ENERGY-EFFICIENT SIGNAL CONVERSION AND IN-MEMORY COMPUTING USING EMERGING SPIN-BASED DEVICES

by

SOHEIL SALEHI MOBARAKEH
M.S. University of Central Florida, 2016
B.S. Isfahan University of Technology, 2014

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
2020

Major Professor: Ronald F. DeMara

ABSTRACT

New approaches are sought to maximize the signal sensing and reconstruction performance of Internet-of-Things (IoT) devices while reducing their dynamic and leakage energy consumption. Recently, Compressive Sensing (CS) has been proposed as a technique aimed at reducing the number of samples taken per frame to decrease energy, storage, and data transmission overheads. CS can be used to sample spectrally-sparse wide-band signals close to the information rate rather than the Nyquist rate, which can alleviate the high cost of hardware performing sampling in low-duty IoT applications. In my dissertation, I am focusing mainly on the adaptive signal acquisition and conversion circuits utilizing spin-based devices to achieve a highly-favorable range of accuracy, bandwidth, miniaturization, and energy trade-offs while co-designing the CS algorithms. The use of such approaches specifically targets new classes of Analog to Digital Converter (ADC) designs providing Sampling Rate (SR) and Quantization Resolution (QR) adapted during the acquisition by a cross-layer strategy considering both signal and hardware-specific constraints. Extending CS and Non-uniform CS (NCS) methods using emerging devices is highly desirable. Among promising devices, the 2014 ITRS Magnetism Roadmap identifies nanomagnetic devices as capable post-CMOS candidates, of which Magnetic Tunnel Junctions (MTJs) are reaching broader commercialization. Thus, my doctoral research topic is well-motivated by the established aims of academia and industry. Furthermore, the benefits of alternatives to von-Neumann architectures are sought for emerging applications such as IoT and hardware-aware intelligent edge devices, as well as the application of spintronics for neuromorphic processing. Thus, in my doctoral research, I have also focused on realizing post-fabrication adaptation, which is ubiquitous in post-Moore approaches, as well as mission-critical, IoT, and neuromorphic applications.

Dedicated to my wonderful, caring, and supportive family. To my mother, Manijeh, to my father, Mehrdad, and to my brother, Sina, for believing in me and giving me their endless support so that I can achieve my goals and complete this incredible milestone.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Dr. Ronald F. DeMara who provided me with a unique opportunity of being part of the Computer Architecture Lab family and research group. He has kindly supported me all along and his insightful comments helped me conduct high-quality research, leading to several publications in prestigious journals and conference proceedings. Additionally, I would like to thank my committee members Dr. Amro Awad, Dr. Deliang Fan, Dr. Nazanin Rahnavard, and Dr. Annie Wu for supporting me and my research.

Furthremore, 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, and by NSF through ECCS-1810256.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION AND MOTIVATION[1]

Internet of Things (IoT) utilizes autonomous and trustworthy ambient-powered devices with small area footprints, which provide intermittent operations and low-power data acquisition and processing capabilities while maintaining a low maintenance cost. In particular, achieving low-power, high-reliability, and high-performance data acquisition and processing is of utmost importance within IoT applications due to the limited energy budget and challenges caused by the device scaling [8, 23, 24].

Furthremore, Compressive Sensing (CS) methods aim to maximize signal sensing and reconstruction performance while reducing energy consumption for IoT applications. The goal of CS methods are to reduce the number of samples per frame in order to decrease energy consumption, storage requirements, and data transmission overheads. However, CS techniques are mostly designed and implemented oblivious to specifics and limitations of hardware platform that performs the sampling operation. Moreover, in the rate and resolution trade-off within CS approaches, both signal dependent constraints, such as sparsity rate and noise levels, and hardware dependent constraints, such as energy budget, bandwidth, and battery capacity, play an important role. Moreover, in any practical scenario, sensing operations are required to maximize the sensing performance by satisfying energy and bandwidth constraints. Additionally, in real-world applications, signals may contain a Region of Interest (RoI) and uniform sampling will not be efficient given the fact that signal's sparsity may be non-uniform. Thus, cornerstone to achieving high-accuracy and efficient CS is utilization of an adaptive measurement matrix that changes according to the signal characteristics that are extracted from the observations of the signal components in the previous time frames.

[1] ©IEEE. Part of this chapter is reprinted, with permission, from [1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 13, 22]

On the hardware side, Von-Neumann architectures have been facing challenges such as increased static energy consumption, large access latencies, and, limited scalability. Recently, researchers have focused on in-memory computing paradigms by utilizing non-volatile spin-based devices to realize non-Von-Neumann architectures. Additionally, there is an increasing demand for energy- and area-efficient Analog to Digital Converters (ADCs) in IoT applications, especially for applications such as image processing where each pixel sensor requires a compact ADC. However, integrating CMOS ADCs in sensor nodes is challenging due to the large area of analog CMOS-based circuits. Additionally, increased static energy consumption and increased reliability challenges due to high process variations in scaled technology nodes, exacerbates these challenges.

Thus, there is a need for low-complexity, ultra-low-power circuits for signal conversion for IoT applications. Furthermore, there is a demand for CS solutions that consider hardware constraints and signal constraints for IoT intelligent sampling and edge processing. As shown in Figure 1.1, the primary motivation of my research is to find an answer to the question of "How can spin-based devices assist Compressive Sensing within IoT Applications?"

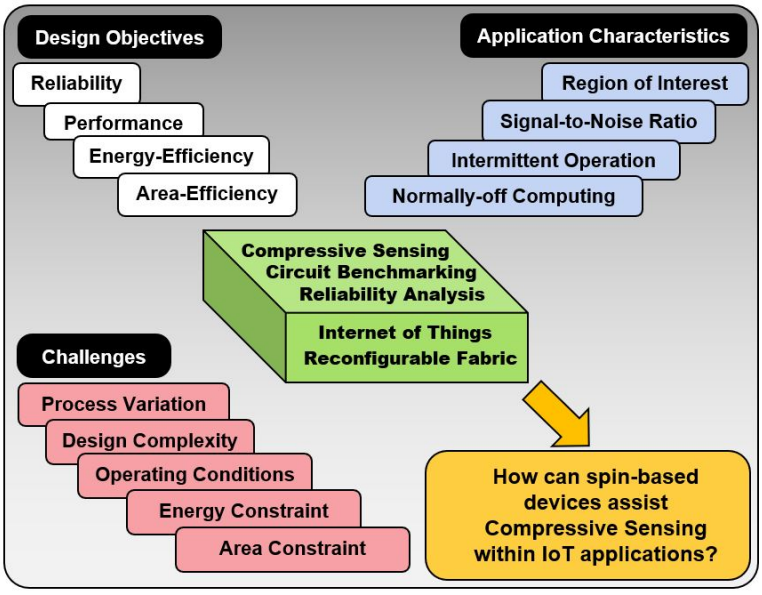


Figure 1.1: Research Motivation.

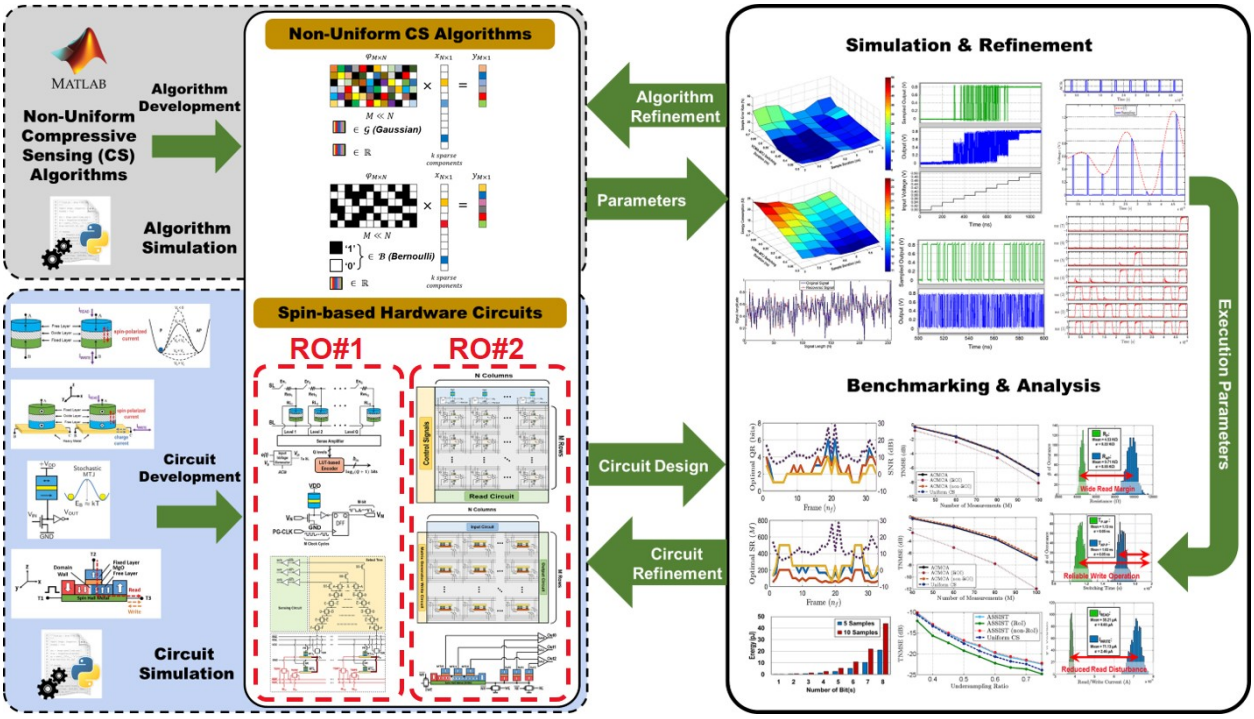


Figure 1.2: Research Objectives (ROs) context diagram.

Herein, the aforementioned challenges have been studies and the following Research Questions (RQs) are designed to address them. Furthermore, for each of the RQs, Research Objectives (ROs) are defined to provide answers to the RQs. Figure 1.2 illustrates the context diagram of the research objectives that are addressed within this document.

- RQ#1: How can spin-based devices advance compressive sensing?
  - RO#1: Integrate adaptive Compressive Sensing techniques with beyond-CMOS hardware to minimize overall cost of data acquisition and processing
- RQ#2: How can spin-based devices advance beyond Von-Neumann architectures with intelligent signal conversion and processing?
  - RO#2: Devise a framework for signal acquisition, conversion, and edge processing in IoT where energy and area are significantly constrained

1.1 Need for Reliable and Energy-Efficient Nanoscale Computing Architectures

As technology scales down along with increased demands of greater on-chip integration for larger memory capacities, researchers and designers have responded to the resulting fabrication and operational challenges by embracing new device technologies along with new memory cell designs which leverage their unique advantages. A collection of innovative methods has been developed to increase their reliability and performance.

In addition to addressing scalability to technologies beyond 10nm where traditional memory elements such as Static Random Access Memory (SRAM) and Dynamic Random Access Memory (DRAM) face significant scaling challenges [25, 26], innovations to mitigate the power wall and reduce leakage power consumption occupy the forefront of on-chip memory design considerations [24, 27]. Power consumed by memory elements can become a significant portion of total power in active modes whereby the processing cores [28, 29] rely on these memory arrays that are significant contributors to standby mode power consumption. These concerns motivate the research into balancing energy and reliability effectively.

To attain these goals and deliver the necessary operational characteristics, emerging memory devices such as Resistive RAM (RRAM), Phase Change Memory (PCM), and Magnetic RAM (MRAM) offer several potential advantages. Among promising devices, the 2014 ITRS Magnetism Roadmap identifies nanomagnetic devices such as Spin-Transfer Torque Magnetic Random Access Memory (STT-MRAM), as capable post-CMOS candidates, of which Nano Magnetic Logic (NML), Domain Wall Motion (DWM), Spin-Transfer Torque (STT)-MTJ/Spin Hall Effect (SHE)-MTJ [30, 31, 32] whose attributes are depicted in Figure 1.3. Additionally, emerging devices such as Quantum Cellular Automata (QCA) [33, 34, 35] logic designs have demonstrated performance improvements. However, herein, we focus on the MTJ devices due to their commercial availability. STT-MRAM can offer low read access time, near-zero standby power consumption, and small

area requirement. STT-MRAM also offers integration with backend CMOS processes. To embrace their adoption in anticipated applications, a palette of cooperating reliability techniques is identified and compared at the bit-cell level.

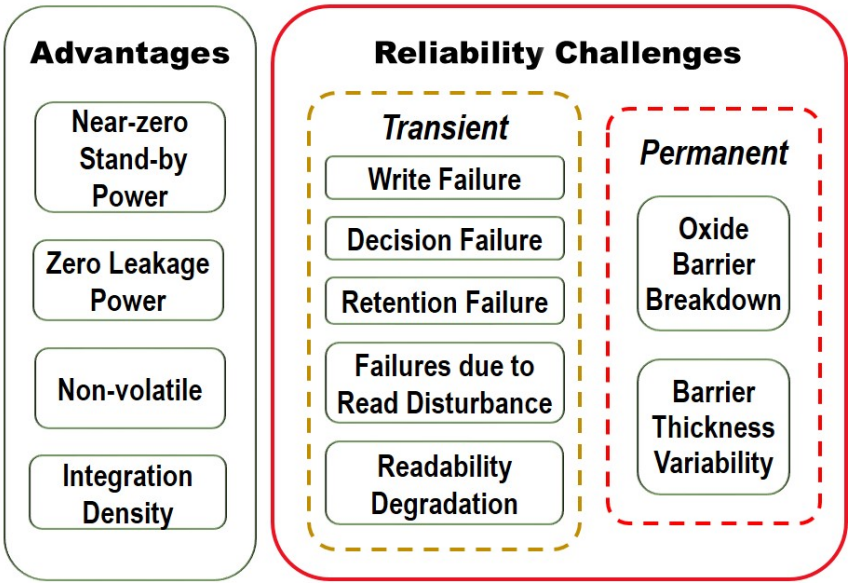


Figure 1.3: Advantages and reliability challenges of STT-MRAM. [1]

1.1.1 Need for Reliable and Energy-Efficient Sense Amplifiers

High reliability and energy efficient switching of STT-based devices is highly sought and active area of emerging device research. Thus, we examine improvements to the Sense Amplifier (SA) design which can achieve both of these objectives on a continuum of energy versus reliability trade-offs. Due to increase in process variation as technology shrinks, STT memory cell reliability has become a significant concern in high density memory arrays and cache designs. While a collection of innovative methods have been previously proposed to increase reliability and performance, each incurs costs and challenges which may lead to a sub-optimal performance profile. Of particular urgency is the need to reduce the effects of device mismatch and variation due to scaling of the devices, especially with respect to the use of different SAs.

1.1.2 Reliable and High-Performance Last Level Cache Design

Complimentary Metal Oxide Semiconductor (CMOS) device scaling continues to increase the need to identify viable approaches for reducing leakage power. An alternative to CMOS-based memory devices is offered by emerging technology memory devices that contribute inherent features of Non-Volatile Memory (NVM) capabilities. With attributes of non-volatility, near-zero standby energy, and high density, Spin Transfer Torque Magnetic RAM (STT-MRAM) has emerged as a promising alternative post-CMOS technology for embedded memory applications. In order to practically implement these NVMs, various techniques to mitigate the specific reliability challenges associated with STT-MRAM elements are surveyed, classified, and assessed in [1]. In [1], we identified various solutions to the reliability issues within a taxonomy of current and future approaches to reliable STT-MRAM designs.

Despite the range of approaches available to mitigate Process Variation (PV), it remains as one of the most negatively influential factors impacting STT-MRAM technology performance from the perspectives of delay and energy consumption [1]. Furthermore, the Sense Margin (SM), which is an important parameter of the tolerance in sensing the resistive state of emerging NVM devices, varies considerably in the presence of PV of the devices which comprise the bit-cell and their associated sensing circuits [1]. SM is also known as the difference between bit-line voltage and reference voltage. These variations may then result in erroneous data sensing operations, read disturbance, readability degradation at scaled technology nodes, and retention failure [1]. These reliability issues have increased the demand for designing advanced low-power approaches with reliable sensing circuits to mitigate and leverage PV for improved performance and reliability of NVMs, including increasing the SM and finding the optimum read current and latency [1].

Using NVM can increase energy efficiency via a significant reduction in leakage energy. However, effects of Process Variation (PV) still put a limit on the scalability and applicability of NVM

devices. PV mainly impacts the CMOS peripheral circuits [24] and emerging technology NVM elements such as Magnetic Tunnel Junction (MTJ) devices [10]. Effects of PV on MTJ devices manifests itself as variation in oxide thickness and MTJ geometry, which in turn results in deviations of MTJ resistance and severe fluctuations of the *Sense Margin (SM)*, resulting in possible false detection scenarios and increased bit error rates [36, 37]. Furthermore, PV negatively impacts the performance consistency of memory operation, since the threshold voltage, V_{th}, and gate length, L_{eff}, of CMOS peripheral circuit fluctuates in presence of PV, which result in read and write delays, driving current variations, and increased energy consumption [38]. Additionally, a survey of reliability challenges and mitigation techniques for emerging NVM elements is presented in [1]. As a result of PV impacts and the performance limitation it dictates on NVMs, there exists an increased demand for advanced reliable and energy-efficient read and write circuits, which can be integrated into PV-resilient system architectures to provide high performance NVMs with reliable read and write operations.

The work herein proposes a novel approach for read and write operations of emerging NVMs used as Last Level Cache (LLC). One of the main focuses of this dissertation is on increasing the energy-efficiency and reliability of the read operation in STT-/SHE-MRAM and is motivated by the observation that in the PARSEC suite [39] using STT-/SHE-MRAM-based LLC and in the presence of PV, approximately 27.5% of the sensed data has the potential to be read incorrectly. However, out of 27.5%, roughly 21% of the incorrectly sensed data requires to be handled since up to 6% of the incorrectly sensed data on average will be overwritten prior to being used by the processor or to be committed to the main memory. Additionally, such a significant percentage of incorrect data sensing requires close attention before they cause wrong outputs, application crashes, or prolonged program executions [29].

In order to improve the reliability of the read operation based on the aforementioned observations, we propose a circuit-architecture cross-layer solution suitable for multi-core processors as well

as Internet of Things (IoT) devices. Our proposed technique, referred to as *Self-Organized Sub-bank (SOS)*, partitions STT-/SHE-MRAM data arrays into several sub-banks to directly access the requested data while introducing individualized sensing resolution. In our proposed approach, two Sense Amplifiers (SAs) are assigned to sub-banks, one energy-efficient SA and one high-resilient SA. Initially during an evaluation phase each sub-banks is evaluated using a *Power-On Self-Test (POST)* and then a preferred SA will be assigned to each sub-bank based on the results of the POST. Our results indicate that SOS increases reliability of read operations, which in turn reduces fault propagation, as well as reducing the risk of contaminating the application's data structure.

1.1.3 Need for Reliable and Energy-Efficient Write Circuits

Another focus of this dissertation is on increasing energy efficiency and reliability of write operations in STT-MRAM and is motivated by the observation that the STT switching technique suffers from high dynamic energy consumption [40]. SHE-MTJ has been recently studied as an energy-efficient alternative for STT-MTJ due to its improved performance. Several write circuits have been studied in recent years in order to achieve optimum energy while maintaining high reliability. Herein, we explore SHE-MTJ write circuits and compared those with conventional STT-MTJ write circuits in terms of performance and reliability. Furthermore, a high-resilient write circuit as well as an energy-efficient write circuit are selected in order to be utilized in the SOS approach for further performance and reliability improvements of SHE-MRAM. In particular, the SOS approach is implemented once with the high-resilient write circuit and once with the energy-efficient write circuit.

Our results indicate that the energy-efficient write circuit provides significant energy and delay improvements over the conventional STT-MTJ write circuit and high-resilient SHE-MTJ write circuit. On the other hand, the high-resilient write circuit for SHE-MTJ offers reliability improvement

over the energy-efficient SHE-MTJ write circuit.

1.1.4 Need for Reliable and Energy-Efficient Non-Volatile SRAM

Spin-based devices have been extensively researched as promising companions to Complimentary Metal on Oxide Semiconductor (CMOS) devices. As CMOS scaling trends continue, the need to identify viable approaches for reducing leakage power increases [26]. Especially, the increase of leakage power in normally-off computing applications and sleep power critical systems, such as mobile System-on-Chips and Internet of Things (IoT), has become a major challenge [17, 41, 42, 43, 44, 45, 46, 47, 48, 49].

Several approaches, such as Dynamic Voltage Scaling (DVS) [50], multiple threshold voltage levels [51], and Power Gating (PG) [52], have been utilized to address the high-leakage power dissipation of highly-scaled CMOS devices. Conventional PG is one of the most widely-used approaches where parts of the system will be turned off when there are no activities taking place in those parts. In other words, when parts of the system are going into stand-by mode, the power supply to those parts will be cut-off in order to reduce the power dissipation. A commonly-used method regarding PG approaches for normally-off computing applications and sleep power critical systems, is utilizing two-macro architecture. In two-macro architectures, a volatile memory, such as Static Random Access Memory (SRAM), is accompanied by a Non-Volatile Memory (NVM), such as FLASH, and whenever the SRAM device is going to stand-by mode, the data will be stored in the NVM and then stand-by mode will be activated via PG approach [17, 42]. Consequently, when the volatile memory returns to its normal operation, the data stored in NVM will be restored to the volatile memory and then normal operation will be resumed.

Despite the advantages that the two-macro architecture offers, such as reduced NVM accesses and considering endurance criteria for NVM devices, one of the major drawbacks of the two-macro

approach is the data transfer between the volatile memory and NVM, which induces a significant data restoration delay and dynamic power dissipation overheads. Additionally, in case of a power failure, long restoration delay might incur data loss and reliability challenges.

