7.6 and 7.7, respectively. Figure 7.9(a) shows the effect of ΔR_W on the recognition accuracy and power consumption of our default $784 \times 200 \times 10$ DBN implementation. As it can be seen in the figure, the error rate is reduced from 53% to 24% by increasing the ΔR_W from 100% to 400%, however a significant change in the error rate cannot be observed for ΔR_W values larger than 400%. These results are particularly beneficial for magnetic tunnel junction (MTJ)-based weighted connections [180, 210], in which the difference between maximum and minimum resistance is defined by the tunneling magneto-resistance (TMR) effect. The results obtained show that a TMR of 400% could be adequate to achieve the desired error rate.

However, it is worth noting that this is quite application specific and can vary for different datasets. These results are worthy since the realization of higher TMR values would impose more complex fabrication processes [211], of which 700% [212] have been demonstrated experimentally and others of 250% [213] via current scalable means. Moreover, as it is shown in Figure 7.9(a), increasing the ΔR_W results in reduced power dissipation in the weighted array, while the power dissipated in activation functions remains almost unchanged. The higher resistance range for the weighted connections increases the overall resistance of the weighted array. Therefore, since the input voltages remain unchanged the current flowing through the synapses will be decreased, which consequently reduces the power dissipated in the weighted array.

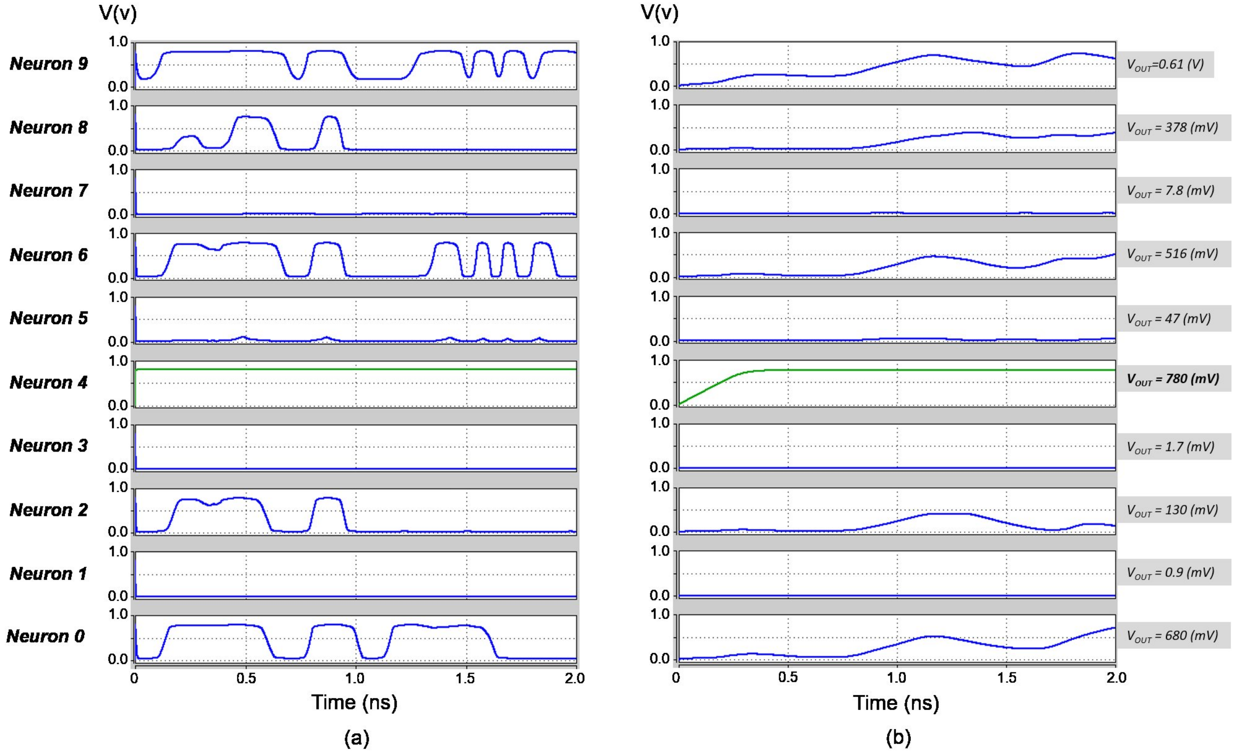
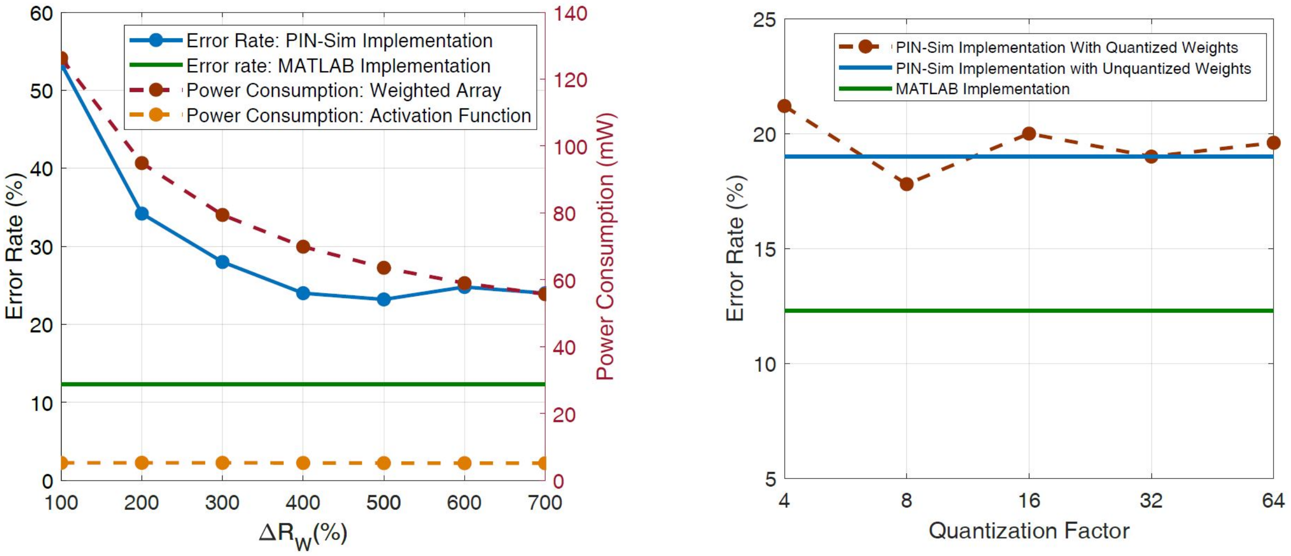


Figure 7.8: Output of a $784 \times 200 \times 10$ DBN for a sample digit of “4” in the MNIST dataset: (a) Probabilistic output of the p-bit devices, (b) Output of the integrator circuit. The output voltage of the *neuron-4*, which represents the digit “4” in the output classes, is greater than the other output voltages verifying a correct evaluation operation [9].

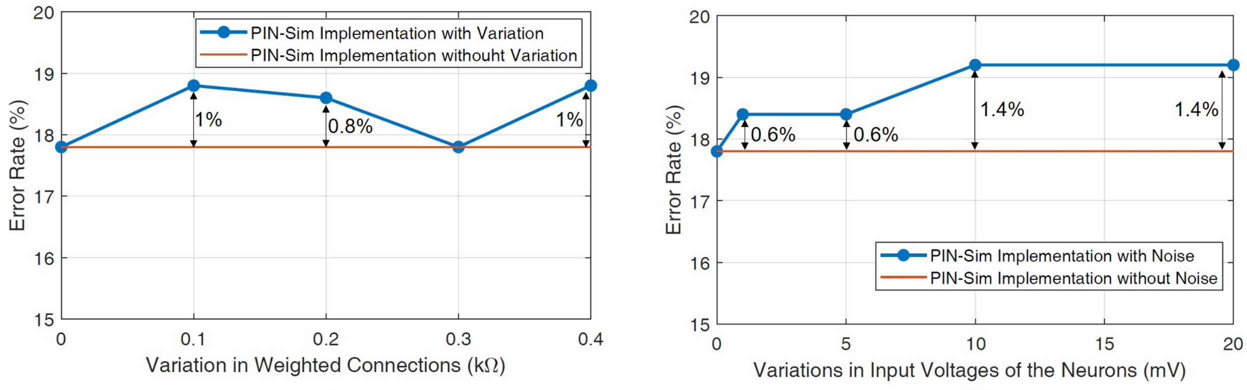


(a)

(b)

Figure 7.9: (a) Error rate and power consumption versus ΔR_W, and (b) error rate versus quantization factor (Q) for a $784 \times 200 \times 10$ DBN trained by 3,000 training images. The software implementation is technology-independent, in which the ideal sigmoid activation function and weight values are utilized in MATLAB to calculate the error rate. Thus, the changes in the tunable parameters used in the circuit-level SPICE implementation do not affect the measured error rates [9].

In practice, providing an accurate and continuous range of weight resistances at nanoscale is not attainable due to the fabrication complexities and process variation. Therefore, a realistic circuit-level model of the resistive crossbar architecture should leverage quantized weights. Thus, leveraging PIN-Sim framework for design space exploration, we have assigned a quantization factor (Q) parameter, which can be tuned by the user based on the application requirements. Figure 7.9(b) shows the effect of weight discretization on the recognition accuracy of a $784 \times 200 \times 10$ DBN with ΔR_W of 400% that is trained with 3,000 training samples. As shown, the error rate for the hardware implementation with $Q = 4$, which means the weights are discretized into four equal intervals between R_{min} and R_{max}, is increased to 21.2% from the 19% error rate that is achieved by the DBN with unquantized weights. As it is expected, this increase in the error rate is mainly caused by the information loss that occurs during the discretization. Moreover, implementations with larger Q values result in error rates closer to that of the DBN with unquantized weights, which can also be expected since the discretization intervals are so small that the weight values are getting close to their unquantized values. However, an interesting phenomenon can be observed in the hardware implementation with $Q = 8$, where the error rate of 17.8% is realized which is lower than the error rate of the unquantized DBN. We have performed multiple tests to ensure that this is a repetitive behavior for the DBNs with $Q = 8$, and in all of the cases the error rate obtained was lower than that of the DBN with unquantized weights. These results can be particularly interesting in the hardware-implementation, since for instance in our examined case there is a 0.5 $k\Omega$ gap between various weight resistances, considering the $R_{min} = 1k\Omega$ and $\Delta R_W = 400\%$, which can provide some robustness against process variations without incurring a significant increase in the error rate. In particular, we have investigated the impacts of the variations in the input voltages of neurons, which can be induced by different noise sources, as well as variations in the resistance of the weighted connections on the recognition accuracy of the network. According to the results shown in Figure 7.10 (a), a $784 \times 200 \times 10$ DBN trained by 3,000 images loses 1% accuracy in presence of variations in weighted connections ranging from 0.1 $k\Omega$ to 0.4 $k\Omega$. Moreover, Figure 7.10 (b) exhibits 1.4% increase in the error rate for variations in the input voltages of neurons with a standard deviation of 20 mV.



(a)

(b)

Figure 7.10: (a) Error rate versus the variation in the resistance of weighted connections, and (b) error rate versus the variations in the input voltages of the neurons for a $784 \times 200 \times 10$ DBN trained by 3,000 training images [9].

7.3.3 Discussion

Some of the previous hardware implementations of DBNs are listed in Table 6.3. The designs proposed in [96, 97] leverage FPGAs to achieve speedups of 25-145 compared to software implementations, however these approaches suffer from constrained clock frequencies and routing congestion, as well as major resource deficiencies due to the significant embedded memory utilization for both weighted connections and activation functions. In [188], those authors have proposed optimization methods to reduce memory requirements for weights and biases, however implementing each activation function still requires dedicated piecewise linear approximator (PLA), random number generator (RNG), and comparator circuits which lead to increased area and energy consumption per neuron than the embedded MRAM-based approach herein. In [98], the low-complexity characteristics of stochastic CMOS-based arithmetic units are leveraged to implement RBM with reduced area and power consumption. However, the large number of linear feedback shift registers (LFSRs) that are required to generate the long input and weight bit-streams results in increased latencies that considerably limits the energy savings.