Recently, researchers have proposed a one-macro architecture, where the NVM device is integrated within each volatile memory cell [17, 41, 42, 43, 44, 45, 46, 47, 48, 49]. This will enable fast and more energy-efficient back-up and restore operations. Recently, a significant amount of research is done on integration of SRAM-based volatile memory with emerging NVM devices, such as Resistive RAM (RRAM), Phase Change Memory (PCM), Spin Transfer Torque Magnetic RAM (STT-MRAM), and Spin-Hall Effect Magnetic RAM (SHE-MRAM) [17, 41, 42, 43, 44, 45, 46, 47, 48, 49]. SRAM devices offer a compact on-chip storage that utilizes a low minimum supply voltage and provides fast read and write operations. However, the volatile nature and high leakage power dissipation of SRAM devices have become a major challenge. On the other hand, the emerging NVMs offer zero leakage power dissipation, which can increase energy efficiency via a significant reduction in static power consumption and leakage energy consumption [2, 15, 53]. This has provided the opportunity to integrate the two types of memories in an approach which is widely known as Non-Volatile SRAM (NV-SRAM).

SRAM devices have very fast read/write access times and relatively low dynamic energy requirements, but their disadvantages include volatility and substantial static operating current draws. SHE-MRAM devices are non-volatile and have the potential for very low static energy requirements, but their write energy demands exceed that of SRAM. We seek to leverage the complementary nature of these two technologies by integrating them into parallel SRAM/SHE-MRAM bit-cell devices.

1.2 Transition from Memory to In-Memory Computing

1.2.1 Need for Reliable and Energy-Efficient Look-Up Tables for In-Memory Computing

Flexibility and runtime adaptability are two of the main motivations for the wide adoption of reconfigurable fabrics. Among the most commonly used reconfigurable fabrics, Field Programmable Gate Arrays (FPGA) have been the primary focus due to their flexibility that allows realization of logic elements at medium and fine granularities while incurring low non-recurring engineering costs and rapid deployment to market. Additionally, FPGAs have been researched as promising platform that can be utilized effectively to increase reliability in case of process-voltage-temperature variation [54]. The main challenge of static random access memory (SRAM)-based FPGAs is their increased area and power consumption to achieve flexible design. The main components of FPGAs are Look-Up Tables (LUTs) and switch boxes that are mainly consisted of SRAM cells [55]. However, SRAM-based LUTs incur limitations such as high static power, volatility, and low logic density.

Innovations using emerging devices within FPGAs have been sought to bridge the gaps needed to overcome the limitations of SRAM-based FPGAs. High-endurance non-volatile spin-based LUTs have been studied in the literature as promising alternatives to SRAM-based LUTs, Flash-based LUTs, and other state-of-the-art emerging LUTs such as resistive random access memory (RRAM)-based LUTs and phase change memory (PCM)-based LUTs [56, 57, 58, 59, 60, 61]. Spin-based devices offer non-volatility, near-zero static power, high endurance, and high integration density [1, 62]. The spin-based LUTs presented in the literature [56, 57, 58, 59, 60, 61] require separate read and write operations as well as a clock, which makes these LUTs a suitable candidate for sequential logic operations. However, the main challenge that has not been addressed in the literature is providing a spin-based LUT design for combinational logic operation without the need

for a clock. Additionally, proposed spin-based LUTs proposed in the literature fail to maintain a wide sense margin and high reliability without incurring significant area and power dissipation overheads [56, 57, 58, 59, 60, 61].

Herein, in order to address the aforementioned challenges, we develop a clockless 6-input fracturable non-volatile Combinational LUT (C-LUT) with wide read margin using spin Hall effect (SHE)-based Magnetic Tunnel Junction (MTJ) and provide a detailed comparison between the SHE-MRAM and Spin Transfer Torque (STT)-MRAM C-LUTs. Additionally, we provide detailed analysis on the reliability of our proposed C-LUT in the presence of Process Variation (PV).

1.2.2 Need for Reliable and Energy-Efficient Analog to Digital Converters

Adaptive signal acquisition and conversion circuits using emerging spin-based devices offer a new and highly-favorable range of accuracy, bandwidth, miniaturization, and energy trade-offs. The use of such approaches specifically targets new classes of Analog to Digital Converter (ADC) designs providing *sampling rate* (SR) and *quantization resolution* (QR) adapted during acquisition by a cross-layer strategy considering both signal and hardware specific constraints.

Prior works on adaptive rate and resolution ADCs [63, 64, 65, 66, 67, 68, 69, 70] have optimized the rate/resolution trade-off assuming a low-pass signal model and utilizing Complementary Metal Oxide Semiconductor (CMOS) technology. However, in this work, we use the theory of Compressive Sensing (CS) [71, 72] and spin-based devices [1, 10, 30, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83] to advance beyond these limitations. Compressive sensing is a modern signal acquisition paradigm that aims to measure sparse signals close to their *information rate* rather than their *Nyquist rate*. This is specifically critical for *spectrally sparse wide-band signals* in which conventional sampling becomes impractical due to challenges associated with building sampling hardware that operates at prohibitively high Nyquist rates.

Quantized CS [84, 85, 86, 87] aims at addressing the existing trade-off between the number of measurements and the number of bits used to quantize each measurement for a fixed bit budget. Although this trade-off has been studied previously [84, 85, 86, 87], adaptive optimization of the SR and QR during signal acquisition has not been investigated. Moreover, despite the fundamental theoretical discoveries in this field, quantized CS techniques are mostly designed and implemented *oblivious* to the specifics and limitations of the hardware platform that performs the sampling/acquisition. In other words, in the rate/resolution trade-off, both signal dependent constraints (e.g., sparsity and noise level) and hardware dependent constraints (e.g., energy, bandwidth, and battery capacity) play an important role and as these constraints vary during signal acquisition, dynamic and cross-layer optimization of SR and QR is desirable for efficient signal acquisition. While adaptive SR and QR seem to be viable approaches from the signal processing and algorithmic point of view, the actual implementation of them requires a hardware platform that can adapt itself to these variations.

1.2.3 Need for Analog to Digital Converters with In-Memory Computing Capabilities

Spin-based devices have been extensively researched as promising companions to CMOS devices. As CMOS scaling trends continue, the need to identify viable approaches for reducing leakage power increases [26]. With attributes of non-volatility, near-zero standby energy, and high density, the Magnetic Tunnel Junction (MTJ) has emerged as a promising alternative post-CMOS technology for embedded memory and logic applications [1, 10, 30]. Recent studies have shown that conventional Von-Neumann computing architectures, in which the storing elements are distinct from computing elements, incur challenges created by interconnection and busing demands [88]. These challenges include, but are not limited to, increased static energy consumption, large access latencies, and limited scalability. Recent studies have offered in-memory computing paradigms as a potential solution to these challenges. Use of non-volatile memory devices such as spin-based

devices have enabled researchers to design non-Von-Neumann architectures, where processing and memory are integrated [88, 89].

Furthermore, there is an increasing demand for energy and area efficient Analog to Digital Converters (ADCs) as the need for integrating the signal acquisition and processing as well as rapid parallel data conversion in sensor nodes has increased [90, 91, 92]. One of the main challenges of designing such sensors in CMOS is integrating ADCs in each sensor due to the large area of analog circuits, especially in applications such as image processing where each pixel sensor requires a compact ADC [93, 94]. Moreover, another main challenge is the increased static energy consumption due to transistor scaling. Additionally, decreased reliability caused by high process variation can become another major challenge in scaled technology nodes [95].

1.2.4 Need for MRAM Stochastic Oscillators for Adaptive Sampling of Sparse IoT Signals

Recently, non-uniform sampling approaches such as Compressive Sensing (CS) have been proposed to reduce the energy consumption of sampling operation by reducing number of samples in each frame, reduce required storage to save the sampled data, and reduce the data transmission due to lower number of samples taken [84, 91, 96]. Additionally, event-driven sampling, such as level-crossing sampling, has been widely adopted as a promising CS technique to maximize the performance of sampling operation while reducing energy consumption [65]. Furthermore, CS techniques are utilized to sample spectrally sparse wide-band signals close to their information rate rather than their Nyquist rate, which can be a challenge using conventional uniform sampling techniques due to the high cost of the hardware that is capable of performing the sampling operation at a high Nyquist rate.

Despite all the benefits that CS techniques offer, they are typically realized oblivious to the hardware limitations such as energy, bandwidth, and battery capacity. Additionally, signal-dependent

constraints such as sparsity and noise level are ignored while studying the quantization rate and resolution trade-off. The aforementioned hardware-dependent and signal-dependent constraints alter during the sampling operation. Thus, an adaptive quantization rate and resolution optimization circuitry is required to maximize sampling performance while minimizing the number of samples to reduce energy consumption, data transmission, and storage. Adaptive quantization rate and resolution sampling might be readily achieved from the algorithm perspective, however it requires a hardware platform that is capable of real-time adaptation according to certain signal behavior such as sparsity rate. Recently, an adaptive optimization of the quantization rate and resolution during signal acquisition has been investigated in [5].

Previous works on adaptive quantization rate and resolution ADCs have been implemented using Complementary Metal Oxide Semiconductor (CMOS) technology and considering a low-pass signal model [65, 97]. Herein, we propose a spin-based Adaptive quantization rate (AQR) generator circuit that considers the signal dependent constraint as well as hardware limitations. The proposed AQR generator circuit utilized Magnetic Random Access Memory (MRAM)-based stochastic oscillator devices, which offer miniaturization and significant energy savings [7].

Researchers have recently expanded their efforts to maximize the signal sensing and reconstruction performance while reducing energy consumption for Internet of Things (IoT) applications such as sensors and mobile devices [5, 13]. Recently, Compressive Sensing (CS) has been proposed as a sampling technique aimed at reducing the number of samples taken per frame to decrease energy, storage, and data transmission overheads. CS can be used to sample spectrally-sparse wide-band signals close to the information rate rather than the Nyquist rate, which can alleviate the high cost of hardware performing sampling at high Nyquist rates [84, 91, 94, 98].

Implementing non-uniform CS in hardware requires a random number generator (RNG) since CS theory assumes random sampling of data [98]. RNGs can be divided into two classes: true RNGs

(TRNGs) and pseudo-RNGs (PRNGs). PRNGs include Linear Feedback Shift Registers (LFSR), which begin with a seed value and then continuously update this value by means of a linear function in order to create the illusion of randomness; such designs can suffer from limited quality in the randomness of the output as well as high energy and area [99]. TRNGs, on the other hand, rely on truly random events such as thermal noise, oscillator jitter, and metastability; TRNG designs can be challenged by limited generation speed as well as post-processing requirements which impose area and power overheads [100].

Previous attempts at TRNG design using spintronics have included use of bistable superparamagnetic tunnel junctions [99], application of sub-threshold voltages for stochastic switching in magnetic tunnel junctions (MTJs) [100, 101], use of MTJ stack arrangements for precessional switching [102], and by means of the voltage-controlled magnetic anisotropy (VCMA) effect [103]. While these designs have been effective in their quality of randomness, they have also involved relatively complex hardware resulting in power and area overhead. Thus, a spin-based TRNG is sought to minimize the power dissipation and area. Furthermore, previous works on non-uniform compressive sensing have been implemented using Complementary Metal Oxide Semiconductor (CMOS) technology [65, 97].

Herein, we propose a spin-based non-uniform compressive sensing circuit-algorithm solution that considers the signal-dependent constraint as well as hardware limitations called *Adaptive Sampling of Sparse IoT signals via STochastic-oscillators (ASSIST)*. The proposed ASSIST approach utilizes Magnetic Random Access Memory (MRAM)-based Stochastic Oscillator (MSO) devices as the main element in TRNGs, which offer miniaturization and significant energy savings [7, 13]. Additionally, MRAM-based Non-Volatile Memory (NVM) is used to store the output of TRNGs, which are the elements of the CS measurement matrix.

CHAPTER 2: BACKGROUND AND RELATED WORK[1]

Recent approaches to continue the trends associated with Moore's Law have focused on beyond-CMOS devices to supplement conventional computing methods. For instance, hardware integration and realization of highly-efficient Compressive Sensing (CS) methods have inspired novel circuit and architectural-level approaches [5]. The challenge is to design more optimal device-level approaches for IoT applications wherein lifetime energy, device area, and manufacturing costs are highly-constrained. Herein, we have developed a novel adaptive hardware-based approach for NCS of sparse IoT signals.

2.1 Overview of MTJ-based Non-Volatile Memory (NVM) Operation

The basic concept of spin-based Non-Volatile Memory (NVM) devices is to control the intrinsic spin of electrons in a ferromagnetic thin film based solid-state nano-device. Magnetic Tunnel Junction (MTJ) devices are constructed with layered pillars of ferromagnetic and insulating layers to leverage magnetic orientations that can be controlled and sensed in terms of electrical signal levels. The non-volatile MTJ consists of two ferromagnetic (FM) layers , which are called the fixed layer and the free layer, and one tunneling oxide layer between the two FM layers [1]. FM layers could be aligned in two different magnetization configurations, Parallel (**P**) and Anti-Parallel (**AP**). Accordingly, the MTJ exhibits a low resistance ($\mathbf{R}_P$) or high resistance ($\mathbf{R}_{AP}$), respectively [1, 22].

Four switching schemes have been used by researchers in order to write into MTJ cells [104]. The four switching schemes are: Field Induced Magnetization Switching (FIMS), Thermally Assisted

[1] ©IEEE. Part of this chapter is reprinted, with permission, from [1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 13, 22]

Switching (TAS), Current Induced Magnetic Switching (CIMS), which is also called Spin Torque Transfer (STT), and Spin-Hall Assisted STT (SHA-STT). Herein, we have focused on the STT and SHA-STT switching approaches.The **P** or **AP** state of the MTJ is configured by means of the bidirectional current that passes through it, $\mathbf{I}_{MTJ}$, which could readily be produced by simple MOS based circuits. The states of the MTJ are switched when $\mathbf{I}_{MTJ}$ becomes higher than a critical current, $\mathbf{I}_C$. The MTJ resistance in **P** ($\theta = 0°$), and **AP** ($\theta = 180°$) states is expressed by the following equations:

$$\begin{aligned} R(\theta) &= 2R_{MTJ} \times \frac{1 + TMR(V_b)}{2 + TMR(V_b) + TMR(V_b).\cos(\theta)} \\ &= \begin{cases} R_P = R_{MTJ} & ,\theta = 0° \\ R_{AP} = R_{MTJ}(1 + TMR) & ,\theta = 180° \end{cases} , \end{aligned} \tag{2.1}$$

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

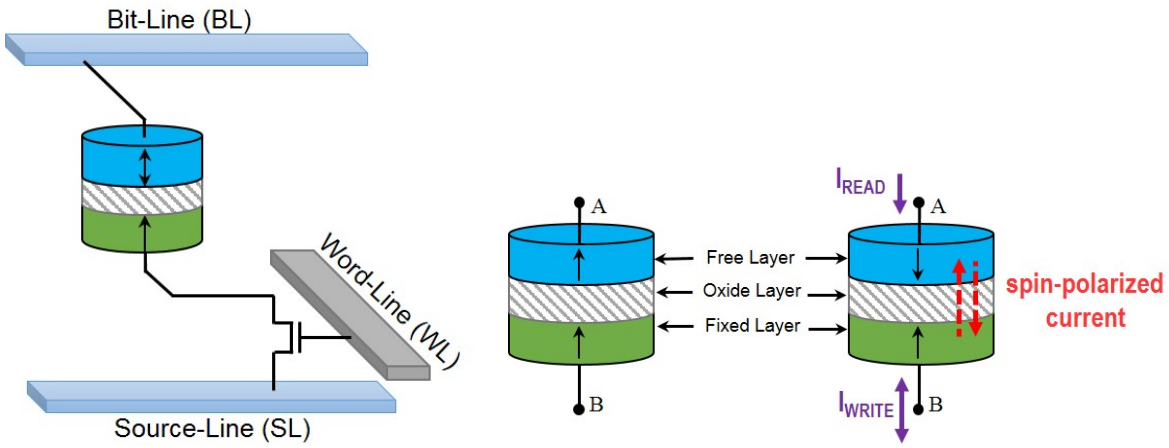
$$TMR(V_b) = \frac{TMR(0)}{(1 + (\frac{V_b}{V_h})^2)}, \tag{2.3}$$

where V_b is the bias voltage, $V_h = 0.5V$ is the bias voltage when Tunnel Magneto-Resistance (TMR) ratio is half of the $TMR0$, t_{ox} is the oxide thickness of MTJ, Factor is obtained from the resistance-area product value of the MTJ that relies on the material composition of its layers, Area is the surface area of the MTJ, and ϕ is the oxide layer energy barrier height [15].

2.1.1 STT Switching Approach

The STT switching method is based on applying a spin polarized current through the MTJ junction which will cause the magnetization of the free layer to change if the current magnitude passes a

certain value known as critical current. As illustrated in Figure 2.1(b), STT-MRAM utilizes an MTJ device as the storage element. Although STT-MRAM provides a high write endurance, the advent of long write latency and high energy consumption exacerbate the energy and reliability implications of STT-MRAM. STT switching is one of the most promising alternatives for data storage since it doesn't require external wires and magnetic fields, requires lower current density for switching operation, and consumes less power compared to other methods introduced herein [104]. Figure 2.1(a) shows an STT-MRAM cell, which has an access transistor. This STT-MRAM cell structure is known as "one-transistor-one-MTJ (1T-1R)" structure [1].



(a) (b)

Figure 2.1: (a) 1T-1R STT-MRAM cell structure, (b) Right: Anti-parallel (high resistance), Left: Parallel (low resistance) [2].

Based on the STT approach, a bidirectional spin-polarized current ($\mathbf{I}_{MTJ}$) is required for switching the MTJ nanomagnet configuration, which can be readily generated through simple MOS-based circuits. STT switching behavior can be categorized into precessional region ($\mathbf{I}_{MTJ} > \mathbf{I}_C$), and thermal activation region ($\mathbf{I}_{MTJ} < \mathbf{I}_C$). To achieve higher switching speed, STT-MRAM should operate in the precessional region, which is described by the Sun model [105] as shown below:

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

where τ_{STT} is the mean duration for precessional switching region, $C = 0.577$ is the Euler's constant, $\Delta = \frac{\mathrm{E}}{4k_BT}$ is the thermal stability factor, and *m* is the free layer magnetic moment. While the STT approach offers significant advantages in terms of read energy and speed, a significant incubation delay due to the pre-switching oscillation [106] incurs high switching energy. Recently, the 3-terminal Spin-Hall Effect MTJ (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 [107], as shown in Figure 2.2(b).

2.1.2 SHA-STT Switching Approach

Despite all of the merits that STT-MRAM offers, violation of reliability tolerances may result in read and/or write failures [1]. Thermal fluctuations and other issues such as MTJ PV and the CMOS peripheral circuit PV have severely limited the scalability of STT-MRAM devices [1]. Also, as a result of these issues, there is an increased demand for advanced sensing circuits that can provide an adequate Sense Margin (SM) along with low power operation.

Due to the large incubation delay of write operation in the STT approach, which makes STT-MRAM not a perfect candidate for LLC, Spin Hall Assisted STT (SHA-STT) is recently proposed [21, 108, 109]. Since SHA-STT reduces the incubation delay and due to the fact that SHE-MTJ offers separate read and write paths, SHE-MRAM provides a faster and more reliable write operation compared to STT-MRAM. Furthermore, there is no need for an external magnetic field in order to switch the magnetization direction of the free layer. As shown in Figure 2.2(b), in a SHE-MRAM Cell a Heavy Metal (HM) stripe is placed next to the free layer. In order to write into the MTJ using SHA-STT, a charge current should be applied between write terminals B and C, as shown in Figure 2.2(b), which will produce SHE. Due to the SHE, pure spin current will be produce in an upward or downward direction perpendicular to the charge current in the HM, which

determines the magnetization orientation of the free layer of the MTJ. Reading the SHE-MTJ uses the same operation as STT-MTJ. Figure 2.2(a) depicts a cell structure for SHE-MRAM bit-cell. This SHE-MRAM cell structure is known as "two-transistor-one-MTJ (2T-1R)" structure [109].

In [107], ratio of the injected spin current to the applied charge current, called Spin Hall Injection Efficiency (SHIE), is defined as shown below:

$$\text{SHIE} = \frac{I_{sz}}{I_{cx}} = \frac{\pi.MTJ_{width}}{4HM_{thick}}\theta_{SHE}\left[1 - sech\left(\frac{HM_{thick}}{\lambda_{sf}}\right)\right], \tag{2.5}$$

where MTJ_{width} is the width of the MTJ, HM_{thick} is the thickness of the HM, λ_{sf} is the spin flip length of the HM, and θ_{SHE} is the SHE angle. This equation is valid for SHE-MTJ devices in which the length of the MTJ equals the width of the HM. The critical spin current required for switching the free layer magnetization orientation is expressed by (2.6) [110]:

$$I_{S,critical} = 2q\alpha M_S V_{MTJ}\left(H_k + 2\pi M_S\right)/\overline{h}, \tag{2.6}$$

where V_{MTJ} is the MTJ free layer volume. Thus, the SHE-MTJ critical charge current can be calculated using (2.5) and (2.6). The relation between SHE-MTJ switching time and the voltage applied to the HM terminals is shown in (2.7), in which the Critical Voltage (v_c) is given by (2.8) [107].