On the other hand, emerging technologies such as resistive RAM (RRAM) and phase change memory (PCM) have been recently utilized within the crossbar arrays to implement matrix multiplication within RBMs [100, 99, 101]. In particular, [100] has achieved $100\times$ and $10\times$ improvement in terms of operation speed and energy consumption, respectively, compared to single-threaded cores by using RRAM devices as weighted connections. In all of the above-mentioned designs, CMOS-based circuits such as multipliers and RNGs are utilized to realize the probabilistic behavior of activation functions. In [8], authors have utilized low energy barrier spin-orbit torque (SOT) MTJs to implement the probabilistic sigmoidal activation function, which realizes significant area and energy reductions. However, the current-mode behavior of the SOT-MTJ devices imposes significant power consumption to the activation functions, while requiring weighted con-

Table 7.3: Various DBN hardware implementations with a focus on activation function structure [8, 9].

Design	Weighted Connection	Activation Function	Energy per Neuron	Normalized area per neuron
[96]	Multipliers	CMOS-based LUTs	N/A	N/A
[97]	Multipliers	- 2-kB BRAM - PLA - RNG	$\sim$10-100 nJ	$\sim 3000\times$
[188]	- Multiplier - Adder tree	- PLA - RNG - Comparator	$\sim$10-100 nJ	$\sim 2000\times$
[98]	- LFSR - bit-stream - AND/OR gates	-LFSR - Bit-wise AND - tree adder - FSM-based *tanh* unit	$\sim$10-100 nJ	$\sim 90\times$
[99]	RRAM Memristor	Off-chip	N/A	N/A
[100]	RRAM	- 64×16 LUTs - Pseudo Random Number Generator - Comparator	$\sim$1-10 nJ	$\sim 1250\times$
[101]	PCM	Off-chip	N/A	N/A
[8]	SOT-DWM $M\Omega$ resistances	Low-energy barrier SOT-MTJ	$\sim$1-10 fJ	$\sim 1.25\times$
Proposed Herein	Memristive Devices	MRAM-based Stocahstic Neuron	Neuron: $\sim$1-10 fJ Integrator: $\sim$10-20 fJ	Neuron: $1\times$ Integrator: $\sim 3\times$

nections in $M\Omega$ resistances which can incur significant area overhead and fabrication complexity [180, 185]. The work presented herein utilizes a voltage-driven embedded MRAM-based neuron with low energy barrier unstable nanomagnets, which leverages the intrinsic thermal noise to generate sigmoidal probabilistic activation functions required for RBMs within a power-efficient package. As listed in Table 6.3, the proposed RBM implementation using embedded MRAM-based neuron can achieve approximately three orders of magnitude energy reduction compared to the previous energy-efficient CMOS-based implementations, while realizing at least $90\times$ device count reduction. However, as it was described in previous sections, the embedded MRAM based neuron requires an RC circuit to integrate its output voltage. The SPICE circuit simulation results exhibits an approximate average energy consumption of 10-20 fJ for the RC circuit as listed in Table 6.3. Moreover, the area required to implement the RC circuit with 100 $K\Omega$ resistor and $20fF$ capacitor is approximately three times larger than that of the MRAM-based neuron

[214, 215]. Thus, the proposed MRAM-based activation function can achieve approximately $20\times$ and $300\times$ area reduction compared to the CMOS-based stochastic neurons proposed in [98] and [100], respectively. The area of the MRAM-based neuron, which is utilized as the baseline for the area comparisons, is approximately equal to $32\lambda \times 32\lambda$, that is obtained by the layout design, in which λ is a technology-dependent parameter. Herein, we have used the 14nm FinFET technology, which leads to the approximate area consumption of $0.05\mu m^2$ per neuron. MRAM devices can be fabricated on top of the transistors, thus incurring near-zero area overhead.

CHAPTER 8: CONCLUSION

8.1 Summary

The proposed HSC-FPGA offers an intriguing feasible architecture for the next generation of configurable fabrics, which allows embracing the advantages of both CMOS and beyond-CMOS technologies without requiring significant modification to the routing structure, programming paradigms, and synthesis tool-chain of the commercial FPGAs. In the HSC-FPGA's fabric structure, hybrid spin/CMOS CLBs are used to implement both sequential and combinational logic circuits. Within its CLBs, the intrinsic characteristics of the MTJ and its corresponding sensing circuit make MRAM-LUT a suitable alternative for CMOS-based LUT-FF pairs to implement sequential logic, while combinational logic circuits can be implemented by SRAM-LUTs.

SPICE simulation results indicate at least 40% and 83% reduction can be realized in average read power and standby power, respectively, for MRAM-LUT compared to SRAM-based LUT-FF pairs. However, these advantages are achieved at the cost of significantly larger write energy, which fortunately occurs rarely, as well as more than 20% increase in the LUT circuit area that is mainly caused due to the large size of the CMOS transistors in the MTJ's write circuit. Thus, device-level optimizations were proposed, according to which STT-MTJ based devices were replaced by SHE-MTJ devices in MRAM-LUT circuits realizing approximately 67% and 61% reductions in terms of write energy consumption and area, respectively. Next, fabric-level analysis for the developed HSC-FPGA show that the HSC-FPGA can achieve at least 18%, 70%, and 15% reduction in terms of area, standby power, and read power consumption, respectively, for various ISCAS-89 and ITC-99 benchmark circuits.

Moreover, the impact of process variation on MRAM-LUT and SRAM-LUT circuits is investigated using the Monte Carlo SPICE circuit simulations. The results obtained exhibited an average error rate of 44% for the MRAM-LUT in presence of variations in both CMOS and MTJ devices. The detailed analyses recognized the sense amplifier circuits as the most susceptible portion of the MRAM-LUT circuit to PV. Therefore, we have used a modular redundancy circuit-level approach to improve the PV-tolerance of the MRAM-LUT. The average error rate of the developed MR-based MRAM-LUT circuit was reduced to 12%, while further device-level innovations could reduce the error rate to less than 0.1% as provided in the literature. The PV-tolerance is achieved at the cost of 24% and 6% read power consumption and area overheads compared to regular MRAM-LUT, while the standby and write power consumptions remain unchanged. The fabric-level simulations show that the MR-MRAM based HSC-FPGA realizes at least 9% and 17% read power and area reductions compared to conventional SRAM-based FPGAs, while maintaining the 70% reduction in standby power, which can be further decreased by the power-gating allowed by the non-volatility feature of MRAM-LUTs.

An orthogonal dimension of fabric heterogeneity is also non-determinism enabled by either low-voltage CMOS or probabilistic emerging devices. It can be realized using probabilistic devices within a reconfigurable network to blend deterministic and probabilistic computational models. Herein, we developed a hybrid CMOS/spin-based DBN implementation using p-bit based activation functions modeled to produce a probabilistic output that can be modulated by an input current. The device-level simulations exhibited a sigmoid relation between the input currents and output probability. The SPICE model of the p-bit is used to design a weighted array structure to implement RBM. The circuit simulations showed that the performance of the array can be improved by enlarging the array size, as well as reducing the resistance of the weighted connections. However, these improvements are achieved at the cost of increased area and power consumption. For instance, the lowest power dissipation among the examined designs belongs to an 8×8 array with the maximum resistance of $1M\Omega$ for weighted connections. However, this structure can only provide the output probabilities ranging from 0.175 to 0.77, which is the narrowest range among the examined designs resulting in a DBN implementation with lowest accuracy.

Next, we simulated a DBN for digit recognition application in MATLAB using the device and circuit-level behavioral models. Trade-offs include the relations between the recognition accuracy of the DBN and the number of training samples, which are comparable to conventional hardware implementations. The recognition error rate decreased substantially for the first thousand training samples, regardless of the size of the array, while benefits continue through several thousand inputs. However, at least two hidden layers are desirable to achieve suitable error rates. Finally, we have provided a comparison between previous hardware-based RBM implementations and our design with an emphasis on the probabilistic activation function within the neuron structure. The results exhibited that the p-bit based activation function can achieve roughly three orders of magnitude energy improvement, while realizing at least 90X reduction in terms of device count, compared to the previous most energy-efficient designs.

Finally, we proposed a spintronic neuromorphic reconfigurable array (SNRA) that offers an intriguing architectural approach to realize beyond von-Neumann paradigms which embrace both probabilistic and Boolean computation. As developed herein, the inclusion of in-field programmability offers several practical benefits beyond simulation towards a feasible post-Moore fabric. Most importantly, it can accommodate process variation issues that would otherwise preclude the validity of the baseline training values that differ from the manufactured component.

To coordinate training, a four-state FSM is shown to be sufficient to implement the contrastive divergence (CD) algorithm, as well as the control circuitry for the test operation of DBNs with various topologies. The proposed FSM is capable of unsupervised training of an RBM in N+3 clocks where N denoted the number of nodes in the hidden layer of RBM. Interpolating the synthesis results indicate a conventional FPGA footprint can accommodate training circuitry for significantly deeper belief networks. This is facilitated using the flexible allocation and routing of layers and their downstream destinations which is a central tenant of CD training. For instance, it was shown that the FSM for both 784×500×10 and 784×500×500×10 DBN topologies can be implemented with 1,771 LUTs, since the size of the largest RBM in both networks is 784×500.

Beyond the flexible architectural approach, within the SNRA fabric, the device parameters are tuned to realize either stochastic switching or deterministic behavior. In particular, near-zero energy barrier SHE-MTJ devices are used to provide a natural probabilistic sigmoidal function required for implementation of the neuron's activation function within an RBM structure. Meanwhile, non-volatile SHE-MTJ devices with high energy barrier ($\Delta \geq 40kT$) can be used to implement LUTs. Use of SHE-MTJ based LUTs achieves more than 80% and 50% reduction in terms of power dissipation and area, respectively, compared to conventional SRAM-based reconfigurable fabrics. These improvements are achieved at the cost of higher energy consumption during the reconfiguration operation, which occurs rarely and can be tolerated due to the significant area and power reductions realized during the normal operation of the SNRA.

Next, it was shown that a volatge-based embedded MRAM-based neurons with thermally unstable superparamagnetic MTJs can realize a probabilistic output that can be modulated by an input voltage. The device-level simulations exhibited a desired sigmoidal relation between the input voltages and output probability of the neuron. Once the functionality of the proposed stochastic neuron was verified, we have developed an embedded MRAM-based RBM leveraging two resistive crossbar arrays with differential amplifiers to implement the matrix multiplication operation for both positive and negative weights. SPICE circuit simulations for a 2×2 weighted array validated the functionality of the proposed embedded MRAM-based RBM.

To provide a circuit-level implementation of DBN, we have developed a PIN-Sim framework which is a transportable framework for rapid, automated, and accurate design space exploration of hybrid

CMOS and post-CMOS neuromorphic circuits. PIN-Sim is composed of five main modules to train the network, map the trained weights to their corresponding resistances, create the SPICE model of the RBMs, and measure the accuracy and energy consumption. MNIST dataset is utilized to investigate the accuracy and energy tradeoffs for seven distinct DBN topologies implemented by the developed PIN-Sim framework. The simulation results showed that at least two hidden layers are required to achieve suitable error rates. In particular, a $784 \times 200 \times 10$ DBN can realize 5% error rate while consuming less than 500 pJ energy. The error rates could be decreased to 2.5% by using a $784 \times 500 \times 500 \times 500 \times 10$ DBN topologies trained by 10,000 input training samples at the cost of $\sim 10\times$ higher energy consumption and significantly larger area overheads. Moreover, PIN-Sim can be used to optimize network topologies based on different application requirements for energy versus accuracy tradeoffs.

Finally, we have focused on the effect of various hardware-level parameters that can be adjusted in the PIN-Sim tool on the performance of the network. One particular parameter which is specifically important for MTJ and RRAM based crossbar architectures is the difference between the largest and smallest possible resistance values in a weighted connection (ΔR_W). It was shown that at least a ΔR_W of 400% is required to realize suitable error rates, however it is worth noting that increasing the ΔR_W to values larger than 400% does not lead to a significant reduction in error rate. Therefore, some fabrication complexities for increasing the ΔR_W in MTJ-based weighted connections can be avoided. Moreover, to realize a realistic hardware implementation we have studied the effect of weight quantization on the accuracy of our network. It was shown that a quantization factor of eight, which provides eight different resistive levels in each weighted connection, can lead to even lower error rates compared to a network with unquantized weights. This also shows the robustness of our proposed circuit-level DBN implementation to minor variations in the resistance of the weighted connections, which is inevitable during the fabrication process. Finally, the comparison results exhibited that the embedded MRAM-based neuron can contribute to several orders of magnitude energy reduction, and reduce the area requirement by 20-fold, with respect to recent energy-optimized designs. Although this is a simulation-based result, hardware realization may endure significant process variation and impacts of sneak currents in large crossbar arrays. To address these further, the development of the PIN-Sim framework provides several possibilities for future work, including: (1) leveraging optimization techniques to reduce the performance gap between the ideal implementation of the DBN using simulation tools such as MATLAB, and the realistic circuit-level implementation of DBN using PIN-Sim framework, (2) training DBNs with binary weights which can be implemented by MTJs or RRAMs, (3) implementing convolutional DBNs using PIN-Sim for more complex pattern recognition applications.

8.2 Future Directions

8.2.1 Pinpoint Vertical Integration of Spintronic Devices for Reconfigurable Resiliency

Aggressive CMOS technology scaling in digital circuits has resulted in significant increase in manufacturing defects and transient fault rates that consequently reduces the performance and reliability of the emerging VLSI circuits. By the extensions to sub-10nm regimes, error resiliency has become a major challenge for microelectronics industry, particularly mission critical systems, e.g. space and terrestrial applications. Research to overcome these challenges has been focused on measuring and analyzing electrical parameters within the structure of a very large scale logic circuit. However, the exponential growth in the complexity of the modern digital circuits has made these extensive measurement approaches impractical, due to the extreme computation cost and effort. Utilizing reconfigurable computing by applying hardware and time redundancy to the digital circuits offers promising and robust technique for addressing the aforementioned reliability challenges. However, in addition to the induced energy consumption and area overheads due to the applied logic-level redundancy, approaches relying on golden elements that are assumed to be fault-free, while in practice they employ conventional MOS technology in their structure, and also suffer from mentioned scaling challenges, as well as susceptibility to radiation-induced soft errors. Meanwhile spintronic devices offer radiation immunity and incur near-zero standby energy consumption, making them ideal elements for these golden components. They also realize a fabric of amorphous resources in standby mode to regain lost functionality from hard or intermittent faults, and PV. The challenge is to provide a flexible yet effective mapping of spintronic devices to adapt their functionality for resilience at run-time.

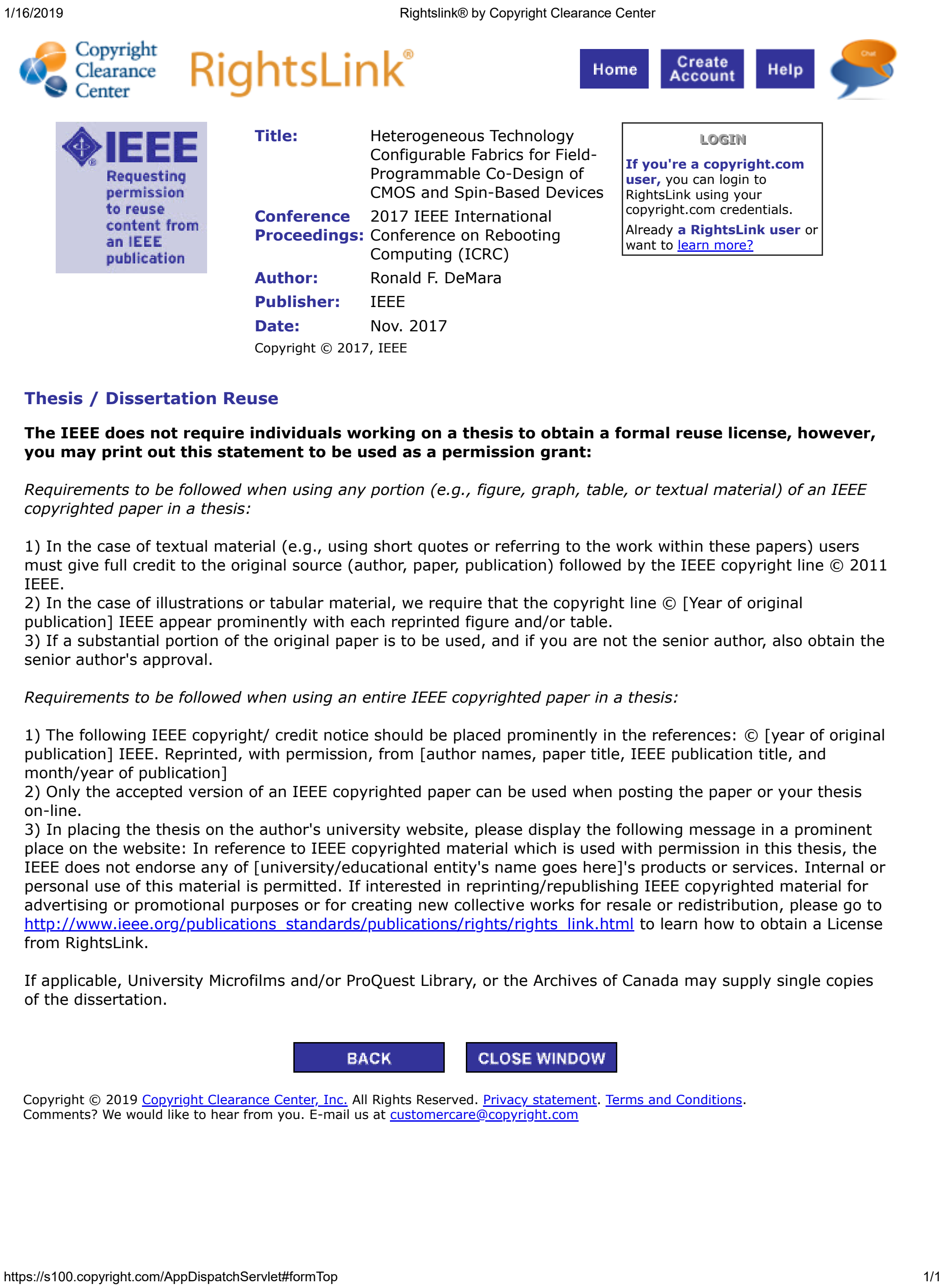
MTJ-based LUTs can be placed at the critical points of an ASIC to implement various logic functions as a run-time adaptable fabric under middleware control. Radiation immunity of MTJ devices decrease the susceptibility of these golden elements of the design to radiation-induced errors. Moreover, the pinpointed magnetic LUTs provides the fabric with sufficient reconfigurability features to mitigate PV. The fabric will be leveraged for fault detection and recovery using the adaptive self-healing approaches. Synthesis tools should be utilized to determine the composition of the fabric, as well as synthesizing and optimizing the HDL codes. MTJs comprising the storage elements in the adaptable LUTs are vertically-integrated as a backend process of typical CMOS fabrication. In addition to soft error immunity, this significantly reduces the area cost of the redundancy. Therefore, the research presented in this dissertation can serve as a framework to investigate innovations in device, circuit, and architecture levels to address the following Research Questions:

- *How can the cooperating advantages of CMOS and spin-based devices be leveraged within golden elements, both at design-time and at recovery-time?*
- *What is the necessary and sufficient quantity of reconfigurability to instill operationally-significant resilience without incurring detrimental overheads?*
- *How can spintronic devices facilitate scalable and cost-effective techniques to address the reliability challenges of highly-scaled FPGA and ASIC platforms?*

8.2.2 Device-Cognizant Design Space Exploration for Neuromorphic Hardware

Benefits of alternatives to von Neumann architectures for neuromorphic applications include avoidance of the processor-memory bottleneck, reduced energy consumption, and area-sparing computation. Viable solutions to the challenge of designing neuromorphic architectures span the interrelated fields of machine learning, computer architecture, circuit design, and the potential to leverage the complementary characteristics of emerging device technologies. Using emerging technologies within neuromorphic architectures attempts to utilize the more complex intrinsic switching behaviors of the devices at-hand to achieve significant reductions in energy and execution time. However, in order to surmount the device-level and circuit-level challenges such as susceptibility to noise and process variation that are introduced when developing emerging neuromorphic circuits, more flexible, powerful, and intelligent simulation frameworks will be required. Thus, machine intelligence techniques are sought to optimize neuromorphic architectures by placing these aspects within-the-loop of the design process. Advanced Cross-Layer neuromorphic simulation frameworks and design methodologies are proposed to be researched to explore the neuromorphic hardware design space in various architecture-to-device granularities to realize an optimized circuit-level implementation of Deep Neural Networks (DNNs).

APPENDIX : COPYRIGHT PERMISSIONS



1/16/2019 Rightslink® by Copyright Clearance Center

Copyright Clearance Center RightsLink®

IEEE Requesting permission to reuse content from an IEEE publication

Title:	Heterogeneous Technology Configurable Fabrics for Field-Programmable Co-Design of CMOS and Spin-Based Devices
Conference Proceedings:	2017 IEEE International Conference on Rebooting Computing (ICRC)
Author:	Ronald F. DeMara
Publisher:	IEEE
Date:	Nov. 2017

Copyright © 2017, IEEE

If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials. Already **a RightsLink user** or want to learn more?