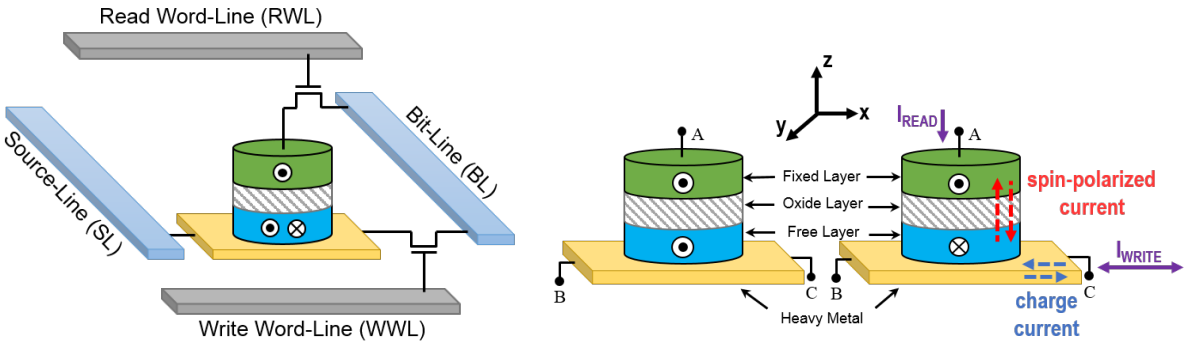
$$\tau_{SHE} = \frac{\tau_0 ln\left(\pi/2\theta_0\right)}{\left(\frac{v}{v_c}\right) - 1}, \tag{2.7}$$

$$v_c = 8\rho I_c\left\{\theta_{SHE}\left[1 - sech\left(\frac{HM_{thick}}{\lambda_{sf}}\right)\right]\pi HM_{length}\right\}^{-1}, \tag{2.8}$$

where, $\theta_0 = \sqrt{(k_B/2E_b)}$ is the effect of stochastic variation, E_b is the thermal barrier of the magnet of volume V, HM_{length} is the length of the HM, and I_C is the critical charge current for spin-torque induced switching. In order to model the SHE-MTJ, the HM resistance is also required, which is

expressed by (2.9), where ρ_{HM} is the electrical resistivity of HM.

$$R_{HM} = (\rho_{HM}.HM_{length}) / HM_{width} \times HM_{thick} \tag{2.9}$$



(a) (b)

Figure 2.2: (a) 2T-1R SHE-MRAM cell structure, (b) Right: Anti-parallel (high resistance), Left: Parallel (low resistance). A positive current along the $+x$ axis induces a spin injection current along the $+z$ axis. The injected spin current produces the required spin torque for aligning the magnetic direction of the free layer along the $+y$ axis, and vice versa [3].

We have utilized the approach proposed in [109] to model the behavior of SHE-MTJ and STT-MTJ devices, in which a Verilog-AMS model is developed using the aforementioned equations. Then, the model is leveraged in SPICE circuit simulator to validate the functionality of the designed circuits. A qualitative summary and comparison for all of the MTJ bit-cells described in herein are listed in Table 2.1.

Table 2.1: Summary of NVM Bit-Cells using Different MTJ Switching Approaches [3].

Memory Bit-Cell	Bit-Cell Area	External Magnetic Field	Energy Consumption	Reliability	Scalibility
FIMS-MRAM	✓	YES	- -	- -	- -
TAS-MRAM	✓	YES	- -	- -	- -
STT-MRAM	✓ ✓	NO	✓	✓	✓ ✓
SHE-MRAM	✓	NO	✓ ✓	✓ ✓	✓

2.1.3 Differential Spin Hall Effect MRAM (DSH-MRAM)

Differential Spin Hall Effect MRAM (DSH-MRAM) bit-cell was proposed in [16] to store both the bit value and its complimentary value with a single write operation. Using DSH-MRAM can reduce the write operation's energy and delay compared to two SHE-MRAM devices [16]. The write and read operations for DSH-MRAM are similar to SHE-MRAM, as shown in Figure 2.3. During the write operation for the DSH-MRAM devices, the charge current, $\mathbf{I}_{SHE}$, is applied through the terminals **T2** and **T4**, and a strong spin-orbit coupling is generated, which results in generation of a spin currents, $\mathbf{I}_{Spin-P}$ and $\mathbf{I}_{Spin-N}$, perpendicular to the charge current, $\mathbf{I}_{SHE}$, and along the positive and negative directions of z-axis of the Cartesian coordinate system, respectively [16]. The generated complimentary spin currents, $\mathbf{I}_{Spin-P}$ and $\mathbf{I}_{Spin-N}$, will result in differential data being stored in top and bottom MTJ devices, shown in Figure 2.3, by changing the magnetization orientation of their free-layer simultaneously. Terminals **T1** and **T4** are used during the read operation for the top MTJ device and terminals **T3** and **T4** are used during the read operation for the bottom MTJ device.

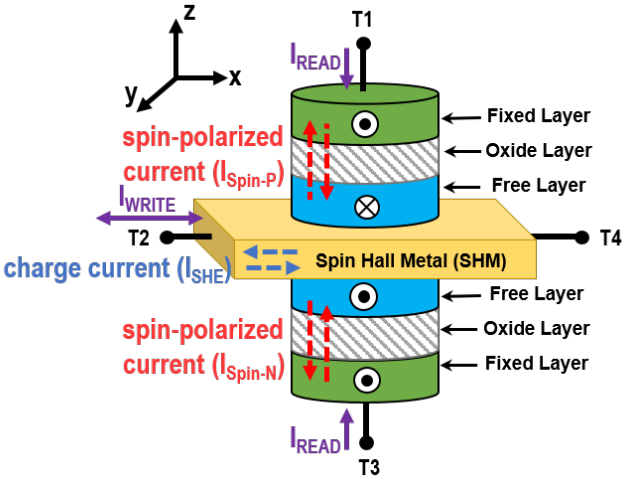


Figure 2.3: DSH-MRAM device structure in P (top) and AP (bottom) states. [4]

2.1.4 VCMA-MTJ Devices for Energy-Efficient Architectures

Although MTJs offer non-volatility, near zero stand-by power dissipation, area efficiency, and fast read operation, their write energy is still significantly higher than volatile switching devices. Thus, it is proposed here to address energy-inefficient and slow write operation by investigating a new approach to modify the switching energy barrier [73]. Due to the current-driven operation of spin-based devices, the majority of the dynamic power dissipation during the switching is caused by ohmic losses and joule heating [74]. In order to solve this issue, researchers have studied the magnetoelectric effect to enable new switching mechanism as an alternative to conventional approaches. The magnetoelectric effect is achieved via utilizing an electric field in order to change the state of the magnetic devices such as MTJs. Using the magnetoelectric effect, MTJ devices will benefit from *faster and more efficient switching while consuming less energy* [73, 74, 75]. Recent research studies have shown that use of the VCMA effect facilitates the use of an electric field to ease or eliminate the demand of charge current for switching the state of MTJ devices. VCMA generates an electric field that causes an accumulation of electron charge and results in a change of occupation of atomic orbitals at the interface, which causes a change in the magnetic anisotropy of the MTJ. Using a VCMA approach can result in a deterministic change of the magnetic state of the MTJ in an *energy-efficient* and *rapid* manner. In other words, use of VCMA can lower the energy barrier between the P and AP states and facilitate the MTJ to switch states using a voltage applied across its terminals. The effective Perpendicular Magnetic Anisotropy (PMA) of an MTJ in the presence of VCMA effect can be modeled using the following equations [73]:

$$K_{eff}(V_b) = \frac{M_s H_{eff}(V_b)}{2} = \frac{K_i(0) - K_i(V_b)}{t_f} - 2\pi M_s^2, \tag{2.10}$$

$$\Delta(V_b) = \frac{E_b(V_b)}{k_B T} = \Delta(0) - K_i(V_b)\frac{A}{k_B T}, \tag{2.11}$$

$$V_c = \Delta(0)\frac{k_B T t_{ox}}{A\xi}, \tag{2.12}$$

where V_b is the bias voltage applied via VCMA effect, $K_{eff}(V_b)$ is the effective PMA, $H_{eff}(V_b)$ is the effective magnetic field in the presence of bias voltage, M_s is the saturation magnetization, $K_i(0)$ is the initial interfacial PMA energy, $K_i(V_b)$ is the interfacial PMA energy after applying the bias voltage, t_f is the MTJ's free-layer thickness, A is the sectional area of the MTJ, $\Delta(0)$ is the thermal stability factor under zero bias voltage, $\Delta(V_b)$ is the thermal stability factor under bias voltage of V_b, $E_b(V_b)$ is the voltage-dependent energy barrier, k_B is the Boltzmann constant, T is the temperature, V_c is the critical voltage required by VCMA effect to modify the energy barrier, ξ is the VCMA coefficient, and t_{ox} is the MTJ's oxide thickness.

VCMA-MTJ devices require a bias voltage to lower their energy barrier between the two stable states of Parallel (P) and Anti-Parallel (AP). This will result in a more efficient method of switching the device between the P and AP states. When the energy barrier is lowered, a current with smaller magnitude and pulse duration can switch the magnetic orientation or the state of the MTJ devices. As a result, the energy consumption of the write operation will be reduced. The VCMA bias voltage that is required to modify the energy barrier can be found using (2.12). Additionally, as experimental results in [73, 74, 75] have shown, $K_i(V_b)$ demonstrates a linear dependency to the electric field, hence, we can simplify it as $K_i(V_b) = \xi\frac{V_b}{t_{ox}}$ [73] through modifying the Landau-Lifshitz-Gilbert (LLG) equation, shown in (2.13), while updating $\vec{H_{eff}}(V_b)$. As shown in (2.15), the voltage dependent anisotropy field, $\vec{H_{ani}}(V_b)$, changes with the VCMA bias voltage. The changes in (2.15) will then result in the modification of $\vec{H_{eff}}(V_b)$ in (2.14), which presents the effective magnetic field vector. As a result, the VCMA effect will enable the MTJ devices to switch faster and with reduced switching currents due to the lowered energy barrier caused by the VCMA bias voltage.

By requiring reduced current magnitude for a shorter pulse duration, this approach will reduce the overall energy consumption of MTJ devices during the write operation. Additionally, in order to observe the switching of the magnetic orientation of the MTJ devices in the z-axis of the Cartesian coordinate system, we need to solve the LLG equation shown in (2.13). The modifications made in the LLG equation to model the VCMA effect are shown below [73]:

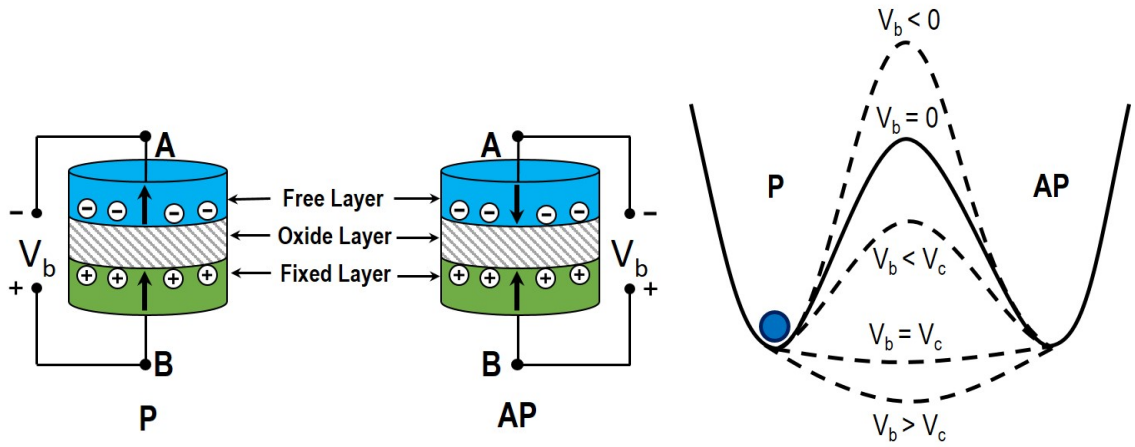
$$\frac{\mathrm{d}\vec{m}}{\mathrm{d}t} = -\gamma\vec{m} \times \vec{H}_{eff}(V_b) + \alpha\vec{m} \times \frac{\mathrm{d}\vec{m}}{\mathrm{d}t} - \rho_{stt}\vec{m} \times (\vec{m} \times \vec{m}_r), \tag{2.13}$$

$$\vec{H}_{eff}(V_b) = \vec{H}_{ext} + \vec{H}_{dem} + \vec{H}_{th} + \vec{H}_{ani}(V_b), \tag{2.14}$$

$$\vec{H}_{ani}(V_b) = \left(\frac{2K_i(0)t_{ox} - 2\xi V_b}{\mu_0 t_f M_s t_{ox}}\right) m_z, \tag{2.15}$$

where $\vec{m}$ is the magnetization vector of the MTJ's free-layer $\{m_x, m_y, m_z\}$, $\vec{m}_r$ is the polarization vector, γ is the gyromagnetic ratio, α is the Gilbert damping factor, $\vec{H}_{eff}(V_b)$ is the effective magnetic field vector in the presence of bias voltage, ρ_{stt} is the STT factor, $\hbar$ is the reduced Planck constant, P is the STT polarization factor, J_{stt} is the driving current density inducing STT, e is the elementary electron charge, μ_0 is the vacuum permeability, $\vec{H}_{ext}$ is the external magnetic field vector $\{H_x, H_y, H_z\}$, $\vec{H}_{dem}$ is the demagnetization field vector, $\vec{H}_{th}$ is the thermal noise field vector, and $\vec{H}_{ani}(V_b)$ is the voltage-dependent anisotropy field vector.

As it can be observed from (2.15), using the VCMA effect, by applying a positive bias voltage across the MTJ, the PMA will be reduced and will result in reduction of its coercivity. On the other hand, by applying a negative voltage across the MTJ, the PMA will be increased and as a result, its coercivity will increase as well. Figure 2.4 shows the effects of VCMA on an MTJ device.



(a) (b)

Figure 2.4: (a) Structure of the VCMA-MTJ. (b) Modification of energy barrier ($E_b(V_b)$) using the VCMA effect. When $V_b > V_c$, the energy barrier is completely eliminated. Additionally, if $0 < V_b < V_c$, the energy barrier will be reduced to facilitate the switching of the state of the MTJ. On the other hand, for $V_b < 0$ the energy barrier will increase [5].

2.1.5 SHE-enabled Domain Wall MTJ Devices

In recent studies, researchers have exploited the use of emerging devices for signal processing applications. In particular, they have explored designing ADCs using emerging devices such as SHE-MTJ [79], Domain Wall Motion (DWM) [82, 111], and Racetrack Memory [112]. The basic concept of spin-based devices is to control the spin of electrons in a ferromagnetic solid-state nano-device.

Figure 2.5 shows a SHE-DWM device [18, 113]. The non-volatile MTJ consists of a Ferromagnetic (FM) layer, which is called the fixed-layer, a FM nano-wire layer, which is called the free-layer, a tunneling oxide layer between the fixed-layer and the free-layer, and a heavy metal to realize Spin-Hall assisted switching. FM layers could be aligned in two different magnetization configurations according to the position of the Domain Wall (DW), Parallel (P) and Anti-Parallel (AP). Accordingly, the MTJ exhibits low resistance (R_P) or high resistance (R_{AP}) states, respectively

[18]. The read and write process for the memory cells are like SHE devices.

Based on Spin-Transfer Torque (STT) switching principles, the P or AP state of the SHE-DWM device is configured by means of the bidirectional current that passes through the Spin-Hall heavy Metal (SHM) from terminal T1 to terminal T3, I_{SHE}. When the I_{SHE} is applied, a strong spin-orbit coupling is generated, which results in generation of a spin current, I_{Spin}, along the z-axis of the Cartesian coordinate system and perpendicular to the I_{SHE} current [113]. The DW will move if I_{SHE} exceeds the critical current, I_C.

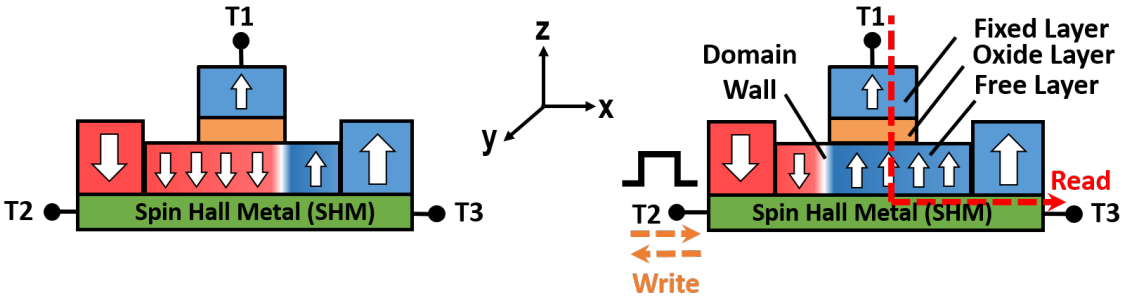


Figure 2.5: SHE-enabled Domain Wall Motion device structure. [6]

Additionally, in order to read the data stored in these devices, a Sense Amplifier (SA) [10] is used to sense the difference between the resistance of the SHE-DWM device that is used to store the data and a reference MTJ device with a known resistance. Terminals T2 and T3 are used during the read operation, meaning that the read and write paths are separate, which is one of the reasons for high reliability of these devices regarding read disturbance failures.

The main reason for using SHE-DWM devices is due to slow and high energy switching of DWM devices alone. Utilizing SHE approach will help reduce the write energy consumption and increase the DW velocity. Conventionally, the DWM was achieved using STT which could switch the domain wall due to the coupling between local magnetic moments of the DW and spin-polarized currents. However, it has been practically shown in recent studies that SHE-DWM devices offer significantly lower energy consumption and faster switching [18]. The spin current, I_{Spin},

generated due to the charge current applied through the SHM, I_{SHE}, can be described using the following equations:

$$I_{Spin} = \theta_{SHM} \times \frac{A_{Spin}}{A_{SHM}} \times I_{SHE} \times \sigma, \tag{2.16}$$

$$A_{Spin} = L_{DW} \times W_{DW}, \tag{2.17}$$

$$A_{SHM} = W_{SHM} \times t_{SHM}, \tag{2.18}$$

$$\sigma = 1 - sech(\frac{t_{SHM}}{\lambda_{sf}}), \tag{2.19}$$

where, L_{DW} and W_{DW} are the length and width of the DW free-layer, W_{SHM} and t_{SHM} are the width and thickness of the SHM, θ_{SHM} is the spin-Hall angle, and λ_{sf} is the spin flip length.

2.1.6 MRAM-based Stochastic Oscillator Devices

Recently, researchers have studied theoretically and experimentally the utilization of thermally unstable superparamagnetic MTJs to realize a variety of functional spintronic devices [7, 114, 115]. Herein, we intend to demonstrate that a recently proposed building block with embedded MRAM technology can enable the hardware realization of a stochastic bitstream generator. The structure of the MRAM-based Stochastic Oscillator (MSO) is depicted in Figure 2.6.Due to the low energy-barrier (i.e. $E_B \ll 40kT$), the MTJ's resistance level fluctuates between the two resistance states of R_{AP} and R_P, which results in the non-uniform stochastic output at the drain of the NMOS transistor shown in Figure 2.6. We can amplify the NMOS drain output to provide full-swing signal, i.e. $[0.0 \rightarrow 0.8]$V, using a single inverter circuit. The probability of the output being '1' can be controlled using the input signal connected to the gate of the NMOS transistor. Thus,

by increasing the gate voltage of the NMOS transistor, V_{IN}, its drain-source resistance, r_{ds}, will decrease, which will result in the drain voltage to be closer to the GND. On the other hand, by decreasing the gate voltage of the NMOS transistor, V_{IN}, its drain-source resistance, r_{ds}, will increase, which will result in the drain voltage to be closer to the VDD.

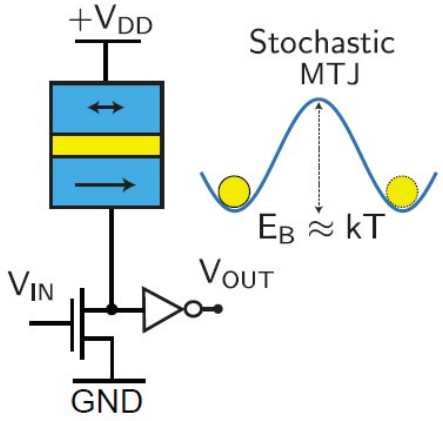


Figure 2.6: The building block of the proposed MRAM-based Stochastic Oscillator (MSO) [7].

Considering the MTJ conductance of the MSO, we can observe the behavior of the circuit shown in Figure 2.6 [7]:

$$G_{MTJ} = G_0 \left[1 + m_z \frac{TMR}{(2 + TMR)}\right], \tag{2.20}$$

where m_z is the free layer magnetization, G_0 is the average MTJ conductance, $(G_P + G_{AP})/2$, and TMR is the tunneling magnetoresistance ratio. The drain voltage of the NMOS transistor shown in Figure 2.6 can be expressed as:

$$V_{DRAIN}/V_{DD} = \frac{(2 + TMR) + TMR\, m_z}{(2 + TMR)(1 + \alpha) + TMR\, m_z}, \tag{2.21}$$

where α is the ratio of the transistor conductance, G_T, to the average MTJ conductance, G_0. When $\alpha \approx 1$ maximum fluctuations can be achieved. This means, when $V_{IN} = V_{DD}/2$, the MTJ resistance is approximately equal to r_{ds}. Herein, we use a circular nanomagnet with near-zero energy barrier without shape anisotropy. Such magnets have been fabricated and characterized in [116].