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:

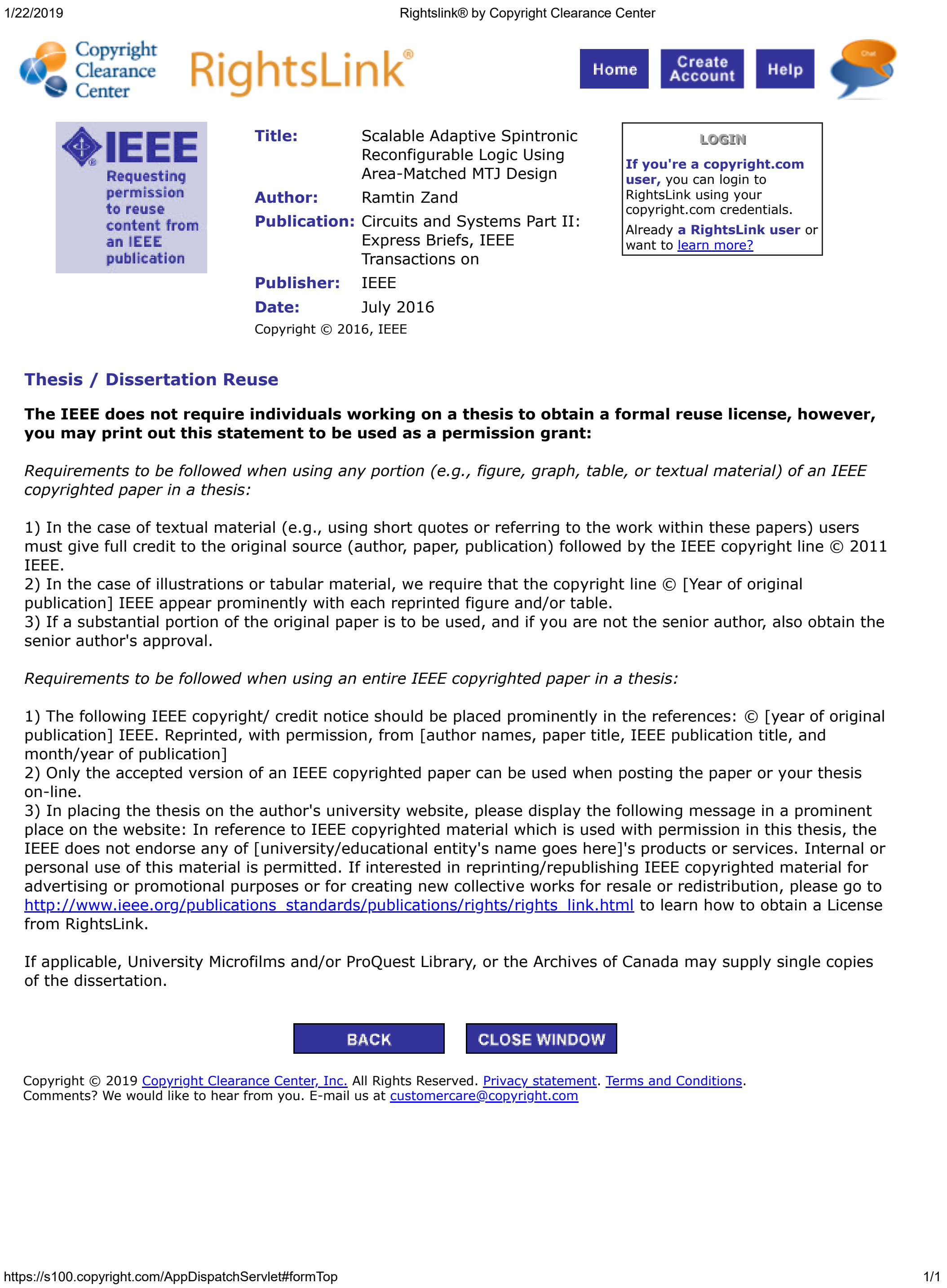
1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that 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.

https://s100.copyright.com/AppDispatchServlet#formTop 1/1


Copyright Clearance Center
RightsLink®
Home
Create Account
Help
Chat
IEEE
Requesting permission to reuse content from an IEEE publication
LOGIN
BACK
CLOSE WINDOW

1/22/2019 Rightslink® by Copyright Clearance Center

Title:	Scalable Adaptive Spintronic Reconfigurable Logic Using Area-Matched MTJ Design
Author:	Ramtin Zand
Publication:	Circuits and Systems Part II: Express Briefs, IEEE Transactions on
Publisher:	IEEE
Date:	July 2016

Copyright © 2016, IEEE

If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials.
Already **a RightsLink user** or want to learn more?

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:

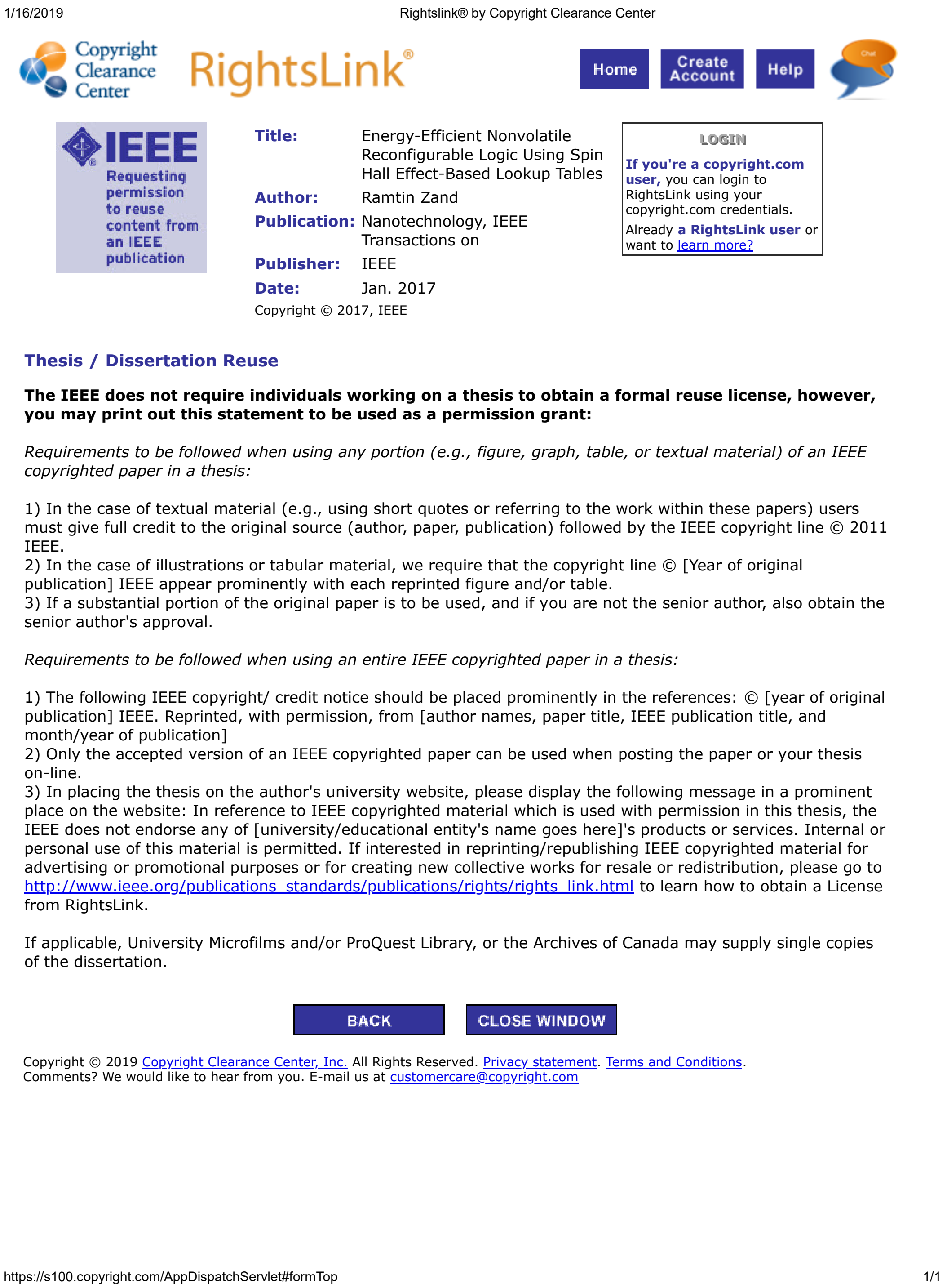
1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that 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.

https://s100.copyright.com/AppDispatchServlet#formTop 1/1


Copyright Clearance Center
RightsLink®
Home
Create Account
Help
Chat
IEEE
Requesting permission to reuse content from an IEEE publication
LOGIN
BACK
CLOSE WINDOW

Title:	Energy-Efficient Nonvolatile Reconfigurable Logic Using Spin Hall Effect-Based Lookup Tables
Author:	Ramtin Zand
Publication:	Nanotechnology, IEEE Transactions on
Publisher:	IEEE
Date:	Jan. 2017

If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials.
Already **a RightsLink user** or want to learn more?

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.

Comments? We would like to hear from you. E-mail us at customercare@copyright.com

Rightslink® by Copyright Clearance Center https://s100.copyright.com/AppDispatchServlet#formTop

Copyright Clearance Center
RightsLink®
Home
Create Account
Help
Chat
IEEE
Requesting permission to reuse content from an IEEE publication

Title:	SNRA: A Spintronic Neuromorphic Reconfigurable Array for In-Circuit Training and Evaluation of Deep Belief Networks
Conference Proceedings:	2018 IEEE International Conference on Rebooting Computing (ICRC)
Author:	Ramtin Zand
Publisher:	IEEE
Date:	Nov. 2018

LOGIN

If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials.

Already **a RightsLink user** or want to learn more?

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.

1 of 1 3/4/2019, 4:23 PM

LIST OF REFERENCES

[1] R. F. DeMara, A. Roohi, R. Zand, and S. D. Pyle, "Heterogeneous technology configurable fabrics for field-programmable co-design of cmos and spin-based devices," in *2017 IEEE International Conference on Rebooting Computing (ICRC)*, Nov 2017, pp. 1–4.

[2] R. Zand, A. Roohi, D. Fan, and R. F. DeMara, "Energy-efficient nonvolatile reconfigurable logic using spin hall effect-based lookup tables," *IEEE Transactions on Nanotechnology*, vol. 16, no. 1, pp. 32–43, Jan 2017.

[3] R. Zand, A. Roohi, S. Salehi, and R. F. DeMara, "Scalable adaptive spintronic reconfigurable logic using area-matched mtj design," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 63, no. 7, pp. 678–682, July 2016.

[4] R. Zand and R. F. DeMara, "Radiation-hardened mram-based lut for non-volatile fpga soft error mitigation with multi-node upset tolerance," *Journal of Physics D: Applied Physics*, vol. 50, no. 50, p. 505002, 2017.

[5] 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, p. 721, 2010.

[6] R. Zand and R. F. DeMara, "Snra: A spintronic neuromorphic reconfigurable array for in-circuit training and evaluation of deep belief networks," in *2018 IEEE International Conference on Rebooting Computing (ICRC)*, Nov 2018, pp. 1–9.

[7] K. Y. Camsari, R. Faria, B. M. Sutton, and S. Datta, "Stochastic p-bits for invertible logic," *Phys. Rev. X*, vol. 7, p. 031014, Jul 2017.

[8] 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.

[9] R. Zand, K. Y. Camsari, S. Datta, and R. F. DeMara, "Composable probabilistic inference networks using mram-based stochastic neurons," *arXiv preprint arXiv:1811.11390*, 2018.

[10] J. Kim, A. Chen, B. Behin-Aein, S. Kumar, J. Wang, and C. H. Kim, "A technology-agnostic mtj spice model with user-defined dimensions for stt-mram scalability studies," in *2015 IEEE Custom Integrated Circuits Conference (CICC)*, Sept 2015, pp. 1–4.

[11] Y. Zhang, W. Zhao, Y. Lakys, J. O. Klein, J. V. Kim, D. Ravelosona, and C. Chappert, "Compact modeling of perpendicular-anisotropy cofeb/mgo magnetic tunnel junctions," *IEEE Transactions on Electron Devices*, vol. 59, no. 3, pp. 819–826, March 2012.

[12] K. Y. Camsari, S. Ganguly, and S. Datta, “Modular approach to spintronics,” *Scientific reports*, vol. 5, p. 10571, 2015.

[13] S. Manipatruni, D. E. Nikonov, and I. A. Young, “Energy-delay performance of giant spin hall effect switching for dense magnetic memory,” *Applied Physics Express*, vol. 7, no. 10, p. 103001, 2014.

[14] S. Motaman, S. Ghosh, and N. Rathi, “Impact of process-variations in sttram and adaptive boosting for robustness,” in *2015 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2015, pp. 1431–1436.

[15] Z. Sun, X. Bi, and H. Li, “Process variation aware data management for stt-ram cache design,” in *Proceedings of the 2012 ACM/IEEE International Symposium on Low Power Electronics and Design*, ser. ISLPED ’12. New York, NY, USA: ACM, 2012, pp. 179–184. [Online]. Available: http://doi.acm.org/10.1145/2333660.2333706

[16] K. J. Kuhn, “Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale cmos,” in *2007 IEEE International Electron Devices Meeting*, Dec 2007, pp. 471–474.

[17] 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.