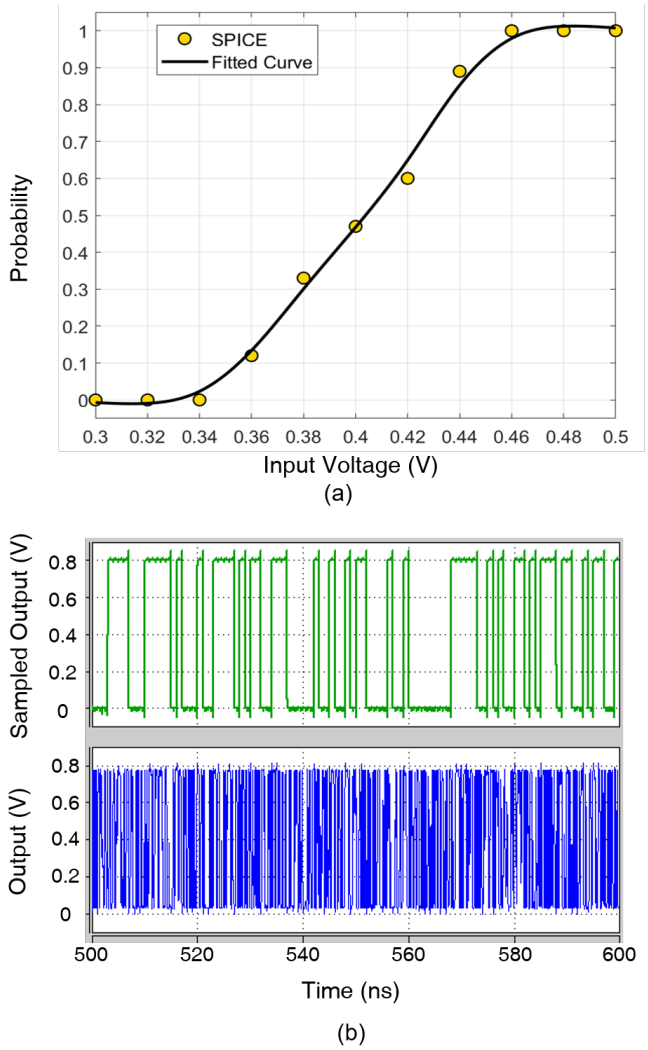


Figure 2.7: (a) Output probability of MSO versus its input voltage, (b) The output and sampled output voltages for $V_{IN} = 0.5V_{DD} = 400$mV. [8]

We use the embedded MRAM-based model developed in [7] to perform SPICE circuit simulations using the parameters listed in Table 2.2 and the nominal voltage of $V_{DD} = 0.8$. The magnetization input for the MTJ conductance elaborated in Equation 2.20 is provided by the stochastic Landau-Lifshitz-Gilbert (LLG) equation:

$$(1+\alpha^2)\frac{d\hat{m}}{dt} = -|\gamma|\hat{m}\times\vec{H} - \alpha|\gamma|(\hat{m}\times\hat{m}\times\vec{H}) + 1/qN(\hat{m}\times\vec{I_S}\times\hat{m}) + \left(\alpha/qN(\hat{m}\times\vec{I_S})\right), \tag{2.22}$$

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. The relation between the probability of output being '1' and V_{IN} is depicted in Figure 2.7(a), where $V_{IN} = V_{DD}/2 = 400$mV generates an output probability of 50%, as shown in Figure 2.7(b).

Table 2.2: Modeling and Simulation Parameters [7].

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

2.2 Reliability Challenges of MTJ Sensing Operation

An ongoing research on reliability issues of STT-MRAM devices resulted in some possible solutions which each of them utilize different properties of the MTJ switching behavior. Based on recent studies conducted on the reliability improvement of STT-MRAM devices, it has been determined that some of the preferred SA designs are able to offer less than 5ns read sensing latency while maintaining wide sensing margins. Furthermore, non-destructive sensing schemes offer lower energy consumption while suffering from narrow sensing margins [25, 27, 36, 38, 83, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126]. Research has also shown that STT-MRAM SAs' performance span in 3 different ranges across all proposed design strategies. The highest performance strategies deliver a sensing margin of approximately above 300mV while incurring low power and energy consumption in the order of pico-Joules and micro-Watts respectively with read or sense latency in the range of pico-Seconds [25, 27, 36, 38, 83, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126].

STT-MRAM has several advantages over other emerging memory technologies, however, it faces some distinct reliability challenges involving read and write failures [27] as listed herein. STT-MRAM scalability is greatly influenced and limited due to thermal fluctuations and issues such as MTJ process variations and the CMOS access transistor have had negative effects on STT-MRAM devices. Also, as a result of these issues, demand for an advanced sensing circuit which can provide required sensing margin along with low power operation has been increased.

STT-MRAM bit errors can be significantly influenced due to process variations [127] which precipitate another important issue that STT-MRAM suffers from as well as suffering from its unique intrinsic thermal randomness. These variations include variation in the access transistor sizes, variation in threshold voltage V_{th}, MTJ geometric variation and initial angle of the MTJ. Whereas the effect of variation involving the access transistor on system performance has been investigated in [38], here we focus on the process variation of the MTJ cell. The difference between the sensed bit-line voltage and the reference voltage which is known as the Sense Margin (SM) will be small due to the wide distribution of MTJ resistance which can also result in a false detection scenario [36]. On the other hand, write speed can be affected and may vary due to the thermal fluctuations during MTJ switching in write operations and this will further aggravate by process variation-induced variability of the switching current [38].

Errors due to the STT-MRAM physical nature's failures will be categorized into transient faults and permanent faults as depicted in Figure 2.8. Transient faults, which can also be described as an incorrect signal condition, is mostly caused by the parameters of free layer such as current density (J_c) and thermal stability factor (Δ). Permanent faults, which can be precipitated by destructive device damage, are initially caused by susceptibility to the sensitive parameters of oxide barrier such as barrier's thickness t_{ox} and Tunnel Magneto-Resistance (TMR) ratio [124], which have been expanded with additional parameters in Table 2.3.

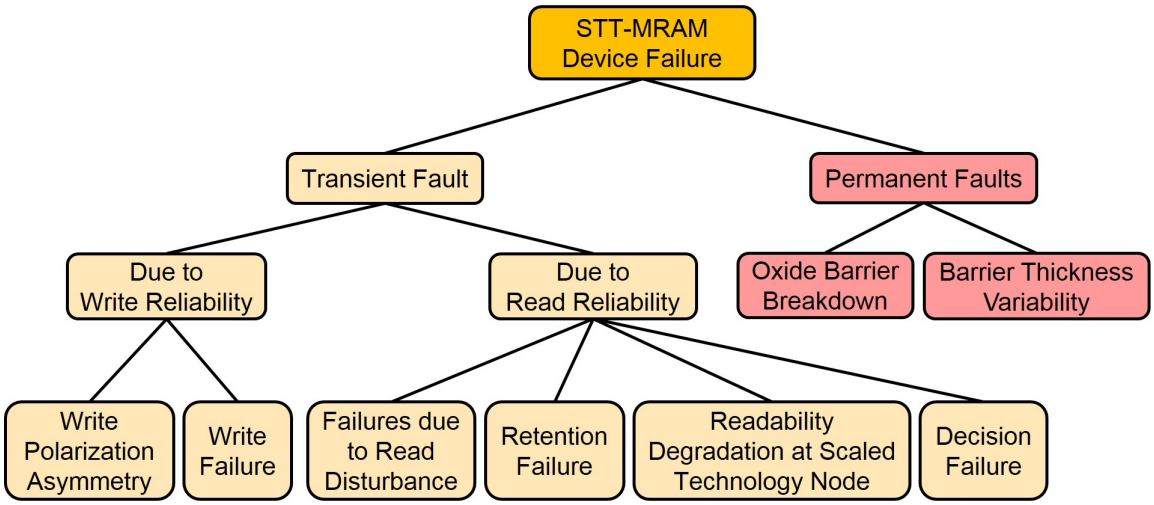


Figure 2.8: Taxonomy of STT-MRAM Device Failures. [1]

2.2.1 STT-MRAM Reliability

Several studies have shown that exposures to permanent and transient faults can be mitigated in various device technologies by employing spatial hardware redundancy, temporal operational redundancy, error correcting codes, and resiliency via active reconfiguration. In order to make reliable STT-MRAM cells which are less vulnerable to alpha-particle-induced transient faults, a variety of design strategies and different considerations have been proposed in [83, 128, 129, 130, 131]. For instance, in [130], concept of '1'/'0' dual-array equalized reference is proposed that reduces the error rate by lowering the read current which introduces a precise reference for stable read operation.

Research studies have shown that STT-MRAM memory devices are vulnerable to radiation induced transient faults due to their CMOS peripheral circuit used to read and write in the MTJ cells [132, 133, 134]. It has been proven experimentally in [134] that the MTJ device itself is resilient to radiation induced transient faults. In order to reduce the vulnerability of the STT-MRAM devices to radiation induced transient faults caused by elements such as alpha particle, Kang *et al.*

have designed and examined a novel area and power efficient sensing circuit in [135]. They have experimentally shown that their design increases the robustness of memory and logic devices using hybrid CMOS/STT-MRAM to radiation induced transient faults such as Single Event Upsets (SEUs) and Multiple Bit Upsets (MBUs) [135].

Table 2.3: STT-MRAM Reliability Issues. [1]

Type			Description	Possible Solution(s)
Transient Faults	Write Reliability Issues	Write Failure	▪ Due to stochastic nature of write process in STT-MRAM ▪ MTJ cell does not switch properly in order to store the required value within write time period	▪ Possible solution can be increasing the write duration ▪ Another possible solution can be increasing write current ▪ These solutions may cause significant amount of power dissipation and area overhead as well as speed degradation
		Write Polarization Asymmetry	▪ P to AP needs higher switching current and suffers from more error rate compared to AP to P switching	▪ Reversed MTJ Connection results in a larger I_{MTJ} for the P to AP switching which alleviates the effect of the I_c asymmetry (I_c (P→AP)/I_c (AP→P) > 1)
	Read Reliability Issues	Failures due to Read Disturb	▪ Read and Write share the same path ▪ Unwanted bit-flip during a read operation ▪ Growing with technology scaling ➢ thermal stability factor Δ decreases ➢ critical switching current decreases	▪ Possible solution can be increasing the margin between read and write currents ▪ Increasing the write current is one of the solutions which may not be feasible since write current maintains a high value in STT-MRAM devices. ▪ Also can be done by decreasing the read current which will Increase the read latency and may result in another reliability issue called decision failure ▪ Error Correction Codes (ECC)
		Readability Degradation at Scaled Technology Node	▪ Due to reduction in switching current which will become a greater concern in scaled technology node ▪ Reduction in switching current will limit the upper-bound of sensing current	▪ High read current is required in order to: ➢ Provide enough sense margin ➢ Ensure reliable sensing by excluding the device variation of the sense amplifier ➢ Maintain fast read and reduce read latency ▪ Low read current is required in order to: ➢ Prevent stored data from being upset
		Decision Failure	▪ When reading an MTJ cell, not being able to distinguish whether the stored bit is zero or one	▪ Possible solution can be increasing the read duration ▪ Another possible solution can be increasing read current
		Retention Failure	▪ STT-MRAM had an intrinsic thermal instability which can result to a bit-flip of an MTJ cell's content	▪ One solution at the device-level is exploiting the thermal stability factor Δ ▪ Increasing Δ results in longer read duration, larger current amplitude and increase in number of bits per word during parallel reading
Permanent Faults		Oxide Barrier Breakdown	▪ Switching current and switching duration are inversely proportional to each other ▪ High current density J_c is normally required in order to achieve high speed based on: $V = R.A \times J_c$ ▪ In order to have a better switching probability we have to provide a large sensing margin and in order to maintain high current density we can reduce the Resistance Area product ($R.A$) or thickness of the oxide or increase the bias voltage V ▪ Each of these solutions can result in the oxide barrier breakdown and shorten the MTJ lifetime	▪ Using Modular Redundancy ▪ In order to prevent permanent faults, oxide thickness variation is required to be less than 5% ▪ Using a low bias voltage for sensing is suggested since the real TMR ratio decreases during the sensing operation
		Barrier Thickness Variability	▪ Maintaining low $R.A$ value, favorably ultra-thin insulator or oxide barrier is required ▪ MTJ's resistance is proportional to the oxide thickness exponentially. Increase in bias voltage will result in decrease in TMR ratio and TMR ratio may become less than the resistance Variation Ratio (VR) ➢ In this case, sensing margin will be upset by VR and permanent faults will occur as a result	

Furthermore, in order to tolerate permanent faults, two solutions have been proposed in [124]. One option is Triple Modular Redundancy using a majority voter and the other option is to resize the active transistors. However, both options introduce area overheads whereas Triple Modular Redundancy requires two additional SAs along with a voting circuit, and meanwhile the second option utilizes a larger transistor. While Triple Modular Redundancy is a popular approach for masking soft errors and providing single-fault coverage in various circuits, it incurs roughly three-fold increases in area and energy.

In a recent research study in [136], Kang *et al.* have proposed a novel area efficient and high speed Error Correcting Code (ECC) circuit utilizing Orthogonal Latin Square Code (OLSC) in order to increase the reliability of STT-MRAM with the option of adaptability that enables the system to adapt the error correction based on its needs. Moreover, in order to increase the yield of STT-MRAM devices, Kang [137] have introduce an innovative method to sustain permanent and transient faults using an integration of ECC along with Fault Masking (FM) methods which they address as sECC [137]. In addition, in their study by combining sECC method with Redundancy Repair (RR) method, they have successfully managed to further improve and optimize the performance of the emerging STT-MRAM devices. Kang *et al.* in [137] have managed to repair the permanent faults in the system using redundant elements so called RR, mask transient faults and Single Isolated Faults (SIFs) using sECC.

Alternatively, results of the study in [124] have shown that only resizing the transistors is sufficient for increasing the reliability of the conventional applications, however, for those applications which require extreme sensing reliability, Triple Modular Redundancy technique can be more useful. Considerable amount of research has been performed on improving reliability and performance of STT-MRAM memory devices. In [138], the write and read performance of STT-MRAM as last-level on-chip cache have been estimated and analyzed over the processor performance.

Moreover, in [139], an early-write-termination scheme was proposed in order to enhance STT-MRAM reliability and reduce STT-MRAM write energy. Furthermore, in [140], an implementation of STT-MRAM under different technology nodes has been discussed in which the corresponding process technology and scaling parameters were presented. Furthermore, in order to reduce the probability of accidental bit-flipping and loss of data caused by the current applied during read period, a disturbance-free read scheme was proposed in Gigabit scale STT-MRAM design in [141], which, due to process variations of MTJs, this scheme is unable to solve the read failures.

In order to reduce the read disturb probability, [142] has proposed the pulsed read method and [143] has proposed the disruptive reading and restoring scheme. Furthermore, in [27], in order to alleviate the read disturb reliability issue, some bit-cell architectures are proposed. Moreover, it has been shown that by increasing the thermal stability factor we can reduce the read disturb rate. However, all of these techniques introduce large area overhead and/or large power dissipation and large delays as mentioned in [27]. In general, sensing schemes can be classified into two categories, Destructive and Non-Destructive [118]. Based on the definition presented in this research, Destructive Schemes are more vulnerable to read reliability failures. Non-Destructive Schemes are more tolerant to process variation of reference cell, however, Destructive Schemes, typically, provide smaller read/sense latency.

2.2.2 Destructive Sensing Schemes

The first category of strategies to mitigate the cost of self-referencing is through consecutive accesses that restore the destroyed value once it has been reliably read. Several self-reference sensing schemes were proposed to overcome reliability concerns due to process variations of MTJs in STT-MRAM. In [25], Sun, *et al.* have analyzed the Conventional Sensing Scheme (CSS) which compares the bit line voltage to a reference voltage to read the value of a memory bit cell under

certain conditions. However, as technology shrinks process variation will be increased and will result in significant standard deviations of sense margin that will lead to large read failure probability. As a result, the chip yield in STT-MRAM design is highly limited due to poor robustness of CSS.

As mentioned in [144, 145], original value stored in an MTJ cell in Conventional Self-Reference Sensing Scheme (CSR), will be compared to a reference value which is stored in the same MTJ in a different write cycle. As it can be found in [144, 145], CSR consumes a large amount of power and also introduces long latency. Sun, *et al.* also analyzed CSR in [25], which needs two write operations that results in long latency and will lead to large power overhead and can also be destructive to the stored value. Comparing CSR to CSS, we can conclude that CSR maintains a higher sense margin with the cost of sacrificing the reliability and performance.

One of the destructive sensing strategies that has been recently used in [30] was proposed by Zhao *et al.* called Pre-Charge Sensing in [83, 124] which uses a Pre-Charge Sense Amplifier (PCSA) and minimizes the read current value and read duration compared with conventional static data sensing. As a result of this action, high reliability will be provided for the STT-MRAM while maintaining the same thermal stability factor. This method, which is called Dynamic Sensing Scheme (DSS) [129], has solved the sensing problem utilizing Dynamic Sensing [124]. In order to reduce the read current required, Lakys *et al.* in [128] have used Dynamic Sensing Scheme using two word selection transistors for each MTJ cell in order to perform read operations and switching operations. Also with this method, the size of the reading transistor can minimized which can lead to reduction of read current far below the disturb margin down to $10\mu A$. This method introduces some area overhead compared to conventional 1T-1R STT-MRAM cell designs, however, this area overhead is negligible to implement cross-point memories since in these designs the selection transistors are shared among several MTJs within the same word [128].

Lakys *et al.* also suggest a method called Self-enable Switching Circuit (SSC) to decrease the impact of stochastic switching in [128] which operates based on relaxing the bias voltage stress on the oxide barrier and utilizing short duration write pulse sequence instead of fixed long writing pulse within switching and sensing operations. This method reduces the probability of oxide breakdown and allows short write pulse durations which will result in reduction in the number of switching operation and this will improve the oxide barrier lifetime [124]. Later in 2012, Ren *et al.* have proposed Body Voltage Sensing Circuit (BVSC) in [122], which like SSC utilizes a short pulse reading scheme that enables fast sensing operations. They have shown that their design provides improved speed at the cost of sacrificing sense margin and also improves reliability of the read operation against read disturbance.

In order to increase the sensing margin while reducing the latency and improving device variation tolerance of the STT-MRAM cell, Zhang *et al.* in [36] have analyzed Regular Differential STT-MRAM Cell Structure (RDAMS). This method uses a differential STT-MRAM cell design including two separate 1T-1R cells which the resistance state of these two are always opposite. This design doubles the maximum sense margin compared to the one of 1T-1R cells (CSS), however, RDAMS capacity is half of the one of the two 1T-1R cells used in this design. In the same research publication, Asymmetric Differential Cell Structure (ADAMS) have been proposed in order to improve the read and write performance of STT-MRAM and also increase the transient fault tolerance compared to RDAMS. This method uses a differential STT-MRAM cell design like RDAMS including two separate 1T-1R cells that one of the MTJ cells is reversely connected to the NMOS transistor. Write latency of ADAMS is the same as RDAMS while maintaining smaller cell area compared to other previous sensing schemes such as CSS, CSR, etc. ADAMS has improved the read latency and has reduced the write error rate [36].

In an extension to a previous research on PCSA, Kang *et al.* proposed Separated Pre-Charge Sense Amplifier (SPCA) [117]. Based on their simulation result, SPCA has a similar performance in

terms of latency and power and provides better reliability with a small area overhead compared to PCSA proposed in [83]. In another interesting research study proposed in 2014, Eken *et al.* in [38] introduce a novel strategy called Field-Assisted STT Self-Referencing Scheme (FA-STT) which utilizes an external magnetic field to generate the self-reference sensing signal. This method offers improved process variation resilience and thermal fluctuation tolerance in STT-MRAM and MTJ switching respectively. It also provides a much better read reliability by improving read sense margin compared to conventional self-reference sensing schemes (CSS, CSR, NDSR, and VDRS) and it significantly reduces the write error rate.

Furthermore, Slope Detection Sensing Scheme was suggested by Motaman *et al.* in [118] which will also be categorized as destructive scheme in which they have claimed makes the STT-MRAM cell design more reliable against device mismatch and variations. They have discussed that their design has a high sensing robustness due to eliminating the reference comparison.

Finally, in a recent research published in 2016, Parallel Reading Sense Amplifier (PRSA) is proposed [125]. This sense amplifier has two sensing steps that the first read step can be performed in parallel with the write operation which reduces the read latency. Authors in [125] have claimed that PRSA offers large sensing margin.

Table 2.4 lists detailed numerical values for all of the destructive sensing schemes reviewed herein as well as a description of their qualitative attributes. Additionally, Figure 2.9 depicts the Read-/Sense latency vs. Sensing Margin of STT-MRAM Destructive Sensing Schemes which have been proposed as time progressed from the initial designs on the left side of the plot to current designs on the right side of the plot.

Table 2.4: Destructive Sensing Schemes and Their Attributes. [1]

#	Reference(s)	Approach	Area per Cell (device count)				Cycle(s) or Stage(s)	Sense Margin ($V_{P/AP}$-V_R)	Read Latency or Sensing Latency	AVG. Energy or Power consumption	Simulation-based or Theoretical
			MTJ	CMOS	Cap	Other					
1	[144]	Self-Reference Scheme (SRS)	1	17	2	0	2	N/A	Large (130ns)	N/A	Simulation-based TSMC 240-nm
2	[25]	Conventional Sensing Scheme (CSS)	1	1	0	0	2	Small (~20mv)	Small (2.5ns)	Low (0.8937pJ)	Theoretical
3		Conventional Self-Reference Sensing Scheme (CSR)	1	1	2	0	2	Large (76.6mv)	Large (40ns)	High (22.05pJ)	
4	[124, 83]	Dynamic Sensing Scheme Using Pre-Charge Sense Amplifier (PCSA)	2	7	0	0	2	Large (N/A)	Small (~164ps)	Low (~3.17fJ)	Simulation-based TSMC 65-nm
5	[122]	Body Voltage Sensing Circuit (BVSC)	3	21	2	1 Res	1	Large (195mv)	Small (1ns)	Low (195.5fJ & 300uW)	Simulation-based TSMC 65-nm
6	[128]	Self-enable Switching Circuit (SSC)	2	13	0	0	2	N/A	Small (<200ps)	Low (N/A)	Simulation-based TSMC 65-nm
7	[36]	Regular Differential STT-MRAM Cell Structure (RDAMS)	2	4	0	0	1	Large (N/A)	Small (321.8ps)	N/A	Simulation-based
8		Asymmetric Differential STT-MRAM Cell Structure (ADAMS)	2	4	0	0	1	Large (N/A)	Small (266.7ps)	N/A	TSMC 45-nm
9	[117]	Separated Pre-Charge Sense Amplifier (SPCSA)	2	7	0	0	2	Large (N/A)	Small (~187ps)	Low (~3.84fJ)	Simulation-based TSMC 65-nm
10	[38]	Field-Assisted STT Self-Referencing Scheme (FA-STT)	1	5	2	External B-Field	2	Large (>20mv)	N/A	N/A	Simulation-based TSMC 45-nm
11	[118]	Slope Detection Sensing Scheme	1	20+4	3	1 Current Source & 2 Switches & 1 Res	2	Large (N/A)	Large (6.8ns)	Low (150uW)	Simulation-based TSMC 22-nm
12	[125]	Parallel Reading Sense Amplifier (PRSA)	1	18+2	2	1 Switch	2	Large (N/A)	Large (<20ns)	N/A	Simulation-based TSMC 180-nm

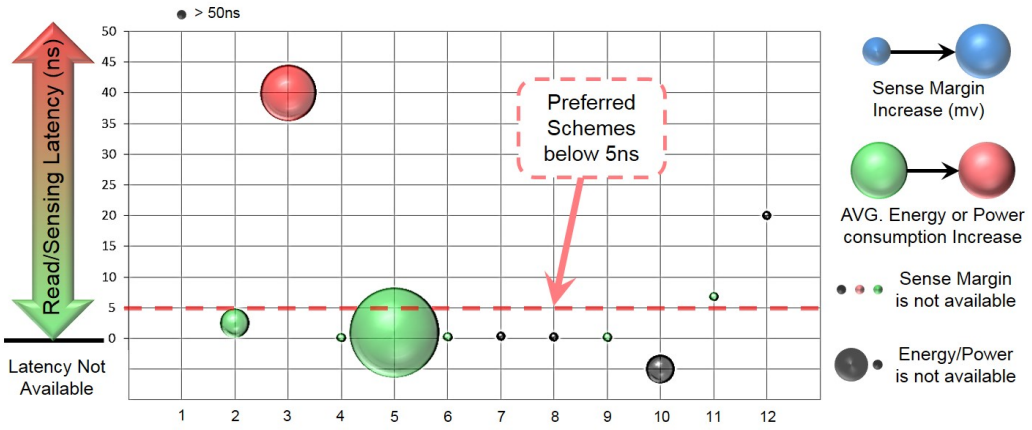


Figure 2.9: Read Sense Latency vs. Sensing Margin of STT-MRAM Destructive Sensing Schemes and Circuit Designs (with the same order as listed in # column in Table 2.4). [1]

2.2.3 Non-Destructive Sensing Schemes

The second category of strategies to mitigate the cost of destructive sensing are being discussed in this Section. Two different design methods that have been demonstrated in [146] are Low-Power Simple Sensing Circuit and High-Sensitivity Switched-Current Sensing Circuit. It has been exhib-

ited that if low power operation is a priority, then Low-Power Simple Sensing Circuit can offer better performance. On the other hand, if a better magneto-resistance ratio and faster reading is required, then High-Sensitivity Switched-Current Sensing Circuit would be preferred. Even though, High-Sensitivity Switched-Current Sensing Circuit faces challenges compared to Low-Power Simple Sensing Circuit in terms of power consumption, it can provide better performance in terms of speed. Reducing the sensing latency and read disturbance faults are of significant importance when designing a sensing circuit. In order to maintain low read latency while improving the device performance, an Adequate-Reference Scheme is proposed in [147] that increases the sense margin of STT-MRAM cells which results in read latency reduction. However, the read disturbance error rate might be increased due to increase in the magnitude of the sensing current. Negative-Resistance Read Scheme (NRRS) which is a current-based sensing scheme has been proposed, fabricated, and tested in [148], and authors have analyzed the non-destructive read operation performed.

Moreover, in [149], taking advantage of the different current dependencies of the high and the low resistance states of an MTJ, another sensing scheme was proposed called Nondestructive Self-Reference Sensing Scheme (NDSR). NDSR utilizes a characteristic of MgO-based MTJ which is the difference between the current roll-off slope of high and low resistance states. Based on this characteristic we can see that if the MTJ has a high resistance state, the current roll-off slope will be steeper than the low resistance state. It has been proven in [149] that although this method has two read steps, NDSR has reduced the power consumption and significantly improved the read latency by removing the two write operation and performing one write operation instead. However, compared to CSR, NDSR has a smaller sensing margin. Furthermore, in [150], they have improved the sensing margin of NDSR utilizing combined magnetic and circuit level enhancements. Researchers have introduced a dual-voltage row decoder with a charge sharing scheme in [123]. This scheme has shown reduced read disturbance while providing short sensing latency, resulting in increase in the yield of STT-MRAM devices.

Device variation tolerance and large sensing margin are other important considerations in designing sensing circuits. In a recent research study published in 2012, a sensing scheme have been proposed and discussed [151], which can tolerate the process variation in the scaled technology nodes. In order to overcome the issues due to process variation and asymmetry of Read Access Pass Yield (RAPY) of the memory cells, respectively they have developed a Source Degeneration Scheme (SDS) and a Balanced Reference Scheme (BRS).

Furthermore, Non-Destructive Variability Tolerant Differential Read Scheme was proposed in [152] which is targeting the device variation while improving other reliability aspects of the device. This design has the advantage of complimentary inputs as well as providing large sense margin along with better reliability by reducing error rate as mentioned in [152]. Later in 2012, due to the small sensing margin of the NDSR proposed in [149], Sun *et al.* came up with a circuit to provide a better sensing margin [25]. This method, which is called Voltage-Driven Non-Destructive Self-Referencing Sensing Scheme (VDRS), is more robust than NDSR and maintains a better tolerance on variation of MTJ devices as well as providing a high sense margin. VDSR has low read latency and low power consumption compared to NDSR.

In [25], authors have shown that this method demonstrates the highest STT-MRAM array yield among all existing sensing schemes of STT-MRAM design. In another research to improve the STT-MRAM reliability, Offset-Tolerant Triple-Stage Sensing Circuit has been proposed by Kang *et al.* in [153]. They have verified that their design can reduce the device errors due to process variation and read disturbance which increases the STT-MRAM reliability in scaled technology nodes while providing large sense margin.

In 2014, Kim *et al.* in another research have introduced Split-Path Sensing Circuit (SPSC) [154] and have shown that using variable reference voltage, their design consumes less energy compared to Highly-Symmetric Cross-Coupled Current Mirror (HSCC) proposed in [155], SDSC [151], and

BVSC [122] while providing a large enough sensing margin.

Another sensing scheme which can tolerate the process variation in the scaled technology nodes and improve the reliability of STT-MRAM, has been proposed and discussed in [119]. In this research study, authors have suggested an Offset-Canceling Triple-Stage Sensing Circuit (OCTS) in order to overcome the issues due to process variation and asymmetry of RAPY of the memory cells that can operate with low currents which will result in avoiding read disturbance.

Moreover, in [156] Self-Body Biasing Sensing Circuit (Self-BB) is proposed, which is thought to also tolerate issues due to process variation and asymmetry of RAPY of the memory cells while providing better sensing margin compared to conventional sensing schemes and performing fast sensing operations.

Read disturb faults and device variation are highly relative to the scaling of the technology node which both can affect the device reliability in a negative way. The first step in order to prevent the read disturb faults effectively is to detect them. In [27], a circuit has been proposed that has the ability to detect the read disturb fault utilizing a self-test mechanism that is supposed to validate its behavior which is called Read Disturb Detection Scheme (RDD). It has been shown that using this method, up-to 95% of the total read disturb faults can be detected while maintaining negligible area and power overhead.

In order to reduce device variation effect on STT-MRAM reliability, a variation-tolerant high-reliability sensing scheme has been introduced and designed which is shown to increase the sensing margin with the cost of more delay and loss of speed [157]. This design includes three stages of sensing. One of these stages is a pre-amplifying stage which utilizes a charge transfer amplifier to amplify the voltage difference between the reference MTJ cell and main MTJ cell. In another research publication, Kang *et al.* [120] utilized a modified charge transfer stage and a source follower amplifier which makes the design more reliable against device mismatch and variations,

prevents read disturbance, and further improves the sense margin.

In a recent research publication, Covalent-Bonded Cross-Coupled Current-Mode Sense Amplifier (CBSA) has been proposed and fabricated in [158]. In this design, the source lines are merged throughout the whole memory array in order to make the design more compact in terms of area efficiency. Authors have shown that this design reduces the mismatch sensitivity of the cross-coupled latch in the sense amplifier design.

Providing reliable solutions are valuable, however if the power consumption of a reliable design is high, then that design cannot be a good alternative. In an effort for maintaining low power while having a reliable design, Lee *et al.* have recently proposed Pre-Read and Write Sense Amplifier (PWSA) which they have shown that due to its pre-read stage, the write error rate can be controlled. Based on their result shown in [126], their design provides a fast reading circuit that consumes a small amount of power and increases the reliability through controlling and reducing the write error rate.

Another source of reliability exposure that results in read disturbance or read failure, is the thermal instability. In [159], a Body-Biasing Feedback Circuit is proposed to improve the sensing margin and reliability of the STT-MRAM against thermal instability which will further reduce read failures and disturbance. In [160], authors have introduced Dynamic Referencing Sensing (DRS) scheme in order to prevent read disturb reliability issue. Based on their design, the area overhead is negligible due to the fact that the sensing circuitry is shared among the memory cells along the bit-line or word-line. In their design, they manipulate the sensing circuit with regards to the sensed signal and adaptively configure the resistance of the load transistor's resistance.

In a recent research [161], another RDD circuit has been demonstrated and examined. Based on the fact that if a read disturb occurs, then the read current will have a sudden change due to change in the resistance of the MTJ and also knowing that this change will be in a unidirectional fashion

either from **P** to **AP** or from **AP** to **P**, they have designed a circuit that can detect the read disturb fault. However, they have also mentioned that there exists a small latency before activation of the RDD circuit which if any read disturb fault occurs within the duration of the latency it will not be detected. They have also exhibited that the probability if detecting the read disturb will increase in their design if TMR ratio increases and/or if the detection time increases. The area overhead of this RDD design is negligible compared to the area of the chip as claimed in [161]. Finally, a Degenerated Cross-Coupled Sensing Circuit (DCCSC) is been proposed in [121] which is proven to have wide sense margin while consuming small amount of energy with a fast sensing time. They have designed a new reference cell that exhibited increased reliability against device mismatch and variations, and ameliorates read disturbances.

Table 2.5 lists detailed numerical values for all of the non-destructive sensing schemes reviewed herein as well as a description of their qualitative attributes.Moreover, Figure 2.10 illustrates the Read/Sense latency vs. Sensing Margin of STT-MRAM Non-Destructive Sensing Schemes which have been proposed as time progressed from the initial designs on the left side of the plot to current designs on the right side of the plot.

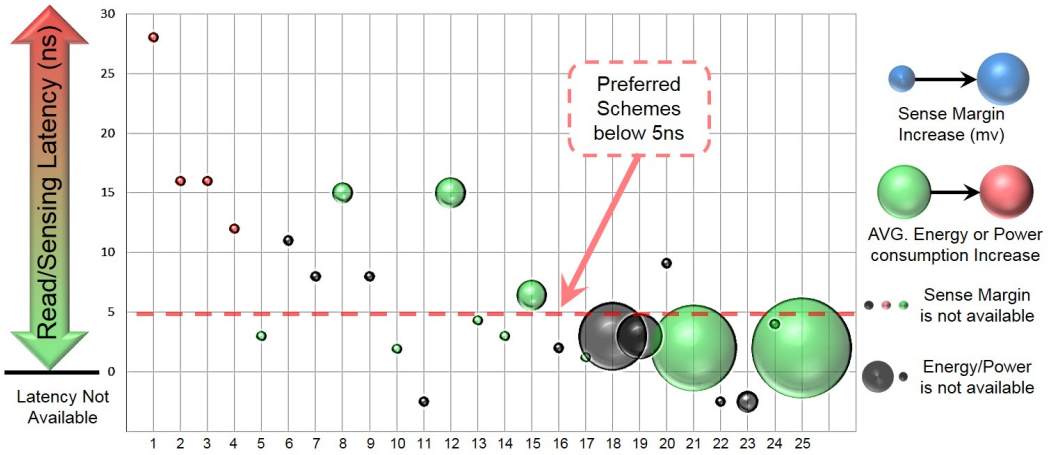


Figure 2.10: Read Sense Latency vs. Sensing Margin of STT-MRAM Non-Destructive Sensing Schemes and Circuit Designs (with the same order as listed in # column in Table 2.5). [1]

Table 2.5: Non-Destructive Sensing Schemes and Their Attributes. [1]

#	Reference(s)	Approach	Area per Cell (device count)				Cycle(s) or Stage(s)	Sense Margin ($V_{P/AP}$-V_R)	Read Latency or Sensing Latency	AVG. Energy or Power consumption	Simulation-based or Theoretical
			MTJ	CMOS	Cap	Other					
1	[146]	Low-Power Simple Sensing Circuit	5	15+ 2×SA	0	0	2	N/A	Large (28ns)	High (6.23mW)	Simulation-based TSMC 600-nm
2									Large (16ns)	High (1.75mW)	Simulation-based TSMC 180-nm
3		High-Sensitivity Switched-Current Sensing Circuit		31+ 2×SA					Large (16ns)	High (8.99mW)	Simulation-based TSMC 600-nm
4									Large (12ns)	High (5.70mW)	Simulation-based TSMC 180-nm
5	[155, 154]	Highly Symmetric Cross-Coupled Current Mirror (HSCC)	5	16	0	0	1	N/A	Small (3ns)	Low (0.76pJ)	Simulation-based TSMC 45-nm
6	[147, 158]	Adequate-Reference Scheme (ARS)	6	26	0	1 Current Source	1	N/A	Large (11ns)	N/A	Simulation-based TSMC 65-nm
7	[148]	Negative-Resistance Read Scheme (NRRS)	1	14+ 2×SA	0	1 Res	1	N/A	Large (8ns)	N/A	Simulation-based TSMC 130-nm
8	[149, 150]	Nondestructive Self-Reference Sensing Scheme (NDSR)	1	5	1	2 Res	2	Large (>20mv)	Large (15ns)	Low (1.04pJ)	Simulation-based TSMC 130-nm
9	[123]	Disturbance-Free Scheme	3	25+SA	0	0	1	N/A	Large (8ns)	N/A	Simulation-based TSMC 45-nm
10	[151]	Source Degenerating Scheme and Balanced Reference Scheme (SDSC)	3	19	0	0	1	N/A	Small (1.9ns)	Low (134.4fJ)	Simulation-based TSMC 65-nm
11	[152]	Non-Destructive Variability Tolerant Differential Read Scheme	4	11+SA	0	0	1	N/A	N/A	N/A	Simulation-based TSMC 22-nm
12	[25]	Voltage-Driven Non-destructive Self-Reference Sensing Scheme (VDRS)	1	9	2	0	2	Large (>45mv)	Large (15ns)	Low (12.08pJ)	Simulation-based TSMC 130-nm
13	[136]	Offset-Tolerant Triple-Stage Sensing Circuit	5	8+2×SA	2	0	3	N/A	Small (4.3ns)	Low (40fJ)	Simulation-based TSMC 40-nm
14	[154]	Split-Path Sensing Circuit (SPSC)	5	16	0	0	1	N/A	Small	Low	Simulation-based TSMC 45-nm
15	[156, 120]	Offset-Canceling Triple-Stage Sensing Circuit (OCTS)	5	15+14	3	7 Switches	3	Large ∼45.21mv	Large (6.4ns)	Low (395.5fJ)	Simulation-based TSMC 45-nm
16	[156]	Self-Body Biasing Sensing Circuit (Self-BB)	3	15	0	0	1	N/A	Small (2ns)	N/A	Simulation-based TSMC 45-nm
17	[27]	Read Disturb Detection Scheme (RDD)	6	37	0	0	1	N/A	Small (1.2ns)	Low (N/A)	Simulation-based TSMC 65-nm
18	[157]	Variation-Tolerant High-Reliability Sensing Scheme	5	17+14	3	7 Switches	3	Large (∼227.54mv)	Small (3ns)	N/A	Simulation-based TSMC 40-nm
19	[120]	Variation-Tolerant and Disturbance-Free Sensing Circuit	5	17+14	3	7 Switches	3	Large (∼102.14mv)	Small (3ns)	N/A	Simulation-based TSMC 40-nm
20	[158, 133, 135]	Covalent-Bonded Cross-Coupled Current-Mode Sense Amplifier (CBSA)	3	31+SA	0	0	2	N/A	Large (9.1 ns)	N/A	Simulation-based TSMC 65-nm
21	[126]	Pre-Read and Write Sense Amplifier (PWSA)	1	46	0	1 Res	4	Large (360mv)	Small (2ns)	Low (18uW)	Simulation-based TSMC 65-nm
22	[159]	Body-Biasing Feedback Circuit	3	12+SA	0	0	1	Large (N/A)	Small (N/A)	N/A	Simulation-based TSMC 40-nm
23	[160]	Dynamic Referencing Sensing Scheme (DRS)	5	11	0	0	1	Large (∼22.31mv)	N/A	N/A	Simulation-based TSMC 40-nm
24	[161]	Read Disturb Detection Scheme (RDD)	5	23	0	0	1	N/A	Small (4ns)	N/A	Simulation-based TSMC 28-nm
25	[121]	Degenerated Cross-coupled Sensing Circuit (DCCSC)	3	33	0	0	1	Large (>495.3mv)	Small (2ns)	Low (0.195pJ)	Simulation-based TSMC 65-nm

2.2.4 Summary of Sensing Schemes and Their Attributes

Based on whether tolerating process variation is of primary importance, some possible inflection points between reliability and performance can occur which in that case, techniques such as [38, 36, 117, 118, 119, 120, 121] are recommended. On the other hand, if tolerating read disturbance is required then [27, 117, 119, 122, 123] techniques can be more effective. Furthermore, if wide sensing margin is a governing requirement then techniques such as [25, 36, 38, 83, 117, 118, 121,

122, 124, 125] could be a preferable alternative, despite increased energy dissipation of some of the approaches. In addition, for robust and reliable designs to reduce write polarization asymmetry, sensing schemes such as [36, 38, 126] are believed to be more promising. Finally, if increasing the yield is the main goal then [25, 123] techniques can be promising candidates for conventional sensing schemes. These suggestions are also listed in Table 2.6, which offer a feasible guide to the circuit designer seeking to trade-off the range of approaches available based on these important parameters of reliability, performance, and energy.

Table 2.6: Sensing Schemes and Their Attributes. [10]

Attribute	**Reference Number**
Process Variation Tolerant	[38, 36, 117, 118, 119, 120, 121]
Read Disturb Reduction	[27, 117, 119, 122, 123]
Wide Sense Margin	[25, 36, 38, 83, 117, 118, 121, 122, 124, 125]
Write Polarization Asymmetry Reduction	[36, 38, 126]
Yield Increase	[25, 123]

2.3 Cache Partitioning Techniques for Energy Reduction

In order to reduce the energy consumption of cache designs, Cache Partitioning is explored in the literature [162, 163, 164, 165]. Two mostly used cache partitioning approaches are introduced in [163], namely, Vertical Cache Partitioning (VCP) and Horizontal Cache Partitioning (HCP). In VCP, the main goal is to increase the cache hierarchy in order to optimize the capacitance of each access. Block buffered cache is an example for VCP presented in [166] where the cache will be accessed only if there is a cache miss, otherwise the data will be accessed from the block buffer which acts as a cache closer to the processor. One of the drawbacks of this approach is that the magnitude of energy saving is highly correlated to the spatial locality of applications and the size of the block.

Furthermore, the main goal of HCP is to provide fine granularity for accessing data via dividing

each cache segment into smaller sub-segments. This will provide flexibility in gating the power to only sub-segments that are being accessed, which will result in energy saving. Cache sub-banking proposed in [162] is an example of HCP where each bank is divided to smaller sub-banks. In this approach, only the sub-bank that holds the data that is currently being accessed is active and all other sub-banks within the bank are inactive. This helps saving unnecessary energy consumed due to accessing the entire bank. As shown in [162], cache sub-banking provides energy savings for instruction and data caches, however block buffering approach is more effective for instruction cache. In particular, more sub-banks can result in more energy saving. Overall, due to the fact that HCP offers more energy saving compared to VCP, herein we have adopted HCP and modified it to fit our approach.