[18] T. M. Conte, E. P. DeBenedictis, P. A. Gargini, and E. Track, “Rebooting computing: The road ahead,” *Computer*, vol. 50, no. 1, pp. 20–29, Jan 2017.

[19] J. Andrade, N. George, K. Karras, D. Novo, V. Silva, P. Ienne, and G. Falcao, “Fast design space exploration using vivado hls: Non-binary ldpc decoders,” in *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines*, May 2015, pp. 97–97.

[20] S. M. Trimberger, “Three ages of fpgas: A retrospective on the first thirty years of fpga technology,” *Proceedings of the IEEE*, vol. 103, no. 3, pp. 318–331, March 2015.

[21] A. Aysu and P. Schaumont, “Hardware/software co-design of physical unclonable function based authentications on fpgas,” *Microprocessors and Microsystems*, vol. 39, no. 7, pp. 589 – 597, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0141933115000447

[22] L. Contreras, S. Cruz, J. Motta, and C. H. Llanos, “Hardware and software co-design for the ekf applied to the mobile robotics localization problem,” *International Journal of Machine Learning and Computing*, vol. 5, no. 2, p. 101, 2015.

[23] E. Kadric, D. Lakata, and A. DeHon, “Impact of memory architecture on fpga energy consumption,” in *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA ’15. New York, NY, USA: ACM, 2015, pp. 146–155. [Online]. Available: http://doi.acm.org/10.1145/2684746.2689062

[24] A. Ahari, M. Ebrahimi, and M. B. Tahoori, “Energy efficient partitioning of dynamic reconfigurable mram-fpgas,” in *2015 25th International Conference on Field Programmable Logic and Applications (FPL)*, Sep. 2015, pp. 1–6.

[25] J. Xie, P. K. Meher, and Z. Mao, “High-throughput finite field multipliers using redundant basis for fpga and asic implementations,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 1, pp. 110–119, Jan 2015.

[26] P. Malk, “High throughput floating-point dividers implemented in fpga,” in *2015 IEEE 18th International Symposium on Design and Diagnostics of Electronic Circuits Systems*, April 2015, pp. 291–294.

[27] R. Chen, S. G. Singapura, and V. K. Prasanna, “Optimal dynamic data layouts for 2d fft on 3d memory integrated fpga,” *The Journal of Supercomputing*, vol. 73, no. 2, pp. 652–663, 2017.

[28] S. Di Carlo, P. Prinetto, P. Trotta, and J. Andersson, “A portable open-source controller for safe dynamic partial reconfiguration on xilinx fpgas,” in *2015 25th International Conference on Field Programmable Logic and Applications (FPL)*, Sep. 2015, pp. 1–4.

[29] N. Imran, R. A. Ashraf, and R. F. DeMara, “Power and quality-aware image processing soft-resilience using online multi-objective gas,” *International Journal of Computational Vision and Robotics*, vol. 5, no. 1, pp. 72–98, 2015.

[30] F. Serrano, J. A. Clemente, and H. Mecha, “A methodology to emulate single event upsets in flip-flops using fpgas through partial reconfiguration and instrumentation,” *IEEE Transactions on Nuclear Science*, vol. 62, no. 4, pp. 1617–1624, Aug 2015.

[31] R. S. Oreifej, R. Al-Haddad, R. Zand, R. A. Ashraf, and R. F. DeMara, “Survivability modeling and resource planning for self-repairing reconfigurable device fabrics,” *IEEE Transactions on Cybernetics*, vol. 48, no. 2, pp. 780–792, Feb 2018.

[32] R. Al-Haddad, R. S. Oreifej, R. Zand, A. Ejnioui, and R. F. DeMara, “Adaptive mitigation of radiation-induced errors and tddb in reconfigurable logic fabrics,” in *2015 IEEE 24th North Atlantic Test Workshop*, May 2015, pp. 23–32.

[33] A. Alzahrani and R. F. DeMara, “Fast online diagnosis and recovery of reconfigurable logic fabrics using design disjunction,” *IEEE Transactions on Computers*, vol. 65, no. 10, pp. 3055–3069, Oct 2016.

[34] R. A. Ashraf, A. Al-Zahrani, N. Khoshavi, R. Zand, S. Salehi, A. Roohi, M. Lin, and R. F. DeMara, “Reactive rejuvenation of cmos logic paths using self-activating voltage domains,” in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2015, pp. 2944–2947.

[35] R. F. DeMara and K. Zhang, “Autonomous fpga fault handling through competitive runtime reconfiguration,” in *2005 NASA/DoD Conference on Evolvable Hardware (EH’05)*, June 2005, pp. 109–116.

[36] R. S. Oreifej, C. A. Sharma, and R. F. DeMara, “Expediting ga-based evolution using group testing techniques for reconfigurable hardware,” in *2006 IEEE International Conference on Reconfigurable Computing and FPGA’s (ReConFig 2006)*, Sep. 2006, pp. 1–8.

[37] J. Lohn, G. Larchev, and R. DeMara, “Evolutionary fault recovery in a virtex fpga using a representation that incorporates routing,” in *Proceedings International Parallel and Distributed Processing Symposium*, April 2003, pp. 8 pp.–.

[38] C. A. Sharma, A. Sarvi, A. Alzahrani, and R. F. DeMara, “Self-healing reconfigurable logic using autonomous group testing,” *Microprocessors and Microsystems*, vol. 37, no. 2, pp. 174 – 184, 2013, digital System Safety and Security. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S014193311200172X

[39] R. A. Ashraf and R. F. DeMara, “Scalable fpga refurbishment using netlist-driven evolutionary algorithms,” *IEEE Transactions on Computers*, vol. 62, no. 8, pp. 1526–1541, Aug 2013.

[40] N. Imran, J. Lee, and R. F. DeMara, “Fault demotion using reconfigurable slack (fadres),” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 7, pp. 1364–1368, July 2013.

[41] M. Alawad, Y. Bai, R. DeMara, and M. Lin, “Energy-efficient multiplier-less discrete convolver through probabilistic domain transformation,” in *Proceedings of the 2014 ACM/SIGDA International Symposium on Field-programmable Gate Arrays*, ser. FPGA ’14. New York, NY, USA: ACM, 2014, pp. 185–188. [Online]. Available: http://doi.acm.org/10.1145/2554688.2554769

[42] D. Sander, S. O. Valenzuela, D. Makarov, C. H. Marrows, E. E. Fullerton, P. Fischer, J. McCord, P. Vavassori, S. Mangin, P. Pirro, B. Hillebrands, A. D. Kent, T. Jungwirth, O. Gutfleisch, C. G. Kim, and A. Berger, “The 2017 magnetism roadmap,” *Journal of Physics D: Applied Physics*, vol. 50, no. 36, p. 363001, 2017. [Online]. Available: http://stacks.iop.org/0022-3727/50/i=36/a=363001

[43] W. Kang, W. Zhao, Z. Wang, J. Klein, Y. Zhang, D. Chabi, Y. Zhang, D. Ravelosona, and C. Chappert, “An overview of spin-based integrated circuits,” in *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*, Jan 2014, pp. 676–683.

[44] B. Behin-Aein, J.-P. Wang, and R. Wiesendanger, “Computing with spins and magnets,” *MRS Bulletin*, vol. 39, no. 8, p. 696702, 2014.

[45] W. Kuang, P. Zhao, J. S. Yuan, and R. F. DeMara, "Design of asynchronous circuits for high soft error tolerance in deep submicrometer cmos circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 3, pp. 410–422, March 2010.

[46] S. Smith, R. DeMara, J. Yuan, D. Ferguson, and D. Lamb, "Optimization of null convention self-timed circuits," *Integration*, vol. 37, no. 3, pp. 135 – 165, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167926003001068

[47] S. Smith, R. DeMara, J. Yuan, M. Hagedorn, and D. Ferguson, "Delay-insensitive gate-level pipelining," *Integration*, vol. 30, no. 2, pp. 103 – 131, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016792600100013X

[48] A. M. Chabi, A. Roohi, H. Khademolhosseini, S. Sheikhfaal, S. Angizi, K. Navi, and R. F. DeMara, "Towards ultra-efficient qca reversible circuits," *Microprocessors and Microsystems*, vol. 49, pp. 127 – 138, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0141933116302435

[49] A. Roohi, R. Zand, S. Angizi, and R. F. DeMara, "A parity-preserving reversible qca gate with self-checking cascadable resiliency," *IEEE Transactions on Emerging Topics in Computing*, vol. 6, no. 4, pp. 450–459, Oct 2018.

[50] S. Angizi, S. Sayedsalehi, A. Roohi, N. Bagherzadeh, and K. Navi, "Design and verification of new n-bit quantum-dot synchronous counters using majority function-based jk flip-flops," *Journal of Circuits, Systems and Computers*, vol. 24, no. 10, p. 1550153, 2015.

[51] A. Roohi, S. Sayedsalehi, H. Khademolhosseini, and K. Navi, "Design and evaluation of a reconfigurable fault tolerant quantum-dot cellular automata gate," *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 2, pp. 380–388, 2013.

[52] K. Roy, D. Fan, X. Fong, Y. Kim, M. Sharad, S. Paul, S. Chatterjee, S. Bhunia, and S. Mukhopadhyay, "Exploring spin transfer torque devices for unconventional computing," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 5, no. 1, pp. 5–16, March 2015.

[53] M. Sharad, C. Augustine, and K. Roy, "Boolean and non-boolean computation with spin devices," in *2012 International Electron Devices Meeting*, Dec 2012, pp. 11.6.1–11.6.4.

[54] D. Fan, Y. Shim, A. Raghunathan, and K. Roy, "Stt-snn: A spin-transfer-torque based soft-limiting non-linear neuron for low-power artificial neural networks," *IEEE Transactions on Nanotechnology*, vol. 14, no. 6, pp. 1013–1023, Nov 2015.

[55] W. kang, W. Zhao, E. Deng, J.-O. Klein, Y. Cheng, D. Ravelosona, Y. Zhang, and C. Chappert, "A radiation hardened hybrid spintronic/cmos nonvolatile unit using magnetic tunnel junctions," *Journal of Physics D: Applied Physics*, vol. 47, no. 40, p. 405003, 2014. [Online]. Available: http://stacks.iop.org/0022-3727/47/i=40/a=405003

[56] Y. Lakys, W. S. Zhao, J. Klein, and C. Chappert, “Hardening techniques for mram-based nonvolatile latches and logic,” *IEEE Transactions on Nuclear Science*, vol. 59, no. 4, pp. 1136–1141, Aug 2012.

[57] W. Wen, Y. Zhang, Y. Chen, Y. Wang, and Y. Xie, “Ps3-ram: A fast portable and scalable statistical stt-ram reliability/energy analysis method,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 33, no. 11, pp. 1644–1656, Nov 2014.

[58] E. Eken, Y. Zhang, W. Wen, R. Joshi, H. Li, and Y. Chen, “A novel self-reference technique for stt-ram read and write reliability enhancement,” *IEEE Transactions on Magnetics*, vol. 50, no. 11, pp. 1–4, Nov 2014.

[59] J. He and J. Rose, “Advantages of heterogeneous logic block architectures for fpgas,” in *Proceedings of the IEEE Custom Integrated Circuits Conference*, vol. 7, 1993, pp. 1–7.

[60] J. Cong and S. Xu, “Delay-optimal technology mapping for fpgas with heterogeneous luts,” in *Proceedings 1998 Design and Automation Conference. 35th DAC. (Cat. No.98CH36175)*, June 1998, pp. 704–707.