2.4 Hybrid Last Level Cache Design

In recent years, several hybrid spintronic-CMOS cache designs have been proposed to improve the write performance while offering much larger cache capacity with low leakage power [167]. Some of these works such as [168, 169, 170] offer solutions for predicting write-intensive blocks and using migration algorithms, place those write-intensive blocks in the SRAM ways to reduce the energy consumption and delay as well as in-crease the performance. While the approach proposed in [168] only works for core-write operations, the Access Pattern Predictor (APP) proposed in [169] and the Prediction Hybrid Cache (PHC) proposed in [170] cover all different write operations. Additionally, [170] offers dynamic threshold adjustment that allows the threshold of write intensity to change based on the characteristics of the application. Some of the recently published works such as [171] suggest frequent movement of written cache blocks to other STT-MRAM or SRAM lines to reduce the write variance of STT-MRAM lines, however such approaches often result in unnecessary energy consumption, which can lower the performance.

2.5 Non-Volatile SRAM Designs for Power Critical Applications

In recent studies, researchers have exploited the use of emerging devices for NV-SRAM applications. In particular, they have explored designing NV-SRAM using emerging devices such as Resistive Random Access Memory (RRAM), Phase-Change Memory (PCM), STT-MRAM, and SHE-MRAM [17, 41, 42, 43, 44, 45, 46, 47, 48, 49]. However, approaches using PCM and RRAM face challenges such as high programming voltages/currents and high back-up/restore delays, which will exacerbate in scaled technology nodes. With attributes of non-volatility, zero stand-by energy consumption, high endurance, and high density, the Magnetic Tunnel Junction (MTJ) has emerged as a promising alternative post-CMOS technology for embedded memory applications [2, 15, 53].

Many of the recently proposed NV-SRAM designs that utilize MTJ devices take advantage of Spin-Transfer Torque (STT) switching approach for write operations. However, due to the large incubation delay of write operations in the STT approach, SHE-MRAM is recently proposed as a viable alternative for improved performance and energy profiles [15, 108]. Since SHE-MRAM reduces the incubation delay and offers separate read and write paths, a faster, more energy-efficient, and reliable write operation can be achieved compared to STT-MRAM.

2.6 Energy-Aware Quantized Compressive Sesnsing via Adaptive Rate and Resolution

Spectrally sparse signals arise in many applications such as cognitive radio networks, frequency hopping communications, radar/sonar imaging systems, and musical audio signals. In many cases, the sparse components are spread over a wide-band spectrum and need to be acquired without prior knowledge of their frequencies. This is a major challenge in spectrum sensing that is an essential block in any spectrum-aware communication system. Spectrum-aware communication

networks require Radio Frequency (RF) and mixed-signal hardware architectures that can achieve very wide-band but energy-efficient spectrum sensing.

Several architectures have already been proposed for wide-band signal acquisition at rates close to its information rate. These include the Random Demodulator (RD) [172, 173, 174], the Multi-coset Sampler [175] and the Modulated Wideband Converter (MWC) [176, 177]. However, the measurements need to be quantized and encoded to bits for subsequent transmission or processing. In many potential applications the available bit budget is constrained, which suggests a trade-off between the SR and QR. This trade-off is well studied in the Quantized Compressive Sensing [84, 85, 86, 87] literature. Generally speaking, in high observation SNR, fewer but fine-quantized measurements yield better reconstruction quality. However, in the low SNR case, more but coarse-quantized measurements are preferred. As the observation noise varies during acquisition, dynamic optimization of the rate/resolution trade-off is favorable, which is a key innovation of our approach.

So far, several algorithms have been proposed for sparse signal reconstruction from quantized measurements [178, 179]. The extreme case of 1-bit compressive sensing has been extensively studied [180, 181, 182, 183]. In the proposed architecture, the input signal is compared with the level signal, and measurements of the error are acquired. The level signal is adaptively predicted in a feedback loop at the ADC. The idea of acquiring sign measurements of level comparisons was applied in [184] to overcome the scale ambiguity in 1-bit CS reconstruction. In [185, 186], the levels were adaptively varied during acquisition.

There has been some effort to investigate the trade-off between resolution and the rate in a sensing system [87, 187, 188, 189]. However, most of the related works do not include the power constraint in their model. For instance, as in [187], Fisher information can be used to quantize the asymptotic performance of the sensing system. More related to the developed scheme, authors in [87] derived an upper bound for error of quantized compressive sensing without any power constraints.

Most recently, authors in [189] derived Cramer-Rao bound for quantized compressed sensing and investigated the trade-off between SR and QR. However, to the best of our knowledge, there is no work on investigating the rate/resolution trade-off under both power and bandwidth constraints for quantized compressive sampling systems.

2.6.1 Fundamentals of Compressive Sensing

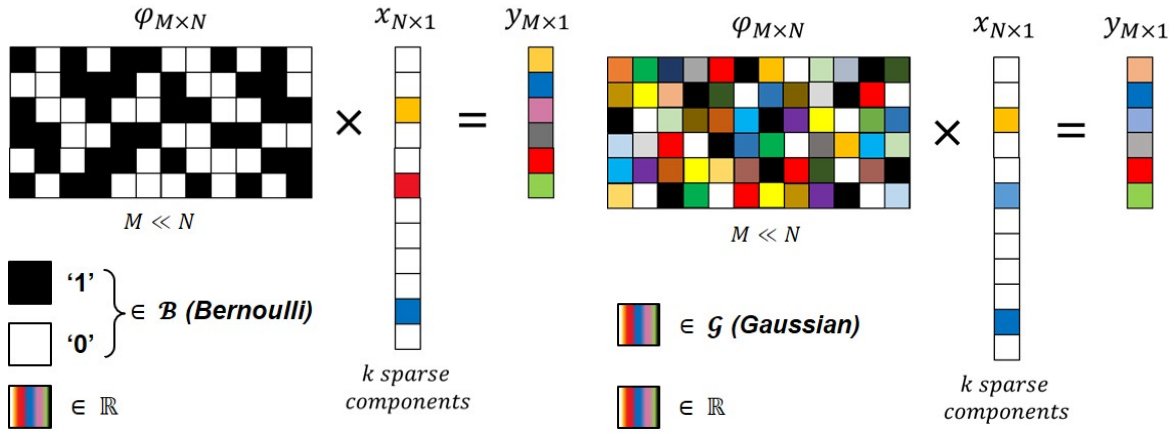
Compressive sensing (CS) is a technique for reconstructing a sparse signal of length N using M measurements, with $M \ll N$. The signal is said to be k-sparse if it has at most k non-zero entries in a given basis; the sparsity rate of the signal is defined as $\frac{K}{N}$. The measurement vector $\mathbf{y} \in \mathbb{R}^M$ is related to the signal vector $\mathbf{x} \in \mathbb{R}^N$ by the measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ through the relation $\mathbf{y} = \mathbf{\Phi x}$. While this is an undetermined system with infinitely many solutions, it has been shown that the signal $\mathbf{x}$ can still be recovered from the M measurements by solving the basis pursuit problem:

$$\hat{\mathbf{x}} = \text{argmin} \left\| \mathbf{x} \right\|_1 \text{ s.t. } \mathbf{y} = \mathbf{\Phi x}, \tag{2.23}$$

where $\left\| \mathbf{x} \right\|_1 = \sum_i |\mathbf{x}|$. It has been shown that $\hat{\mathbf{x}}$ reconstructs the original signal vector if $\mathbf{\Phi}$ satisfies a special condition known as the Restricted Isometry Property (RIP). An $M \times N$ matrix $\mathbf{\Phi}$ satisfies RIP(p) if for any k-sparse vector $\mathbf{x}$:

$$\left\| \mathbf{x} \right\|_p (1 - \delta) \leq \left\| \Phi \mathbf{x} \right\|_p \leq \left\| \mathbf{x} \right\|_p (1 + \delta), \quad 0 < \delta < 1 \tag{2.24}$$

In real-world applications, signals may contain special Regions of Interest (RoI), i.e., subsections of the signal which are more critical to accurately reconstruct than the rest of the signal [91, 94]. Moreover, the sparsity of the signal may be non-uniform. Typically, non-uniform CS measurement matrices utilize Bernoulli and Gaussian distributions as shown in Figure 2.11.



(a) (b)

Figure 2.11: Compressive Sensing with a (a) Bernoulli measurement matrix and (b) Gaussian non-uniform measurement matrix. [9]

Use of a non-uniform measurement matrix allows RoI and parts of the signal with higher sparsity rates to be sampled with higher frequency (i.e., sampled with a sub-matrix containing a higher density of ones). It has been verified that non-uniform measurement matrices satisfy the RIP condition and therefore may be used for sparse signal sampling [91, 94].

Prior work on sparse measurement matrices includes Gilbert and Indyk [190] who described several CS recovery algorithms using sparse measurement matrices and Jafarpour *et al.* [191] who introduced an efficient and low-complexity sparse recovery algorithm. In addition, Kung *et al.* [192] introduced the concept of *neighbor-weighted decoding* as a means of partitioned compressive sensing, i.e. partitioning a signal into blocks which can then be decoded in parallel [96]. Gan [193] proposed to have blocks in the measurement matrix correspond to independent parts of the signal. While [192] and [193] do not take signal non-uniformity into account, Yu *et al.* [194] proposed saliency-based compressive sensing for image processing, where pixels are divided into blocks and the number of measurements applied to a block depends on the saliency of the pixels in that block. Different schemes for non-uniform measurement matrix design have also been reported in [195] and [196].

Recently, researchers have achieved significant performance improvements using sparse signal recovery techniques. Spectrally sparse signals are utilized in many applications such as frequency hopping communications, musical audio signals, cognitive radio networks, and radar/sonar imaging systems [5]. The cornerstone to achieving high-accuracy and efficient CS recovery approaches and non-uniform sampling techniques is the utilization of an adaptive measurement matrix that changes according to the signal characteristics extracted from previous time frames [91, 94]. In most cases, hardware used to implement non-uniform CS sampling and recovery requires a large number of CMOS transistors and incurs significant area overhead and power dissipation [65, 97]. Herein, we propose a low-complexity hardware design to achieve significant power dissipation and area reduction compared to other designs proposed in the literature.

2.6.2 Spectrally-Sparse Signal Model

Similar to [172, 173, 174, 197], we approximate a spectrally sparse signal $x(t)$ by the sum of exponential components $x(t) = \sum_{s \in S} x_s(t)$ in which $S = \{s_1, s_2, ..., s_N\}$ and $x_{s_i}(t+\epsilon) = e^{s_i \epsilon} x_{s_i}(t)$ and assume that only a few number of the components have significant amplitudes $\|x_s(t)_0\|$.

Now consider a frame of the signal as $X_m = \begin{bmatrix} x(m\tau) & x((m-1)\tau) & \cdots & x((m-M+1)\tau) \end{bmatrix}^T$ in which ($\tau = \tau^{(n_f)}$) is the corresponding sample period adapted for the frame and $T = (M-1)\tau$ is the frame length. Let us define Φ by

$$\Phi = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ e^{-s_1\tau} & e^{-s_2\tau} & \cdots & e^{-s_N\tau} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-s_1(M-1)\tau} & e^{-s_2(M-1)\tau} & \cdots & e^{-s_N(M-1)\tau} \end{pmatrix}. \tag{2.25}$$

We can write $X_m = \Phi X'_m$, where $X'_m = \begin{bmatrix} x_{s_1}(m\tau) & x_{s_2}(m\tau) & \cdots & x_{s_N}(m\tau) \end{bmatrix}^T$ is the sparse representation of X_m. Defining a diagonal predictor matrix $P = Diag(e^{Ms_1\tau}, e^{Ms_2\tau}, \cdots, e^{Ms_N\tau})$, we get $X'_m = PX'_{m-M}$, which shows the relation between the sparse representations of the signal for two consecutive frames. This relation later will help us in designing an iterative reconstruction algorithm.

CHAPTER 3: LEVERAGING PROCESS VARIABILITY FOR NON-VOLATILE CACHE RESILIENCE AND YIELD[1]

In this Chapter, in an effort to mitigate and leverage the increased effects of PV in deeply-scaled memory devices, we intriduce the concept of a Self-Organized Sub-bank (SOS) [198]. The proposed SOS approach focuses on leveraging PV in order to provide reliable sensing operation by matching the as-built resource performance with the applications' usage demands while taking the energy budget into consideration. In order to achieve these goals, SOS partitions STT-MRAM data arrays into several sub-banks, which are evaluated using a Power-On Self-Test (POST) phase. The POST assesses the PV impact on the sub-banks, and then, each sub-bank will be assigned an Energy-Aware Sense Amplifier (SA) or a High Resilience SA with regard to a predefined bit error threshold. Based on the results provided in [198], SOS reduces the risk of contaminating the application's data structure by fault propagation as described herein.

In recent years, several hybrid spintronic-CMOS cache designs have been proposed to improve the write performance while offering much larger cache capacity with low leakage power [15, 83, 170, 171, 199, 200]. These methodologies have inspired us to maximize the efficiency of SOS by proposing a dynamic PV-aware and Energy-aware cache block migration policy as a circuit-architecture solution for hybrid memory devices that utilizes a combination of SRAM and STT-MRAM banks in Last Level Cache (LLC).

The proposed approach reorganizes the addresses of the cache blocks within the LLC so that the cache blocks with more frequent write operations are allocated to SRAM cache blocks, whether they are in high-PV impacted regions or not. Additionally, the proposed approach utilizes SOS to transfer the cache blocks with more frequent read operations to STT-MRAM cache blocks that

[1] ©IEEE. Part of this chapter is reprinted, with permission, from [10, 2]

suffer less from PV. As a result, read-intensive operations migrate to low-PV regions of the LLC and sub-banks with less frequent read operations are allocated to high-PV regions of the LLC. We identify herein how an SOS-enabled hybrid cache approach can significantly improve cache utilization and bank accessibility while reducing energy consumption and increasing reliability, since SOS allocates the SA with better energy profile to low-PV regions and the SA with better reliability profile to high-PV regions.

3.1 Proposed Process Variation Immune and Energy Aware Sense Amplifiers for Resistive Non-Volatile Memories

The most common Sense Amplifiers (SAs) which have been studied are Pre-Charge SA (PCSA) [83] and Separated Pre-Charge SA (SPCSA) [117]. While, PCSA offers improved sense latency and power consumption compared to SPCSA, it suffers from increased Bit Error Rate (BER) [117]. SPCSA, on the other hand, offers increased reliability while incurring an acceptable increase in sense latency and power consumption with a negligible area overhead compared to PCSA [117].

In this Section, our focus is performance improvement of the PCSA in terms of Energy Delay Product (EDP) and reliability improvement of SPCSA in terms of Bit Error Rate Reduction (BERR). Additionally, new SA circuits are proposed and simulation result and analysis for the proposed designs are provided. In addition, a new metric is introduced as Sense Error Energy Ratio (SEER) that provides an insight on overall performance and reliability of SAs.

Reducing the amount of resistance in the MTJ devices' paths increases the Sense Margin (SM) and voltage headroom. This reduces the error rate in scaled technology nodes as supply voltage is reduced, which is the case with SPCSA versus PCSA [117]. As shown in Figure 3.1, in PCSA, during the pre-charge stage, **SEN** signal is low, turning **MN2** off while turning **MP0** and **MP3** on.

This will pre-charge the output nodes **OUT** and $\overline{\textbf{OUT}}$ to **VDD**. As a result, **MN0** and **MN1** will turn on while **MP1** and **MP2** are still off. As soon as the sensing stage begins, **MP0** and **MP3** turn off and **MN2** turns on. Thus, based on the difference between **MTJ0** and **MTJ1** resistance, which is determined by the magnetization orientation of their free layer compared to their fixed layer, one of the output nodes begins to discharge more rapidly to **GND**, leading either **MP1** or **MP2** to turn on and charge the other output to **VDD**.

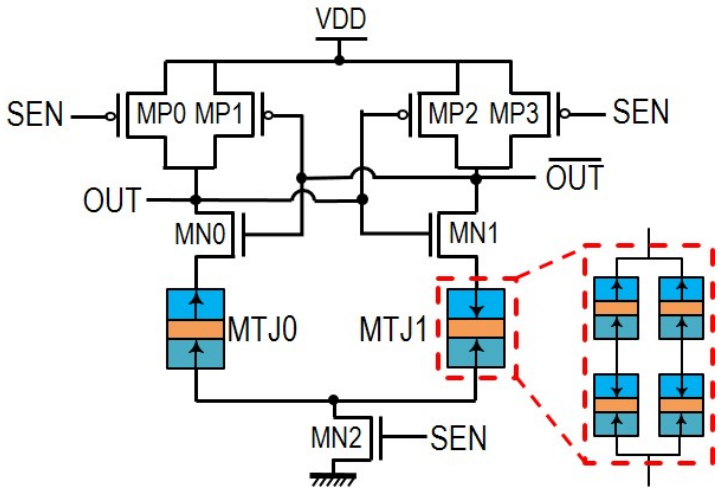


Figure 3.1: PCSA (MTJ1: Reference MTJ). [10]

As depicted in Figure 3.2, in SPCSA, during the pre-charge stage, **SEN** signal is low, turning **MN4** off while turning **MP0**, **MP1**, **MP4**, and **MP5** on. This will pre-charge the output nodes **OUT**, $\overline{\textbf{OUT}}$, **Node0**, and **Node1** to **VDD**. As a result, **MN0** and **MN1** will turn on while **MP2**, **MP3**, **MN2**, and **MN3** are still off. As soon as the sensing stage begins, **MP0**, **MP1**, **MP4**, and **MP5** turn off and **MN4** turns on. Thus, in the secondary discharge path, based on the difference between **MTJ0** and **MTJ1** resistances, one of the two intermediary output nodes, **Node0** or **Node1**, begins to discharge more rapidly to **GND**. This will lead one of the **INV0** or **INV1** output to turn on **MN2** or **MN3**, respectively, which then will cause the primary discharge path to activate and discharge one of the output nodes **OUT** or $\overline{\textbf{OUT}}$ more rapidly to **GND**, resulting in either **MP2** or **MP3** to turn on and charge the other output to **VDD**.

To improve the performance and reliability of PCSA and SPCSA, respectively, Energy Aware SA (EASA) and Variation Immune SA (VISA) are proposed herein as shown in Figure 3.3 and Figure 3.4, respectively. In order to achieve performance and reliability improvements, Transmission Gates (TGs) were utilized to improve the voltage headroom [10]. TGs provide near optimal full-swing switching, and as it has been shown in [201], using TGs, can help reduce the vulnerability to reliability issues caused by PV.

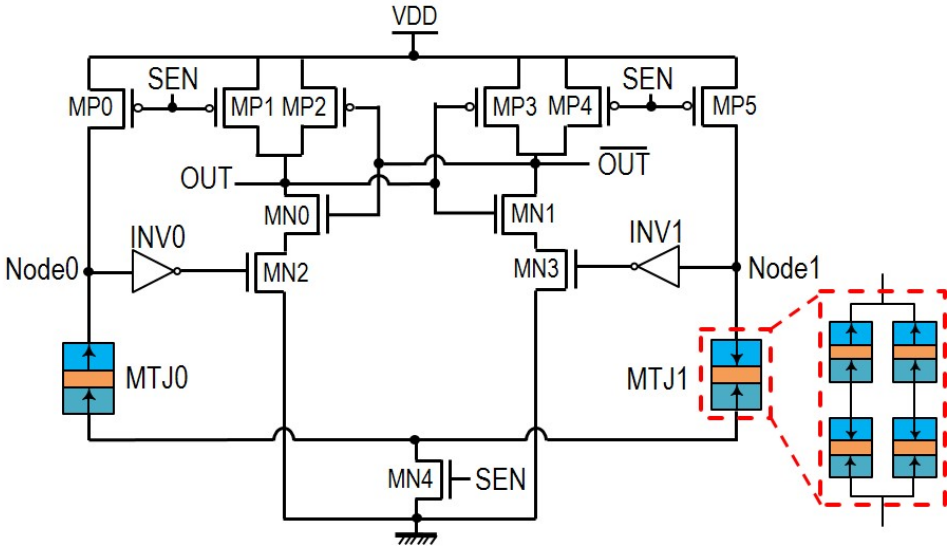


Figure 3.2: SPCSA (MTJ1: Reference MTJ). [10]

In addition, using TGs, as presented in [30], can help reduce the energy consumption by reducing the leakage energy. Thus, **TG0**, **TG1**, and **TG2** are added to improve the performance of the PCSA, as shown in Figure 3.3 [10], and to improve the reliability of SPCSA, as shown in Figure 3.4 [10].

In EASA, during the pre-charge stage, **TG0**, **TG1**, and **TG2** are off, resulting in a reduction of leakage energy from output nodes, **OUT** and $\overline{\textbf{OUT}}$, that are pre-charged to **VDD**. During the sensing stage, **TG0**, **TG1**, and **TG2** turn on and the output nodes start to discharge to **GND**. Based on the resistance difference between the two MTJ branches with regard to the MTJs' states, one of the two output nodes begins to discharge more rapidly, leading the other output to charge to **VDD**.

EASA offers reduced energy consumption by reducing the leakage, however including the TGs on the path of MTJs results in increased resistance of the branches, which will reduce the SM and may result in decreased reliability.

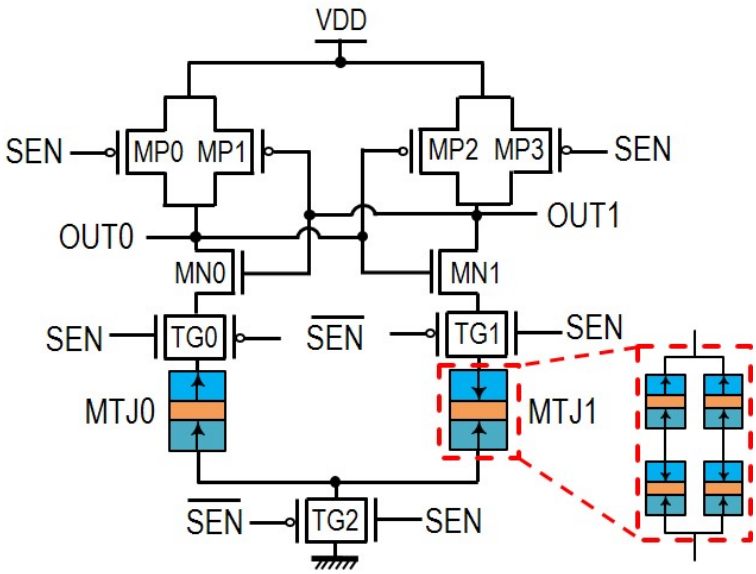


Figure 3.3: EASA (MTJ1: Reference MTJ). [10]

Similar to EASA, in VISA, during the pre-charge stage all the TGs will be turned off, resulting in reduced leakage energy, and both **OUT**, $\overline{\textbf{OUT}}$, **Node0**, and **Node1** will be charged to **VDD**. During the sensing stage, **TG2** will turn on and in the separated part of the SA, based on the resistance difference between the two branches with MTJs with regard to the MTJs' states, one of the two intermediary output nodes, **Node0** or **Node1**, begins to discharge more rapidly. Then, based on the voltage potential of the intermediary outputs, either **TG0** or **TG1** will turn on faster and one of the main branches of the SA begins to discharge quicker, resulting in the output node of that branch to drop and charge the other branch's output node to **VDD**. **INV0** and **INV1** are used to amplify the voltage difference of **Node0** and **Node1** of the SA. Using **TG0** and **TG1** and utilizing **Node0** and **Node1** as well as their amplified value, the authors have reduced the effects of PV by reducing the chance of failure due to device mismatch in the Inverters. Furthermore, by utilizing **TG2**, energy consumption is reduced due to the reduction in the leakage energy [10].

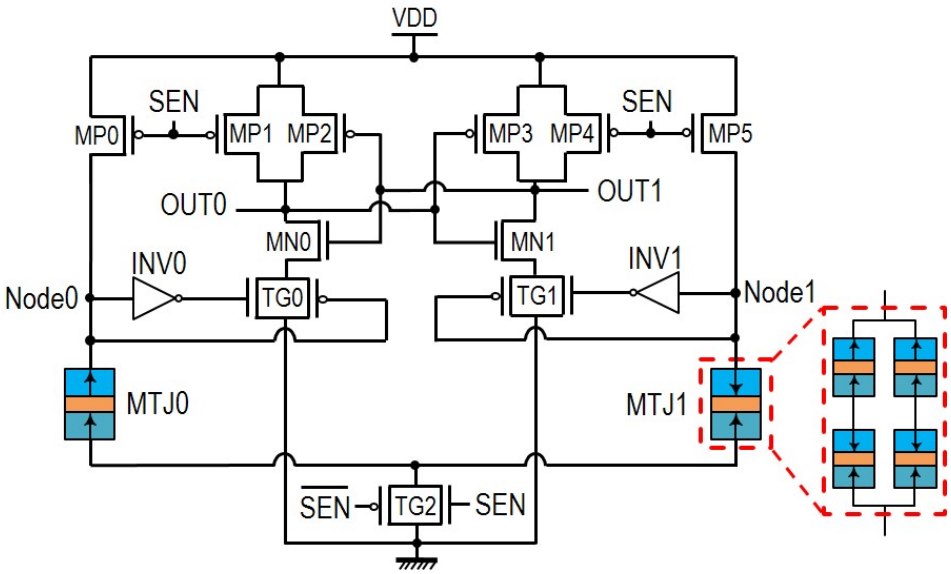


Figure 3.4: VISA (MTJ1: Reference MTJ). [10]

As shown in Figure 3.1, Figure 3.2, Figure 3.3, and Figure 3.4, an alternative referencing configuration is used to further improve the reliability of the SAs. Using $(MTJ_P + MTJ_{AP})||(MTJ_P + MTJ_{AP})$ configuration for the reference MTJ, referred to as **MTJ1** in Figure 3.1, Figure 3.2, Figure 3.3, and Figure 3.4, a reference value of $(MTJ_P + MTJ_{AP})/2$ is achieved, which provides increased SM [10, 120].

3.1.1 Circuit-Level Results and Analysis

Extensive circuit-level simulation results and analysis are provided in this Section. The 22nm Predictive Technology Model (PTM) CMOS library [202] is used alongside the MTJ model used in [15] to calculate the power and performance of a 1-bit MSA and ASA. We have utilized the approach proposed in [15] to model the behavior of STT-MRAM devices, in which a Verilog-AMS model is developed and leveraged in a SPICE circuit simulator to validate the functionality of the designed circuits. Table 3.1 lists the technology parameters and PV values used in the circuit simulations.

Table 3.1: Circuit Simulation Technology Parameters. [10]

Parameter				**Value**	**Std. Dev. (σ)**
PMOS	V_{th} (Threshold Voltage)			$460mV$	10%
	Width/Length $(W/L)_P$			2 & 4	1%
NMOS	V_{th} (Threshold Voltage)			$500mV$	10%
	Width/Length $(W/L)_N$			1 & 2	1%
MTJ	STT-MTJ Area	Data MTJ (MTJ0)		$(\frac{\pi}{4}) \times 40 \times 40\ nm^2$	1%
		Reference MTJ (MTJ1)	MTJ_{AP}	$(\frac{\pi}{4}) \times 30 \times 30\ nm^2$	1%
			$(MTJ_P + MTJ_{AP})/2$	$4 \times [(\frac{\pi}{4}) \times 40 \times 40\ nm^2]$	1%
	t_{ox} (Oxide Thickness)			$0.85nm$	1%
	TMR (Tunnel Magneto Resistance)			100%	1% & 10%
	R×A (Resistance Area Product)			$5\Omega.\mu m^2$	N/A
	ϕ (Potential Barrier Height)			$0.4V$	N/A
	α (Damping Factor)			0.01	N/A
Nominal Voltage (Vdd)				$1.0V$	N/A
SEN Signal Period (T)				$1.0ns$	N/A

All PMOS and NMOS transistors are considered minimum size except transistors used in **INV0** and **INV1** shown in Figure 3.2 and Figure 3.4. Since **INV0** and **INV1** are vital to the reliability of the circuit, we have optimized the size of their transistors to maintain width (W) to length (L) ratio (W/L) of 4 to provide reliable functionality. All of the designs provided in this chapter are simulated and analyzed in a case where no PV is present and in a case where PV is present. Monte Carlo (MC) simulation methods are utilized to model the PV. Table 3.2 lists the results for delay, power consumption, and Energy Delay Product (EDP) where no PV is present and the $TMR = 100\%$ with $MTJ_P = 3.2K\Omega$ and $MTJ_{ref} = 5.7K\Omega$. Table 3.3 lists similar results with $MTJ_P = 3.2K\Omega$ and $MTJ_{ref} = (MTJ_P + MTJ_{AP})/2 = 4.8K\Omega$.

In order to further investigate the effects of PV on the SAs, $10,000$ MC simulations were performed on a single bit memory cell, considering different standard deviations for the CMOS threshold voltage as well as MTJ MgO thickness and surface area. During the simulation, values of V_{th}, W, and L of the CMOS transistors vary in the netlist based on a Gaussian distribution having a mean equal to the nominal model card for PTM and σV_{th} as provided in [203]. For the MTJ variation, the model provided in [15] was used to find the effects of variation on MTJ devices.

Table 3.2: Simulation Results with no PV considering $MTJ_{ref} = 5.7K\Omega$. [10]

Design	Area (Device Count)			Anti-Parallel			Parallel		
	PMOS	NMOS	MTJ	Delay (ps)	Power (μW)	EDP (fJ*ps)	Delay (ps)	Power (μW)	EDP (fJ*ps)
PCSA	4	3	2	17.79	0.7267	12.93	16.86	0.7026	11.85
SPCSA	8	5	2	27.26	2.2960	62.59	25.44	2.2690	57.72
EASA	7	5	2	24.92	0.2445	6.09	27.24	0.2205	6.01
VISA	11	7	2	25.38	1.8560	47.11	23.29	1.7990	41.90

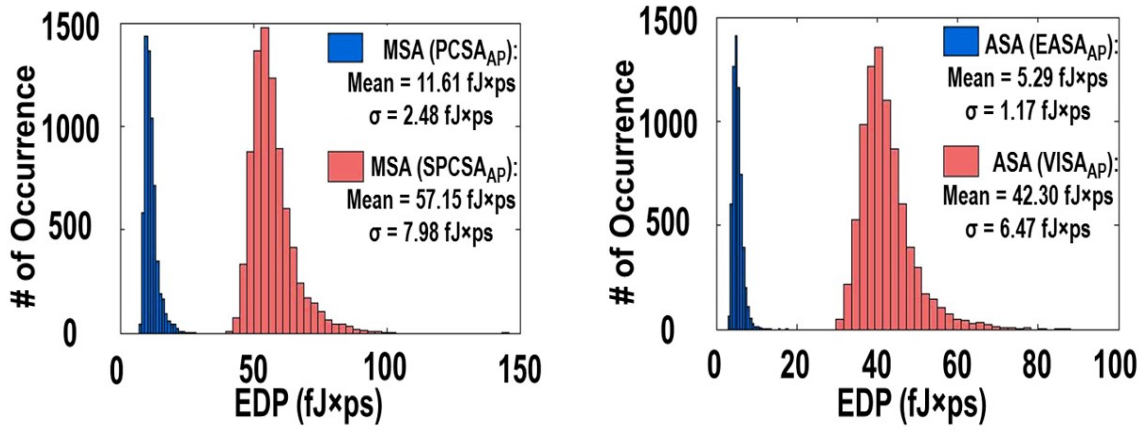
Table 3.3: Simulation Results with no PV considering $MTJ_{ref} = 4.8K\Omega$. [10]

Design	Area (Device Count)			Anti-Parallel			Parallel		
	PMOS	NMOS	MTJ	Delay (ps)	Power (μW)	EDP (fJ*ps)	Delay (ps)	Power (μW)	EDP (fJ*ps)
PCSA	4	3	2	15.56	0.7139	11.11	17.80	0.7026	12.63
SPCSA	8	5	2	24.72	2.2710	56.14	26.51	2.2770	60.36
EASA	7	5	2	22.73	0.2325	5.28	28.38	0.2274	6.45
VISA	11	7	2	22.68	1.8150	41.16	24.28	1.7990	43.68

In [56], authors have fitted the experimental data measured in [204] to an exponential curve to obtain the effect of oxide thickness (t_{OX}) variation on TMR values. The relation between the t_{OX} and TMR is expressed by following equation, $TMR = K1 - \frac{K2}{K3}(1 - e^{(-K3.t_{OX})})$, where $K1 = -8109.436$, $K2 = -37145$, and $K3 = 4.45$ are fitting parameters. We have considered a speculative variation of 1% for oxide thickness, which can result in a range of 1% to 10% TMR variations. This can cover the full range of possible variations enabling a comprehensive PV analysis.

Due to structural limitations of MTJ devices, the TMR ratio is considered 100% as the baseline design herein [83, 117, 205]. Based on the results listed in Table 3.2 and Table 3.3, ASA-EASA provides, on average, 2-fold reduced EDP over MSA-PCSA, 7-fold reduced EDP compared to ASA-VISA, and 9-fold reduced EDP compared to MSA-SPCSA. On the other hand, ASA-VISA provides, on average, 1.4-fold reduced EDP compared to MSA-SPCSA. Figure 3.5(a) depicts the

EDP distribution of MSA-PCSA and MSA-SPCSA for sensing **AP** state, respectively. Figure 3.5(b) exhibit similar results for ASA-EASA and ASA-VISA.



(a) (b)

Figure 3.5: EDP of sensing "1" with $MTJ_{ref} = 5.7K\Omega$ and $TMR = 100\%$, $\sigma TMR = 10\%$ for(a) MSA in PCSA and SPCSA mode and (b) ASA in EASA and VISA mode. [2]

Bit Error Rate (BER) is calculated based on the number of wrong output bits divided by all the sensing operations performed for both P and AP states. The values provided in Figure 3.6 and Figure 3.7 are the average BER values of P and AP states' sensed output obtained from simulating a single bit cell. Figure 3.6(a) lists the $10,000$ MC simulation results, where $MTJ_P = 3.2K\Omega$, $MTJ_{ref} = 5.7K\Omega$, and $MTJ_{AP} = 6.4K\Omega$ for $TMR = 100\%$. Considering 10% variation on TMR, the results show that on average ASA-VISA provides 8.3% reduced BER compared to ASA-EASA, 6.1% reduced BER compared to MSA-PCSA, and 1.6% reduced BER compared to MSA-SPCSA considering $TMR = 100\%$. The results also exhibit further reliability improvement considering $TMR = 150\%$ where ASA-VISA provides 10.6% reduced BER compared to ASA-EASA, 7.2% reduced BER compared to MSA-PCSA, and 1.2% reduced BER compared to MSA-SPCSA.

Furthermore, Figure 3.6(b) shows $10,000$ MC simulation results, where $MTJ_P = 3.2K\Omega$, $MTJ_{AP} =$

$6.4K\Omega$, and $MTJ_{ref} = (MTJ_P + MTJ_{AP})/2 = 4.8K\Omega$ for $TMR = 100\%$. Considering 10% variation on TMR, the results exhibit that on average ASA-VISA provides 10.3%, 5.7%, and 1.3% reduced BER compared to ASA-EASA, MSA-PCSA, and MSA-SPCSA respectively, considering $TMR = 100\%$. The results also indicate additional improvement of reliability for $TMR = 150\%$ where ASA-VISA provides 10.7%, 7.2%, and 1.1% reduced BER compared to ASA-EASA, MSA-PCSA, and MSA-SPCSA respectively.

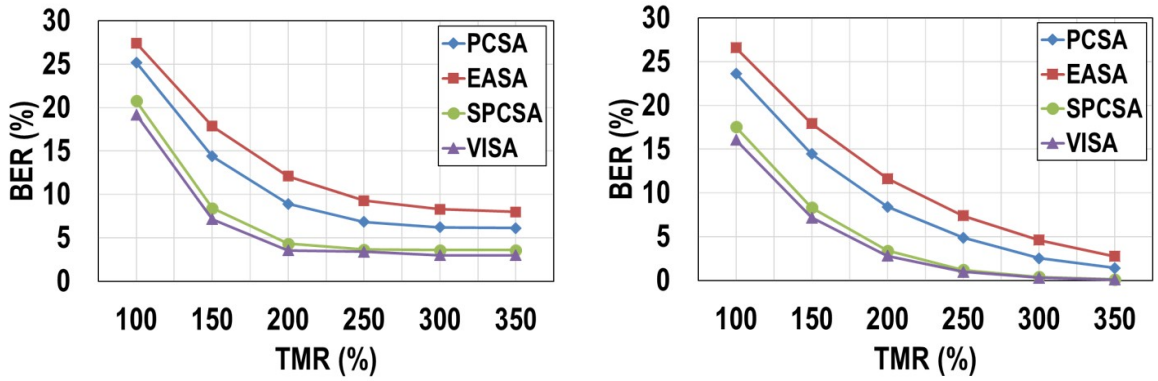
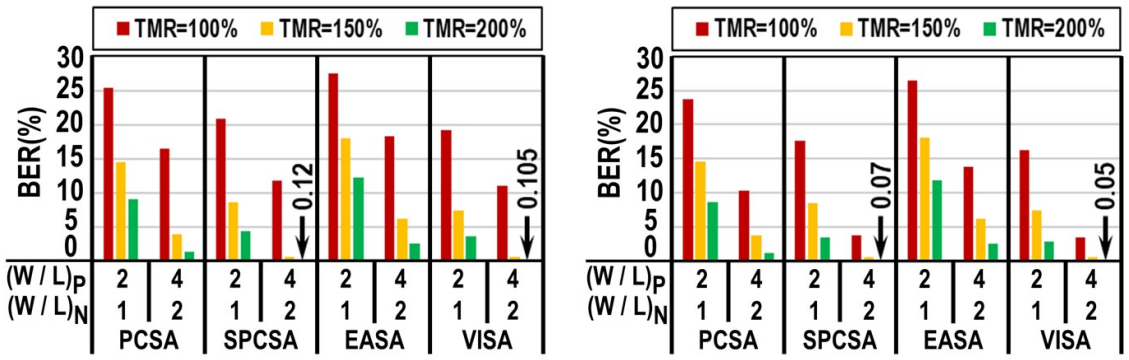


Figure 3.6: Average BER for $\sigma TMR = 1\%$ & 10%, $\sigma V_{th} = 10\%$, $MTJ_P = 3.2K\Omega$, (a) $MTJ_{ref} = 5.7K\Omega$ and (b) $MTJ_{ref} = (MTJ_P + MTJ_{AP})/2$. [2]

Figure 3.7 shows the $10,000$ MC simulation results considering $(W/L)_P$ ratio of 2 and 4, and $(W/L)_N$ ratio of 1 and 2. The results show that in TMR of 100% on average SA designs with increased transistor sizes provide 8.8% and 13.2% reduced BER for $MTJ_{ref} = 5.7K\Omega$ as shown in Figure 3.7(a) and $MTJ_{ref} = (MTJ_P + MTJ_{AP})/2$ as shown in Figure 3.7(b), respectively, compared to minimally-sized transistors. The results also exhibit further reliability improvement considering TMR of 150% where SAs having increased transistor sizes provide 9.3% and 9.4% reduced BER for $MTJ_{ref} = 5.7K\Omega$ as shown in Figure 3.7(a) and $MTJ_{ref} = (MTJ_P + MTJ_{AP})/2$ as shown in Figure 3.7(b), respectively, compared to SAs with minimum transistor sizes. Additionally, considering TMR of 200% further improvements in reliability is observed.

The BER for SAs with increased transistor sizes is reduced by 6.3% and 5.7% on average for $MTJ_{ref} = 5.7K\Omega$ as shown in Figure 3.7(a) and $MTJ_{ref} = (MTJ_P + MTJ_{AP})/2$ as shown in Figure 3.7(b), respectively, compared to SAs with minimum transistor sizes. It can be observed that by optimizing the reference MTJ and using $(MTJ_P + MTJ_{AP})/2$ configuration, the BER can be decreased by 8.9% on average for a TMR of 100% due to increases in the SM for both P and AP states of the MTJ. The distribution of P and AP states of the MTJs and the reference MTJ is depicted in Figure 3.8. Based on the results of MC simulations, it is clear that the larger TMR values results in an increased SM, which reduces the impact of PV.



(a) (b)

Figure 3.7: Average BER for $\sigma TMR = 1\%$ & 10%, $\sigma V_{th} = 10\%$, $(W/L)_P = 2$ & 4, $(W/L)_N = 1$ & 2, $MTJ_P = 3.2K\Omega$, (a) $MTJ_{ref} = 5.7K\Omega$ and (b) $MTJ_{ref} = (MTJ_P + MTJ_{AP})/2$. [2]

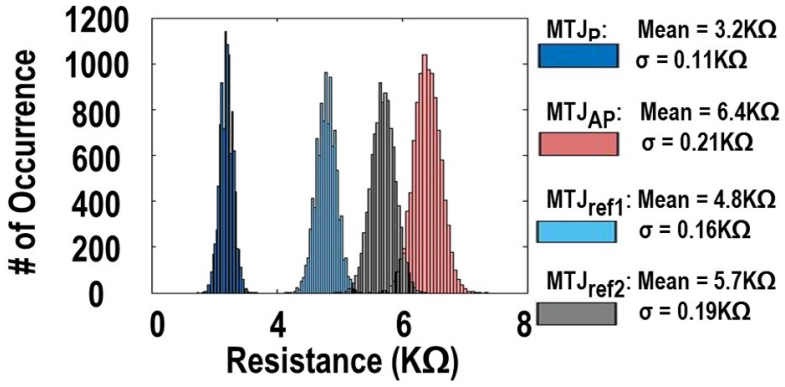


Figure 3.8: Distribution of P and AP states of the MTJ devices, $MTJ_{ref1} = 4.8K\Omega$, and $MTJ_{ref2} = 5.7K\Omega$. [2]

In order to be able to compare the reliability and performance of different SAs more comprehensive, $SEER$ metric is introduced herein, which represents the ratio of $BERR$ to average EDP as demonstrated in (3.1). $BERR$ is calculated as shown in (3.2). $SEER$ will enable the designers to effectively assess the most appropriate SA for their need based on whether they are seeking reliability or energy efficiency. Any increase in $BERR$ or decrease in EDP will cause the $SEER$ to increase. As a result, larger values of $SEER$ imply increased reliability and performance and on the contrary, small values of $SEER$ imply decreased reliability and performance.

$$SEER((fJ \times ps)^{-1}) = \frac{BERR\,(Design\ X)}{EDP\,(Design\ X)} \tag{3.1}$$

$$BERR(\%) = 100 - BER\,(Design\ X\ for\ desired\ TMR) \tag{3.2}$$

Qualitative performance comparison of the designs and metrics proposed in this Section is listed in Table 3.4. Based on physical layout design of PCSA, EASA, SPCSA, and VISA shown in Figure 3.9(a), Figure 3.9(b), Figure 3.9(c), and Figure 3.9(d) respectively, it is clear that the proposed EASA and VISA designs offer small area overhead compared to their counterparts, considering overall size of the memory.