[61] A. Koorapaty, V. Chandra, K. Y. Tong, C. Patel, L. Pileggi, and H. Schmit, “Heterogeneous programmable logic block architectures,” in *Proceedings of the Conference on Design, Automation and Test in Europe - Volume 1*, ser. DATE ’03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 11 118–. [Online]. Available: http://dl.acm.org/citation.cfm?id=789083.1022880

[62] I. Kuon, R. Tessier, and J. Rose, “Fpga architecture: Survey and challenges,” *Foundations and Trends in Electronic Design Automation*, vol. 2, no. 2, pp. 135–253, 2008. [Online]. Available: http://dx.doi.org/10.1561/1000000005

[63] P. Specification, “Virtex-ii pro and virtex-ii pro x platform fpgas: Complete data sheet,” *DS083 v4*, vol. 5, 2002.

[64] D. S. Xilinx, “Spartan-3an fpga family data sheet,” *DS557-2 (v3. 1) June*, vol. 2, pp. 1–8, 2008.

[65] S. Douglass, “Introducing the virtex-5 fpga family,” *Xcell Journal, Xilinx*, no. 59, pp. 8–11, 2006.

[66] D. Lewis, E. Ahmed, D. Cashman, T. Vanderhoek, C. Lane, A. Lee, and P. Pan, “Architectural enhancements in stratix-iii™and stratix-iv™,” in *Proceedings of the ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, ser. FPGA ’09. New York, NY, USA: ACM, 2009, pp. 33–42. [Online]. Available: http://doi.acm.org/10.1145/1508128.1508135

[67] D. E. Nikonov and I. A. Young, “Overview of beyond-cmos devices and a uniform methodology for their benchmarking,” *Proceedings of the IEEE*, vol. 101, no. 12, pp. 2498–2533, Dec 2013.

[68] J. Kim, A. Paul, P. A. Crowell, S. J. Koester, S. S. Sapatnekar, J. Wang, and C. H. Kim, "Spin-based computing: Device concepts, current status, and a case study on a high-performance microprocessor," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 106–130, Jan 2015.

[69] I. Kuon and J. Rose, "Measuring the gap between fpgas and asics," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 2, pp. 203–215, Feb 2007.

[70] P. Pavan, R. Bez, P. Olivo, and E. Zanoni, "Flash memory cells-an overview," *Proceedings of the IEEE*, vol. 85, no. 8, pp. 1248–1271, Aug 1997.

[71] J. Greene, S. Kaptanoglu, W. Feng, V. Hecht, J. Landry, F. Li, A. Krouglyanskiy, M. Morosan, and V. Pevzner, "A 65nm flash-based fpga fabric optimized for low cost and power," in *Proceedings of the 19th ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, ser. FPGA '11. New York, NY, USA: ACM, 2011, pp. 87–96. [Online]. Available: http://doi.acm.org/10.1145/1950413.1950434

[72] J. Yang, X. Wang, Q. Zhou, Z. Wang, H. Li, Y. Chen, and W. Zhao, "Exploiting spin-orbit torque devices as reconfigurable logic for circuit obfuscation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. PP, no. 99, pp. 1–1, 2018.

[73] W. Zhao, D. Ravelosona, J. Klein, and C. Chappert, "Domain wall shift register-based reconfigurable logic," *IEEE Transactions on Magnetics*, vol. 47, no. 10, pp. 2966–2969, Oct 2011.

[74] W. Zhao, N. B. Romdhane, Y. Zhang, J. Klein, and D. Ravelosona, "Racetrack memory based reconfigurable computing," in *2013 IEEE Faible Tension Faible Consommation*, June 2013, pp. 1–4.

[75] D. Suzuki, M. Natsui, A. Mochizuki, S. Miura, H. Honjo, H. Sato, S. Fukami, S. Ikeda, T. Endoh, H. Ohno, and T. Hanyu, "Fabrication of a 3000-6-input-luts embedded and block-level power-gated nonvolatile fpga chip using p-mtj-based logic-in-memory structure," in *2015 Symposium on VLSI Technology (VLSI Technology)*, June 2015, pp. C172–C173.

[76] D. Suzuki, M. Natsui, T. Endoh, H. Ohno, and T. Hanyu, "Six-input lookup table circuit with 62% fewer transistors using nonvolatile logic-in-memory architecture with series/parallel-connected magnetic tunnel junctions," *Journal of Applied Physics*, vol. 111, no. 7, p. 07E318, 2012. [Online]. Available: https://doi.org/10.1063/1.3672411

[77] S. M. Williams and M. Lin, "Architecture and circuit design of an all-spintronic fpga," in *Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '18. New York, NY, USA: ACM, 2018, pp. 41–50. [Online]. Available: http://doi.acm.org/10.1145/3174243.3174256

[78] D. Suzuki, M. Natsui, and T. Hanyu, “Area-efficient lut circuit design based on asymmetry of mtj’s current switching for a nonvolatile fpga,” in *2012 IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug 2012, pp. 334–337.

[79] X. Xue, J. Yang, Y. Lin, R. Huang, Q. Zou, and J. Wu, “Low-power variation-tolerant nonvolatile lookup table design,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 3, pp. 1174–1178, March 2016.

[80] X. Tang, G. Kim, P. Gaillardon, and G. D. Micheli, “A study on the programming structures for rram-based fpga architectures,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 4, pp. 503–516, April 2016.

[81] Y. Chen, H. Li, and W. Zhang, “A novel peripheral circuit for rram-based lut,” in *2012 IEEE International Symposium on Circuits and Systems*, May 2012, pp. 1811–1814.

[82] P.-E. Gaillardon, X. Tang, J. Sandrini, M. Thammasack, S. R. Omam, D. Sacchetto, Y. Leblebici, and G. De Micheli, “A ultra-low-power fpga based on monolithically integrated rrams,” in *Proceedings of the 2015 Design, Automation &#38; Test in Europe Conference &#38; Exhibition*, ser. DATE ’15. San Jose, CA, USA: EDA Consortium, 2015, pp. 1203–1208. [Online]. Available: http://dl.acm.org/citation.cfm?id=2755753.2757090

[83] C. Y. Wen, J. Li, S. Kim, M. Breitwisch, C. Lam, J. Paramesh, and L. T. Pileggi, “A non-volatile look-up table design using pcm (phase-change memory) cells,” in *2011 Symposium on VLSI Circuits - Digest of Technical Papers*, June 2011, pp. 302–303.

[84] K. Huang, Y. Ha, R. Zhao, A. Kumar, and Y. Lian, “A low active leakage and high reliability phase change memory (pcm) based non-volatile fpga storage element,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 9, pp. 2605–2613, Sep. 2014.

[85] M. K. G. Krishna, A. Roohi, R. Zand, and R. F. DeMara, “Heterogeneous energy-sparing reconfigurable logic: spin-based storage and cnfet-based multiplexing,” *IET Circuits, Devices Systems*, vol. 11, no. 3, pp. 274–279, 2017.

[86] A. Roohi, R. Zand, and R. F. DeMara, “A tunable majority gate-based full adder using current-induced domain wall nanomagnets,” *IEEE Transactions on Magnetics*, vol. 52, no. 8, pp. 1–7, Aug 2016.

[87] U. Legat, A. Biasizzo, and F. Novak, “Seu recovery mechanism for sram-based fpgas,” *IEEE Transactions on Nuclear Science*, vol. 59, no. 5, pp. 2562–2571, Oct 2012.

[88] R. Rajaei, “Radiation-hardened design of nonvolatile mram-based fpga,” *IEEE Transactions on Magnetics*, vol. 52, no. 10, pp. 1–10, Oct 2016.

[89] ——, “Single event double node upset tolerance in mos/spintronic sequential and combinational logic circuits,” *Microelectronics Reliability*, vol. 69, pp. 109–114, 2017.

[90] C. Constantinescu, "Trends and challenges in vlsi circuit reliability," *IEEE Micro*, vol. 23, pp. 14–19, 07 2003. [Online]. Available: doi.ieeecomputersociety.org/10.1109/MM.2003.1225959

[91] S. Ghosh and K. Roy, "Parameter variation tolerance and error resiliency: New design paradigm for the nanoscale era," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1718–1751, Oct 2010.

[92] L. Buesing, J. Bill, B. Nessler, and W. Maass, "Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons," *PLoS computational biology*, vol. 7, no. 11, p. e1002211, 2011.

[93] 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.

[94] 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*, vol. 22, no. 4, pp. 778–784, April 2014.

[95] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

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

[97] D. L. 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, Nov 2010.

[98] 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, 2017.

[99] 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.

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

[101] S. B. Eryilmaz, E. Neftci, S. Joshi, S. Kim, M. BrightSky, H. L. Lung, C. 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, Dec 2016.

[102] 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.

[103] 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.

[104] D. E. Nikonov and I. A. Young, "Benchmarking of beyond-cmos exploratory devices for logic integrated circuits," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 1, pp. 3–11, Dec 2015.

[105] S. Ghosh, A. Iyengar, S. Motaman, R. Govindaraj, J. W. Jang, J. Chung, J. Park, X. Li, R. Joshi, and D. Somasekhar, "Overview of circuits, systems, and applications of spintronics," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 3, 2016.

[106] B. Liu, M. Hu, H. Li, Z.-H. Mao, Y. Chen, T. Huang, and W. Zhang, "Digital-assisted noise-eliminating training for memristor crossbar-based analog neuromorphic computing engine," in *2013 50th ACM/EDAC/IEEE Design Automation Conference (DAC)*, 2013.

[107] A. Roohi, R. Zand, D. Fan, and R. F. DeMara, "Voltage-based concatenatable full adder using spin hall effect switching," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 36, no. 12, pp. 2134–2138, Dec 2017.

[108] R. Zand, A. Roohi, and R. F. DeMara, "Energy-efficient and process-variation-resilient write circuit schemes for spin hall effect mram device," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 9, pp. 2394–2401, Sept 2017.

[109] R. Zand and R. F. DeMara, "Hsc-fpga," in *Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '19. New York, NY, USA: ACM, 2019, pp. 118–119. [Online]. Available: http://doi.acm.org/10.1145/3289602.3293940

[110] S. Salehi, N. Khoshavi, R. Zand, and R. F. DeMara, "Self-organized sub-bank she-mram-based llc: An energy-efficient and variation-immune read and write architecture," *Integration*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167926017303425

[111] A. Roohi, R. Zand, and R. F. DeMara, "Synthesis of normally-off boolean circuits: An evolutionary optimization approach utilizing spintronic devices," in *2018 19th International Symposium on Quality Electronic Design (ISQED)*, March 2018, pp. 49–54.

[112] A. Roohi, R. Zand, and R. F. DeMara, “Logic-encrypted synthesis for energy-harvesting-powered spintronic-embedded datapath design,” in *Proceedings of the 2018 on Great Lakes Symposium on VLSI*, ser. GLSVLSI ’18. New York, NY, USA: ACM, 2018, pp. 9–14. [Online]. Available: http://doi.acm.org/10.1145/3194554.3194557

[113] F. S. Alghareb, R. Zand, and R. F. Demara, “Non-volatile spintronic flip-flop design for energy-efficient seu and dnu resilience,” *IEEE Transactions on Magnetics*, vol. 55, no. 3, pp. 1–11, March 2019.

[114] F. S. Alghareb, R. Zand, and R. F. DeMara, “High-performance double node upset-tolerant non-volatile flip-flop design,” in *SoutheastCon 2018*, April 2018, pp. 1–6.

[115] A. Roohi and R. F. DeMara, “Nv-clustering: Normally-off computing using non-volatile datapaths,” *IEEE Transactions on Computers*, vol. 67, no. 7, pp. 949–959, July 2018.

[116] S. Salehi, N. Khoshavi, and R. F. Demara, “Mitigating process variability for non-volatile cache resilience and yield,” *IEEE Transactions on Emerging Topics in Computing*, pp. 1–1, 2018.

[117] S. Salehi and R. F. DeMara, “Process variation immune and energy aware sense amplifiers for resistive non-volatile memories,” in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2017, pp. 1–4.

[118] N. Khoshavi, S. Salehi, and R. F. DeMara, “Variation-immune resistive non-volatile memory using self-organized sub-bank circuit designs,” in *2017 18th International Symposium on Quality Electronic Design (ISQED)*, March 2017, pp. 52–57.

[119] E. Deng, W. Kang, Y. Zhang, J. Klein, C. Chappert, and W. Zhao, “Design optimization and analysis of multicontext stt-mtj/cmos logic circuits,” *IEEE Transactions on Nanotechnology*, vol. 14, no. 1, pp. 169–177, Jan 2015.

[120] W. Zhao, E. Belhaire, Q. Mistral, C. Chappert, V. Javerliac, B. Dieny, and E. Nicolle, “Macro-model of spin-transfer torque based magnetic tunnel junction device for hybrid magnetic-cmos design,” in *2006 IEEE International Behavioral Modeling and Simulation Workshop*, Sep. 2006, pp. 40–43.

[121] S. A. Wolf, D. D. Awschalom, R. A. Buhrman, J. M. Daughton, S. von Molnár, M. L. Roukes, A. Y. Chtchelkanova, and D. M. Treger, “Spintronics: A spin-based electronics vision for the future,” *Science*, vol. 294, no. 5546, pp. 1488–1495, 2001. [Online]. Available: http://science.sciencemag.org/content/294/5546/1488

[122] I. L. Prejbeanu, M. Kerekes, R. C. Sousa, H. Sibuet, O. Redon, B. Dieny, and J. P. Nozires, “Thermally assisted mram,” *Journal of Physics: Condensed Matter*, vol. 19, no. 16, p. 165218, 2007. [Online]. Available: http://stacks.iop.org/0953-8984/19/i=16/a=165218

[123] J. Slonczewski, “Current-driven excitation of magnetic multilayers,” *Journal of Magnetism and Magnetic Materials*, vol. 159, no. 1, pp. L1 – L7, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0304885396000625

[124] X. Fong, Y. Kim, K. Yogendra, D. Fan, A. Sengupta, A. Raghunathan, and K. Roy, “Spin-transfer torque devices for logic and memory: Prospects and perspectives,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 35, no. 1, 2016.

[125] W. Kang, Y. Zhang, Z. Wang, J.-O. Klein, C. Chappert, D. Ravelosona, G. Wang, Y. Zhang, and W. Zhao, “Spintronics: Emerging ultra-low-power circuits and systems beyond mos technology,” *J. Emerg. Technol. Comput. Syst.*, vol. 12, no. 2, pp. 16:1–16:42, Sep. 2015. [Online]. Available: http://doi.acm.org/10.1145/2663351

[126] R. H. Koch, J. A. Katine, and J. Z. Sun, “Time-resolved reversal of spin-transfer switching in a nanomagnet,” *Phys. Rev. Lett.*, vol. 92, p. 088302, Feb 2004. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.92.088302

[127] W. F. Brown, “Thermal fluctuations of a single-domain particle,” *Phys. Rev.*, vol. 130, pp. 1677–1686, Jun 1963. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.130.1677

[128] L. Liu, T. Moriyama, D. C. Ralph, and R. A. Buhrman, “Reduction of the spin-torque critical current by partially canceling the free layer demagnetization field,” *Applied Physics Letters*, vol. 94, no. 12, p. 122508, 2009. [Online]. Available: https://doi.org/10.1063/1.3107262

[129] L. Faber, W. Zhao, J. Klein, T. Devolder, and C. Chappert, “Dynamic compact model of spin-transfer torque based magnetic tunnel junction (mtj),” in *2009 4th International Conference on Design Technology of Integrated Systems in Nanoscal Era*, April 2009, pp. 130–135.

[130] J. Z. Sun, “Spin-current interaction with a monodomain magnetic body: A model study,” *Phys. Rev. B*, vol. 62, pp. 570–578, Jul 2000. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.62.570

[131] T. Devolder, C. Chappert, J. A. Katine, M. J. Carey, and K. Ito, “Distribution of the magnetization reversal duration in subnanosecond spin-transfer switching,” *Phys. Rev. B*, vol. 75, p. 064402, Feb 2007. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.75.064402

[132] H. Zhao, B. Glass, P. K. Amiri, A. Lyle, Y. Zhang, Y.-J. Chen, G. Rowlands, P. Upadhyaya, Z. Zeng, J. A. Katine, J. Langer, K. Galatsis, H. Jiang, K. L. Wang, I. N. Krivorotov, and J.-P. Wang, “Sub-200 ps spin transfer torque switching in in-plane magnetic tunnel junctions with interface perpendicular anisotropy,” *Journal of Physics D: Applied Physics*, vol. 45, no. 2, p. 025001, dec 2011. [Online]. Available: https://doi.org/10.1088%2F0022-3727%2F45%2F2%2F025001

[133] L. Liu, C.-F. Pai, D. C. Ralph, and R. A. Buhrman, “Magnetic oscillations driven by the spin hall effect in 3-terminal magnetic tunnel junction devices,” *Phys. Rev. Lett.*, vol. 109, p. 186602, Oct 2012. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.109.186602

[134] L. Liu, C.-F. Pai, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman, “Spin-torque switching with the giant spin hall effect of tantalum,” *Science*, vol. 336, no. 6081, pp. 555–558, 2012. [Online]. Available: http://science.sciencemag.org/content/336/6081/555

[135] C.-F. Pai, L. Liu, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman, “Spin transfer torque devices utilizing the giant spin hall effect of tungsten,” *Applied Physics Letters*, vol. 101, no. 12, p. 122404, 2012. [Online]. Available: https://doi.org/10.1063/1.4753947

[136] W. Kang, Z. Wang, Y. Zhang, J.-O. Klein, W. Lv, and W. Zhao, “Spintronic logic design methodology based on spin hall effect–driven magnetic tunnel junctions,” *Journal of Physics D: Applied Physics*, vol. 49, no. 6, p. 065008, jan 2016.

[137] S. Rakheja and A. Naeemi, “Graphene nanoribbon spin interconnects for nonlocal spin-torque circuits: Comparison of performance and energy per bit with cmos interconnects,” *IEEE Transactions on Electron Devices*, vol. 59, no. 1, pp. 51–59, Jan 2012.

[138] M. M. Torunbalci, P. Upadhyaya, S. A. Bhave, and K. Y. Camsari, “Modular compact modeling of mtj devices,” *IEEE Transactions on Electron Devices*, pp. 1–7, 2018.

[139] N. S. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J. S. Hu, M. J. Irwin, M. Kandemir, and V. Narayanan, “Leakage current: Moore’s law meets static power,” *Computer*, vol. 36, no. 12, pp. 68–75, Dec 2003.

[140] J. H. Anderson and Q. Wang, “Area-efficient fpga logic elements: Architecture and synthesis,” in *Proceedings of the 16th Asia and South Pacific Design Automation Conference*, ser. ASPDAC ’11. Piscataway, NJ, USA: IEEE Press, 2011, pp. 369–375. [Online]. Available: http://dl.acm.org/citation.cfm?id=1950815.1950894

[141] W. Zhao, C. Chappert, V. Javerliac, and J. Noziere, “High speed, high stability and low power sensing amplifier for mtj/cmos hybrid logic circuits,” *IEEE Transactions on Magnetics*, vol. 45, no. 10, pp. 3784–3787, Oct 2009.

[142] S. Salehi, D. Fan, and R. F. Demara, “Survey of stt-mram cell design strategies: Taxonomy and sense amplifier tradeoffs for resiliency,” *J. Emerg. Technol. Comput. Syst.*, vol. 13, no. 3, pp. 48:1–48:16, Apr. 2017. [Online]. Available: http://doi.acm.org/10.1145/2997650

[143] A. Alzahrani and R. F. DeMara, “Process variation immunity of alternative 16nm hk/mg-based fpga logic blocks,” in *2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug 2015, pp. 1–4.

[144] W. Zhao, E. Belhaire, C. Chappert, F. Jacquet, and P. Mazoyer, “New non-volatile logic based on spin-mtj,” *physica status solidi (a)*, vol. 205, no. 6, pp. 1373–1377, 2008.

[145] D. Lewis, E. Ahmed, G. Baeckler, V. Betz, M. Bourgeault, D. Cashman, D. Galloway, M. Hutton, C. Lane, A. Lee, P. Leventis, S. Marquardt, C. McClintock, K. Padalia, B. Pedersen, G. Powell, B. Ratchev, S. Reddy, J. Schleicher, K. Stevens, R. Yuan, R. Cliff, and J. Rose, “The stratix ii logic and routing architecture,” in *Proceedings of the 2005 ACM/SIGDA 13th International Symposium on Field-programmable Gate Arrays*, ser. FPGA ’05. New York, NY, USA: ACM, 2005, pp. 14–20. [Online]. Available: http://doi.acm.org/10.1145/1046192.1046195

[146] J. U. Horstmann, H. W. Eichel, and R. L. Coates, “Metastability behavior of cmos asic flip-flops in theory and test,” *IEEE Journal of Solid-State Circuits*, vol. 24, no. 1, pp. 146–157, Feb 1989.

[147] D. Lewis, B. Pedersen, S. Kaptanoglu, and A. Lee, “Fracturable lookup table and logic element,” Sep. 13 2005, uS Patent 6,943,580.

[148] A. Percey, “Advantages of the virtex-5 fpga 6-input lut architecture,” *White Paper: Virtex-5 FPGAs, Xilinx WP284 (v1. 0)*, 2007.

[149] W. Zhao and Y. Cao, “Predictive technology model for nano-cmos design exploration,” *J. Emerg. Technol. Comput. Syst.*, vol. 3, no. 1, Apr. 2007. [Online]. Available: http://doi.acm.org/10.1145/1229175.1229176

[150] M. J. Wirthlin, A. M. Keller, C. McCloskey, P. Ridd, D. Lee, and J. Draper, “Seu mitigation and validation of the leon3 soft processor using triple modular redundancy for space processing,” in *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA ’16. New York, NY, USA: ACM, 2016, pp. 205–214. [Online]. Available: http://doi.acm.org/10.1145/2847263.2847278

[151] T. M. Lovelly and A. D. George, “Comparative analysis of present and future space-grade processors with device metrics,” *Journal of Aerospace Information Systems*, vol. 14, no. 3, pp. 184–197, 2017.

[152] D. Malagn, S. Bota, G. Torrens, X. Gili, J. Praena, B. Fernndez, M. Macas, J. Quesada, C. G. Sanchez, M. Jimnez-Ramos, J. G. Lpez, J. Merino, and J. Segura, “Soft error rate comparison of 6t and 8t sram ics using mono-energetic proton and neutron irradiation sources,” *Microelectronics Reliability*, vol. 78, pp. 38 – 45, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0026271417303852

[153] J. L. Barth, C. S. Dyer, and E. G. Stassinopoulos, “Space, atmospheric, and terrestrial radiation environments,” *IEEE Transactions on Nuclear Science*, vol. 50, no. 3, pp. 466–482, June 2003.

[154] Y. Sharma, B. Javadi, W. Si, and D. Sun, “Reliability and energy efficiency in cloud computing systems: Survey and taxonomy,” *Journal of Network and Computer Applications*, vol. 74, pp. 66 – 85, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804516301746

[155] J.-L. Autran, S. Semikh, D. Munteanu, S. Serre, G. Gasiot, and P. Roche, “Soft-error rate of advanced sram memories: Modeling and monte carlo simulation,” in *Numerical Simulation-From Theory to Industry*. InTech, 2012.

[156] P. E. Dodd, “Physics-based simulation of single-event effects,” *IEEE Transactions on Device and Materials Reliability*, vol. 5, no. 3, pp. 343–357, Sep. 2005.

[157] P. Roche, J. M. Palau, G. Bruguier, C. Tavernier, R. Ecoffet, and J. Gasiot, “Determination of key parameters for seu occurrence using 3-d full cell sram simulations,” *IEEE Transactions on Nuclear Science*, vol. 46, no. 6, pp. 1354–1362, Dec 1999.

[158] R. Naseer, Y. Boulghassoul, J. Draper, S. DasGupta, and A. Witulski, “Critical charge characterization for soft error rate modeling in 90nm sram,” in *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2007, pp. 1879–1882.

[159] L. B. Freeman, “Critical charge calculations for a bipolar sram array,” *IBM Journal of Research and Development*, vol. 40, no. 1, pp. 119–129, Jan 1996.

[160] T. Merelle, H. Chabane, J. . Palau, K. Castellani-Coulie, F. Wrobel, F. Saigne, B. Sagnes, J. Boch, J. R. Vaille, G. Gasiot, P. Roche, M. . Palau, and T. Carriere, “Criterion for seu occurrence in sram deduced from circuit and device simulations in case of neutron-induced ser,” *IEEE Transactions on Nuclear Science*, vol. 52, no. 4, pp. 1148–1155, Aug 2005.

[161] S. Yue, X. Zhang, Y. Zhao, L. Liu, and H. Wang, “Modeling and simulation of single-event effect in CMOS circuit,” *Journal of Semiconductors*, vol. 36, no. 11, p. 111002, nov 2015. [Online]. Available: https://doi.org/10.1088%2F1674-4926%2F36%2F11%2F111002

[162] M. Fazeli, S. G. Miremadi, A. Ejlali, and A. Patooghy, “Low energy single event upset/single event transient-tolerant latch for deep submicron technologies,” *IET Computers Digital Techniques*, vol. 3, no. 3, pp. 289–303, May 2009.

[163] W. kang, W. Zhao, E. Deng, J.-O. Klein, Y. Cheng, D. Ravelosona, Y. Zhang, and C. Chappert, “A radiation hardened hybrid spintronic/CMOS nonvolatile unit using magnetic tunnel junctions,” *Journal of Physics D: Applied Physics*, vol. 47, no. 40, p. 405003, sep 2014.

[164] P. E. Dodd and L. W. Massengill, “Basic mechanisms and modeling of single-event upset in digital microelectronics,” *IEEE Transactions on Nuclear Science*, vol. 50, no. 3, pp. 583–602, June 2003.

[165] S. S. Parkin, C. Kaiser, A. Panchula, P. M. Rice, B. Hughes, M. Samant, and S.-H. Yang, “Giant tunnelling magnetoresistance at room temperature with mgo (100) tunnel barriers,” *Nature materials*, vol. 3, no. 12, p. 862, 2004.

[166] B. Zhao, Y. Du, J. Yang, and Y. Zhang, "Process variation-aware nonuniform cache management in a 3d die-stacked multicore processor," *IEEE Transactions on Computers*, vol. 62, no. 11, pp. 2252–2265, Nov 2013.

[167] Z. Sun, X. Bi, and H. Li, "Process variation aware data management for stt-ram cache design," in *Proceedings of the 2012 ACM/IEEE International Symposium on Low Power Electronics and Design*, ser. ISLPED '12. New York, NY, USA: ACM, 2012, pp. 179–184. [Online]. Available: http://doi.acm.org/10.1145/2333660.2333706

[168] K. Tsunekawa, D. D. Djayaprawira, M. Nagai, H. Maehara, S. Yamagata, N. Watanabe, S. Yuasa, Y. Suzuki, and K. Ando, "Giant tunneling magnetoresistance effect in low-resistance cofebmgo(001)cofeb magnetic tunnel junctions for read-head applications," *Applied Physics Letters*, vol. 87, no. 7, p. 072503, 2005. [Online]. Available: https://doi.org/10.1063/1.2012525

[169] Z. Ghaderi, N. Bagherzadeh, and A. Albaqsami, "Stable: Stress-aware boolean matching to mitigate bti-induced snm reduction in sram-based fpgas," *IEEE Transactions on Computers*, vol. 67, no. 1, pp. 102–114, Jan. 2018.

[170] J. A. Walker, M. A. Trefzer, S. J. Bale, and A. M. Tyrrell, "Panda: A reconfigurable architecture that adapts to physical substrate variations," *IEEE Transactions on Computers*, vol. 62, no. 8, pp. 1584–1596, Aug 2013.

[171] J. Yang, X. Wang, Q. Zhou, Z. Wang, H. Li, Y. Chen, and W. Zhao, "Exploiting spin-orbit torque devices as reconfigurable logic for circuit obfuscation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. PP, no. 99, pp. 1–1, 2018.

[172] D. Suzuki and T. Hanyu, "Design of a magnetic-tunnel-junction-oriented nonvolatile lookup table circuit with write-operation-minimized data shifting," *Japanese Journal of Applied Physics*, vol. 57, no. 4S, p. 04FE09, 2018. [Online]. Available: http://stacks.iop.org/1347-4065/57/i=4S/a=04FE09

[173] J. Hayakawa, S. Ikeda, K. Miura, M. Yamanouchi, Y. M. Lee, R. Sasaki, M. Ichimura, K. Ito, T. Kawahara, R. Takemura, T. Meguro, F. Matsukura, H. Takahashi, H. Matsuoka, and H. Ohno, "Current-induced magnetization switching in mgo barrier magnetic tunnel junctions with cofeb-based synthetic ferrimagnetic free layers," *IEEE Transactions on Magnetics*, vol. 44, no. 7, pp. 1962–1967, July 2008.

[174] S. Salehi and R. F. DeMara, "Process variation immune and energy aware sense amplifiers for resistive non-volatile memories," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2017, pp. 1–4.

[175] V. Hung, A. Gonzalez, and R. DeMara, "Towards a context-based dialog management layer for expert systems," in *2009 International Conference on Information, Process, and Knowledge Management*, Feb 2009, pp. 60–65.

[176] V. Hung, M. Elvir, A. Gonzalez, and R. DeMara, “Towards a method for evaluating naturalness in conversational dialog systems,” in *2009 IEEE International Conference on Systems, Man and Cybernetics*, Oct 2009, pp. 1236–1241.

[177] R. F. DeMara and D. I. Moldovan, “Performance indices for parallel marker-propagation.” in *ICPP (1)*, 1991, pp. 658–659.

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

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

[180] 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.

[181] R. Faria, K. Y. Camsari, and S. Datta, “Low-barrier nanomagnets as p-bits for spin logic,” *IEEE Magnetics Letters*, vol. 8, pp. 1–5, 2017.

[182] B. Sutton, K. Y. Camsari, B. Behin-Aein, and S. Datta, “Intrinsic optimization using stochastic nanomagnets,” *Scientific Reports*, vol. 7, 2017.

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

[184] 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.

[185] 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.

[186] 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.

[187] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[188] B. Yuan and K. K. Parhi, “Vlsi architectures for the restricted boltzmann machine,” *J. Emerg. Technol. Comput. Syst.*, vol. 13, no. 3, pp. 35:1–35:19, May 2017. [Online]. Available: http://doi.acm.org/10.1145/3007193

[189] D. Suzuki, M. Natsui, A. Mochizuki, S. Miura, H. Honjo, H. Sato, S. Fukami, S. Ikeda, T. Endoh, H. Ohno, and T. Hanyu, “Fabrication of a 3000-6-input-luts embedded and block-level power-gated nonvolatile fpga chip using p-mtj-based logic-in-memory structure,” in *2015 Symposium on VLSI Technology (VLSI Technology)*, June 2015, pp. C172–C173.

[190] 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.

[191] 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.

[192] G. E. Hinton, T. J. Sejnowski, and D. H. Ackley, *Boltzmann machines: Constraint satisfaction networks that learn*. Carnegie-Mellon University, Department of Computer Science Pittsburgh, PA, 1984.

[193] S. Bhatti, R. Sbiaa, A. Hirohata, H. Ohno, S. Fukami, and S. Piramanayagam, “Spintronics based random access memory: a review,” *Materials Today*, 2017.

[194] N. Locatelli, A. Mizrahi, A. Accioly, R. Matsumoto, A. Fukushima, H. Kubota, S. Yuasa, V. Cros, L. G. Pereira, D. Querlioz *et al.*, “Noise-enhanced synchronization of stochastic magnetic oscillators,” *Physical Review Applied*, vol. 2, no. 3, p. 034009, 2014.

[195] 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.

[196] A. Fukushima, T. Seki, K. Yakushiji, H. Kubota, H. Imamura, S. Yuasa, and K. Ando, “Spin dice: A scalable truly random number generator based on spintronics,” *Applied Physics Express*, vol. 7, no. 8, p. 083001, 2014.

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

[198] 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.

[199] C. M. Liyanagedera, A. Sengupta, A. Jaiswal, and K. Roy, “Magnetic tunnel junction enabled stochastic spiking neural networks: From non-telegraphic to telegraphic switching regime,” *arXiv preprint arXiv:1709.09247*, 2017.

[200] A. Mizrahi, T. Hirtzlin, A. Fukushima, H. Kubota, S. Yuasa, J. Grollier, and D. Querlioz, “Neural-like computing with populations of superparamagnetic basis functions,” *Nature communications*, vol. 9, no. 1, p. 1533, 2018.

[201] 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.

[202] M. Bapna and S. A. Majetich, "Current control of time-averaged magnetization in superparamagnetic tunnel junctions," *Applied Physics Letters*, vol. 111, no. 24, p. 243107, 2017.

[203] 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.

[204] 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 *Electron Devices Meeting (IEDM), 2009 IEEE International*. IEEE, 2009, pp. 1–4.

[205] 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.

[206] 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

[207] M. R. Pufall, W. H. Rippard, S. Kaka, S. E. Russek, T. J. Silva, J. Katine, and M. Carey, "Large-angle, gigahertz-rate random telegraph switching induced by spin-momentum transfer," *Phys. Rev. B*, vol. 69, p. 214409, Jun 2004. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.69.214409

[208] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *nature*, vol. 453, no. 7191, p. 80, 2008.

[209] M. Hu, H. Li, Q. Wu, and G. S. Rose, "Hardware realization of bsb recall function using memristor crossbar arrays," in *Proceedings of the 49th Annual Design Automation Conference*, ser. DAC '12. New York, NY, USA: ACM, 2012, pp. 498–503. [Online]. Available: http://doi.acm.org/10.1145/2228360.2228448

[210] K. Roy, A. Sengupta, and Y. Shim, "Perspective: Stochastic magnetic devices for cognitive computing," *Journal of Applied Physics*, vol. 123, no. 21, p. 210901, 2018.

[211] S. S. Parkin, C. Kaiser, A. Panchula, P. M. Rice, B. Hughes, M. Samant, and S.-H. Yang, "Giant tunnelling magnetoresistance at room temperature with mgo (100) tunnel barriers," *Nature materials*, vol. 3, no. 12, p. 862, 2004.

[212] W. Wang, H. Sukegawa, R. Shan, S. Mitani, and K. Inomata, "Giant tunneling magnetoresistance up to 330% at room temperature in sputter deposited co 2 feal/mgo/cofe magnetic tunnel junctions," *Applied Physics Letters*, vol. 95, no. 18, p. 182502, 2009.

[213] M. Wang, W. Cai, K. Cao, J. Zhou, J. Wrona, S. Peng, H. Yang, J. Wei, W. Kang, Y. Zhang *et al.*, "Current-induced magnetization switching in atom-thick tungsten engineered perpendicular magnetic tunnel junctions with large tunnel magnetoresistance," *Nature communications*, vol. 9, no. 1, p. 671, 2018.

[214] J. Scott, "High-dielectric constant thin films for dynamic random access memories (dram)," *Annual review of materials science*, vol. 28, no. 1, pp. 79–100, 1998.

[215] M. Stengel and N. A. Spaldin, "Origin of the dielectric dead layer in nanoscale capacitors," *Nature*, vol. 443, no. 7112, p. 679, 2006.