Table 3.4: Qualitative performance comparison of Sense Amplifier designs discussed herein. [10]

Design	**Delay**	**Power**	**EDP**	**BERR**	**SEER (Iso-BERR)**	**SEER (Iso-EDP)**
PCSA [83]	✓✓✓	✓	✓✓	- -	✓✓	-
EASA (Proposed herein)	✓	✓✓✓	✓✓✓	- -	✓✓✓	- -
SPCSA [117]	✓	- -	- -	✓✓	- -	✓✓
VISA (Proposed herein)	✓	-	-	✓✓✓	-	✓✓✓

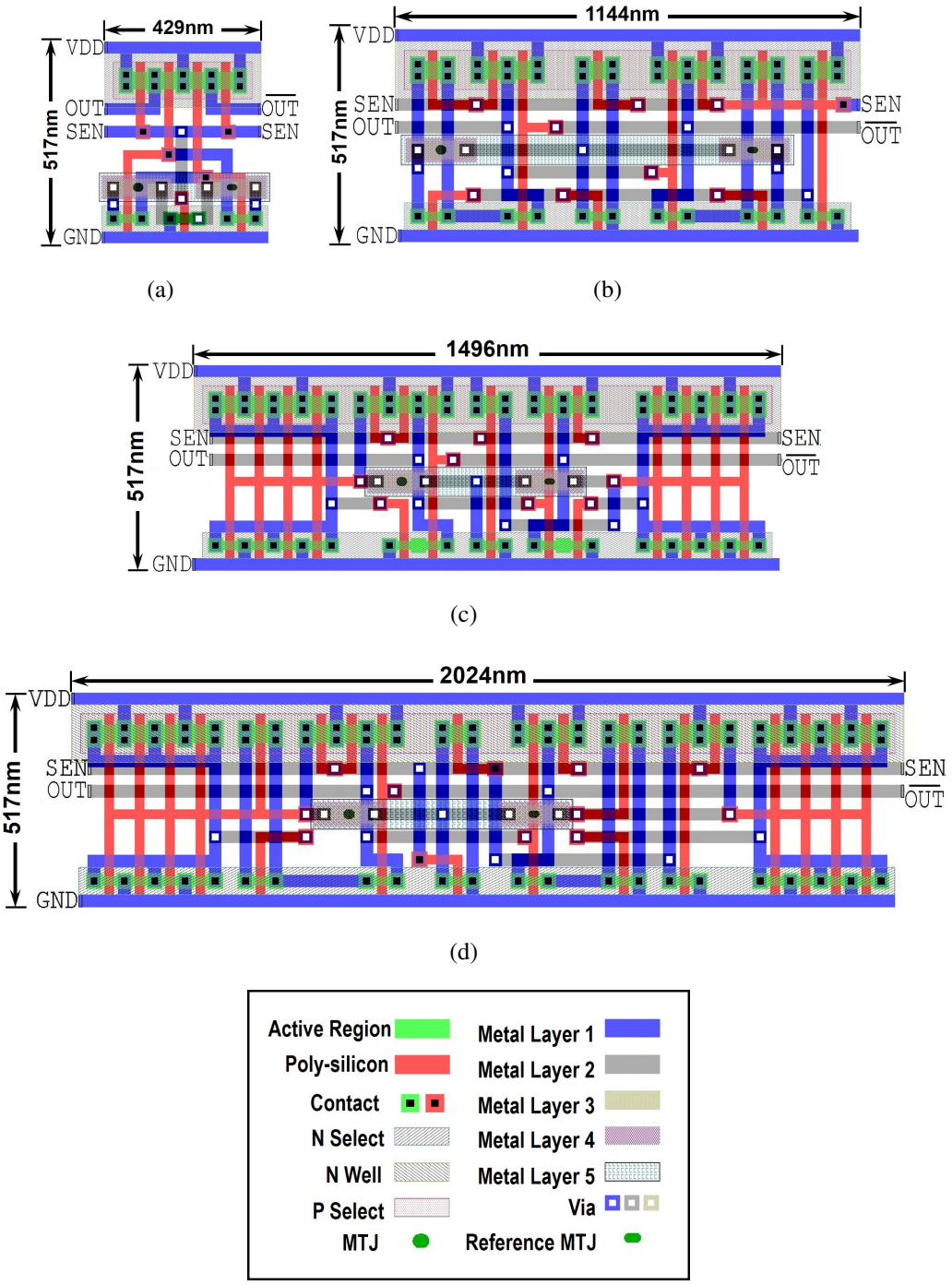


Figure 3.9: (a) PCSA Layout, (b) SPCSA Layout, (c) EASA Layout, (d) VISA Layout, and (e) Layout Legend. [10]

3.2 Proposed SOS Schematic for SA Assignment

By combining PCSA and SPCSA, the Merged Sense Amplifier (MSA) [198] is realized to utilize each SA's properties to increase performance and reliability. In order to improve energy efficiency of MSA proposed in [198], the selectors **MUX1** and **MUX2** are included in order to make sure only one SA is operating and to avoid unnecessary energy consumption by gating the **SEN** signal of the offline SA as shown in Figure 3.10.

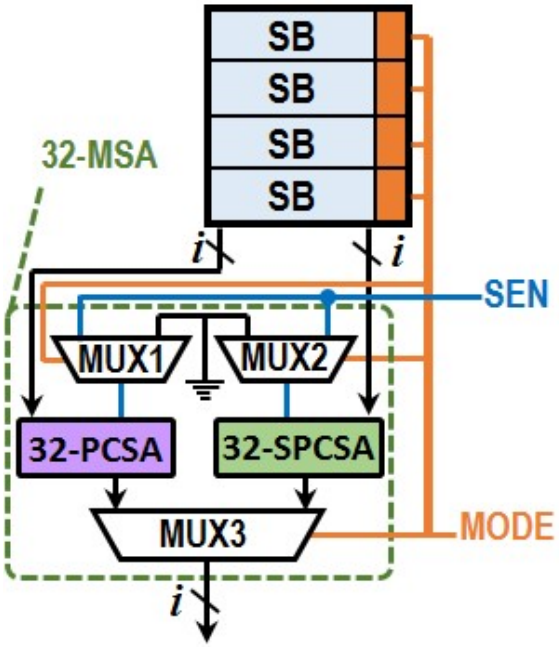


Figure 3.10: MSA (SB: Sub-Bank). [2]

Herein, we propose an alternative for MSA that further improves energy consumption and reliability due to PV. The Adaptive Sense Amplifier (ASA), as shown in Figure 3.11, has a functionality similar to MSA described in [198]. However, by utilizing EASA, as shown in Figure 3.3 [10], and VISA, as shown in Figure 3.4 [10], instead of PCSA and SPCSA, it can achieve better energy and reliability profiles, respectively [10]. Like MSA, **MUX1** and **MUX2** are included in ASA to reduce energy consumption by gating the **SEN** signal of the offline SA so that only one SA is oper-

ating. SPCSA and VISA both increase reliability by reducing the amount of resistance in the MTJ read paths, which increases the SM and voltage headroom of the SA, resulting in a more reliable sensing. Increasing voltage headroom is an important issue in scaled technology nodes since the supply voltage is reduced to 1 volt or below, and even a small voltage drop can result in a sensing error [117].

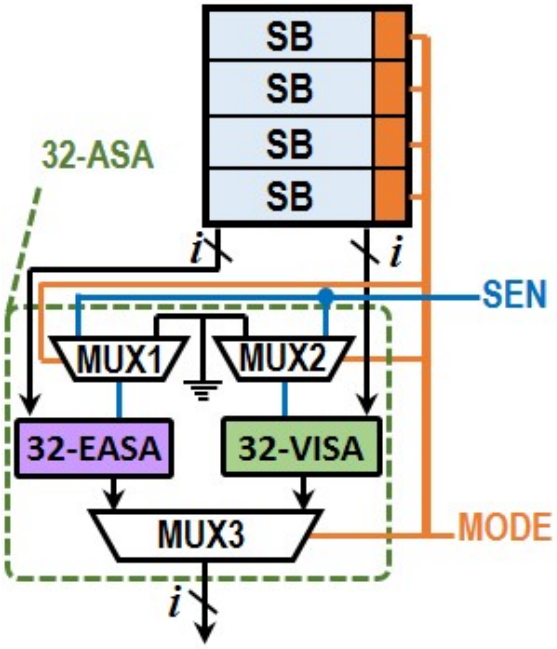
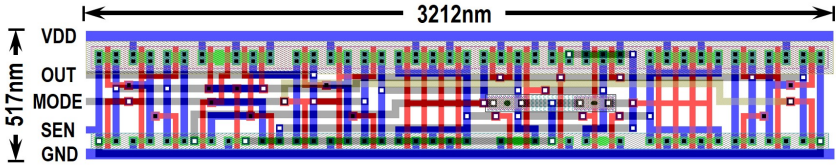


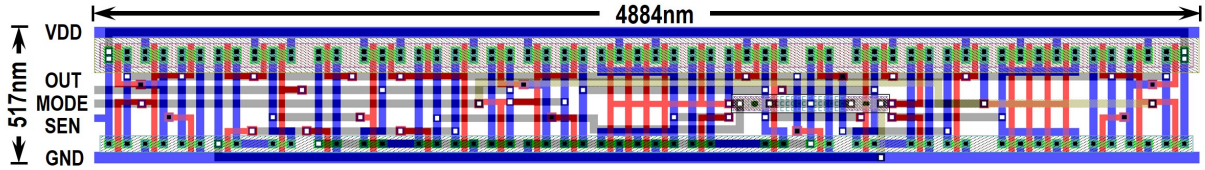
Figure 3.11: ASA (SB: Sub-Bank). [2]

The schematic of different SOS designs is depicted in Figure 3.10 and Figure 3.11, and the process for assigning the preferred SA to each Sub-Bank (SB) is shown in Algorithm 1 for MSA and ASA. As shown in Algorithm 1, SOS starts with a POST function. In both SA designs, after the POST function, an analyzer function is called to determine the preferred SA for that particular SB. A select input is used in the circuit called **MODE** to choose between the two SAs based on the assigned bit set value as shown in Figure 3.10 and Figure 3.11. If the logic 1 is assigned to input **MODE**, then the circuit will operate in PCSA mode in MSA or EASA mode in ASA. On the other hand, if logic 0 is assigned to **MODE**, it will change the operation of the SA to SPCSA mode in

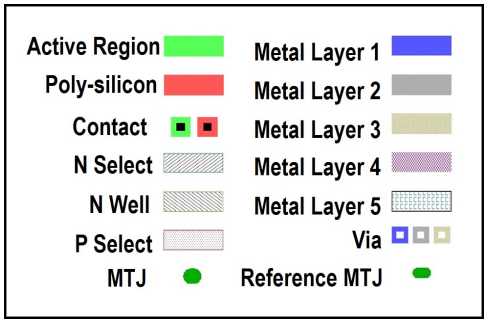
MSA or VISA mode in ASA. As discussed earlier, in both MSA and ASA, the **SEN** signal is gated for the SA that is not in use to increase energy saving of the SA. In other words, only one SA will turn on, and the other SA's **SEN** signal will be connected to **GND**, which results in **OUT** and $\overline{\textbf{OUT}}$ to be 1 at all times. ASA offers improved reliability and performance, while maintaining a small footprint of $2.5\mu m^2$ as depicted in Figure 3.12(a). Additionally, ASA incurs 0.5-fold, 10.4-fold, 2.3-fold, 3.3-fold, and 1.4-fold area overhead compared to the new MSA shown in Figure 3.12(b), PCSA [10], SPCSA [10], EASA [10], and VISA [10], respectively.



(a)



(b)



(c)

Figure 3.12: (a) ASA Layout, (b) MSA Layout, and (c) Layout Legend. [2]

Algorithm 1: SOS Approach to Assign Preferred SA to Sub-bank. [2]

```
1  Function SOS() /*SOS Approach for SA Assignment*/
2  for ∀ cache line ∈ LLC do
3      for ∀ sub − bank ∈ cache line do
4          begin
5              POST() /*Power-On Self-Test*/
6              Analyzer() /*Evaluate the correctness of the outputs*/
7  Function POST() /*Power-On Self-Test*/
8  begin
9      set SEN = 1 /*start the discharge and evaluation stage*/
10     if output != expected-value then
11         ++number-wrong-outputs /*increment number of wrong outputs*/
12     set SEN = 0 /*keep the sense signal in pre-charge stage*/
13 Function Analyzer() /*Evaluate the correctness of the outputs*/
14 begin
15     if number-wrong-outputs > threshold then
16         set MODE = 0 /*assign MSA-SPCSA or ASA-VISA to sub-bank*/
17         /*MUX3 takes sensed data from MSA-SPCSA or ASA-VISA to output*/
18         /*MUX1 selects SEN signal to activate MSA-SPCSA or ASA-VISA and deactivate MSA-PCSA or
             ASA-EASA*/
19     else
20         set MODE = 1 /*assign MSA-PCSA or ASA-EASA to sub-bank*/
21         /*MUX3 takes sensed data from MSA-PCSA or ASA-EASA to output*/
22         /*MUX2 selects SEN signal to activate MSA-PCSA or ASA-EASA and deactivate MSA-SPCSA or
             ASA-VISA*/
```

3.2.1 Extracting the PV Parameters

In our PV modeling process, we assume that the cache tag and peripherals (e.g., row decoder, column decoder, row buffer and SAs) are fabricated at the CMOS layer while memory cells are realized through MTJ devices. Since the MTJs are vertically stacked on top of the CMOS layer and these components are tightly coupled to realize the function of STT-MRAM, the SM varies readily based on the effect of PV on that particular region of the die. Accordingly, we have utilized the approach mentioned in [198] to extract the PV parameters. One PV map is randomly selected from a large pool of PV maps with a resolution of one million $(1,000 \times 1,000)$ sample points that are generated utilizing the approach presented in [198]. The degree of variation is shown by a range of colors. Each color corresponds to a specific value of sample points as shown in Figure 3.13.

In our simulation, we consider the amount of PV for each site based on the location of the LLC components within the floorplan and their associated sample points. Thus, the generated maps are relatively accurate estimation of the impact of PV on the read SM of each SB.

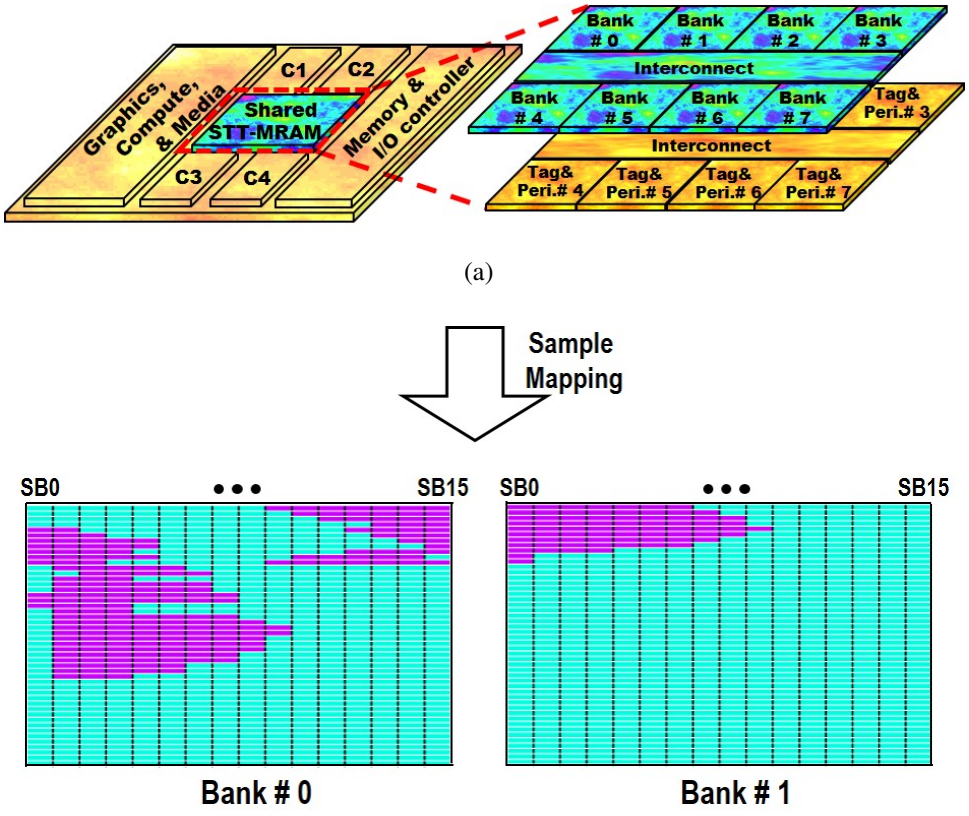


Figure 3.13: (a) PV map of a 4-core CMP, and (b) Determining preferred SA based on post-fabrication SB PV resiliency assessment. [2]

3.2.2 Power On Self-Test (POST)

As shown in Figure 3.13, the cache bank floorplan of the STT-MRAM layer is superimposed on the map. In our SOS approach, each cache bank is partitioned into 16 SBs. The size of each SB is matched with the word size to maintain the energy consumption of the tag to be as low as possible, e.g. 32 bits in our case study. We consider one additional bit per SB to identify the

preferred SA for that particular SB during post-fabrication resiliency assessment to PV. The POST phase is basically a March Test that targets PV-induced faults in STT-MRAM [206]. Similar to the widely-used March test, during the POST operation, first we write 0 to all memory cells, then we read the memory cells and then we write 1 to all memory cells and then read again. Based on the outcome of all read operations we will be able to find the number of erroneous outputs and based on that, it is possible to recognize the high-PV regions. We assume the proposed SRAM March Test with $O(n)$ test length can be utilized for our purpose because the tag and peripherals of STT-MRAM are considered to be implemented in the CMOS layer. Thus, variation-induced delay faults in both SRAM and STT-MRAM manifests itself as the same fault model as an insufficient pre-charge period, insufficient discharge and evaluation period, insufficient amplify time, disturbance of sensing operation, and simultaneously activation of multiple word lines.

In this regard, PV-aware March Test examines all STT-MRAM data arrays and performs a sequence of operations (e. g., exhaustive pair-wise address transitions) to identify PV-induced delay faults in each cell [206]. If the error rate of the impacted STT-MRAM cells in a SB exceeds the predefined threshold, the extra bit is set to 0 indicating that an array of reliable SAs are re-quired for sensing the data of this SB. Otherwise, the extra bit is set to 1, which indicates that an array of low-power SAs offering reduced delay and power consumption can be considered for that particular SB. Since POST is a one-time operation, it will not impact the performance of the memory as a whole, resulting in a negligible overhead.

3.2.3 Fault Models Associated with Sensed Data

In the PARSEC suite [207], when considering the presence of PV, around 27.5% of the sensed data when utilizing a STT-MRAM based LLC has the potential to be incorrect, 6% of which will be overwritten prior to being used by the processor or to be committed to the main memory,

on average. Despite the fact that 6% might not be significant, a substantial portion of incorrectly sensed data requires handling before manifesting themselves as wrong outputs, application crashes, or prolonged program executions [29]. To be specific, we classify the outcomes of SA operation to the following categories for broad adaption according to [198]:

- True Data Sensing (TDS): The sensed data value is identical to the value stored in the STT-MRAM cell.
- Vulnerable False Data Sensing (VFDS): The sensed data value differs from the value stored in the STT-MRAM cell, which propagates out of cache to be either used by the process or committed to oth-er levels of memory [29].
- Non-Vulnerable False Data Sensing (NVFDS): The sensed data value differs from the value stored in the STT-MRAM cell, however the replica copy of the sensed false data in the upper levels of cache will be overwritten by a write operation prior to being used. During a block eviction, replica data becomes written back to the lower levels of cache because it is a dirty victim block. Thus, this benign fault does not threaten the semantic correctness.

Based on these categories, the experiment concentrates on the faults that are caused by incorrectly sensed data rather than alternative fault models that can impact the stored value in STT-MRAM cells [208].

3.2.4 Proposed Hybrid SRAM and STT-MRAM LLC Design

Figure 3.14 illustrates the scheme of a hybrid 8-way set associative SRAM and STT-MRAM LLC design, where way-0 and way-1 are implemented within SRAM-based banks while way-2 through way-7 are built in STT-MRAM-based banks. This configuration is selected based on our exper-

imental results, whereby the average number of write-intensive blocks in each set was approximately 2 across all workloads. Since the peripherals required for read and write operations in NVM arrays occupy a relatively larger portion of the cache footprint than peripherals required by SRAM arrays, it is beneficial to build the tag array with SRAM cells. Thus, we assume that the entire tag array is built with SRAM. With cache tags residing in CMOS, erroneous SRAM-based tags lie outside of the scope of this study.

Unlike conventional cache design approaches, where the tag and data array are accessed simultaneously to reduce access latency while incurring significant power overhead, we propose to split the cache access into two stages similar to the work presented in [209], but with adjustments in favor of high SOS throughput. If LLC is accessed with a read operation, the tag array and all STT-MRAM banks are accessed in parallel. Thus, assuming that data is found in STT-MRAM banks, the unnecessary accesses to SRAM banks can be skipped. Upon a LLC miss on STT-MRAM banks, but hit on a tag corresponding to a SRAM bank, the associated SRAM data array of the bank in LLC is accessed. Even though this mechanism incurs additional latency if the data is stored in SRAM banks, we argue that this incident occurs rarely since our insertion/migration policy maintains the read-dominant cache blocks in STT-MRAM banks while write-intensive blocks are transferred to SRAM banks.

If the cache set is accessed by a write operation, the tag arrays and SRAM banks are searched in the first stage. If the data is not found in SRAM banks but found in a STT-MRAM bank, the corresponding STT-MRAM banks is accessed in the next stage. Unlike the insertion strategy in [209] where SRAM banks are selected for inserting fetched data from memory upon an LLC miss, our insertion policy allocates a way from either SRAM or STT-MRAM banks according to the miss type. In particular, the SRAM and STT-MRAM banks are allocated upon an LLC write miss and read miss, respectively.

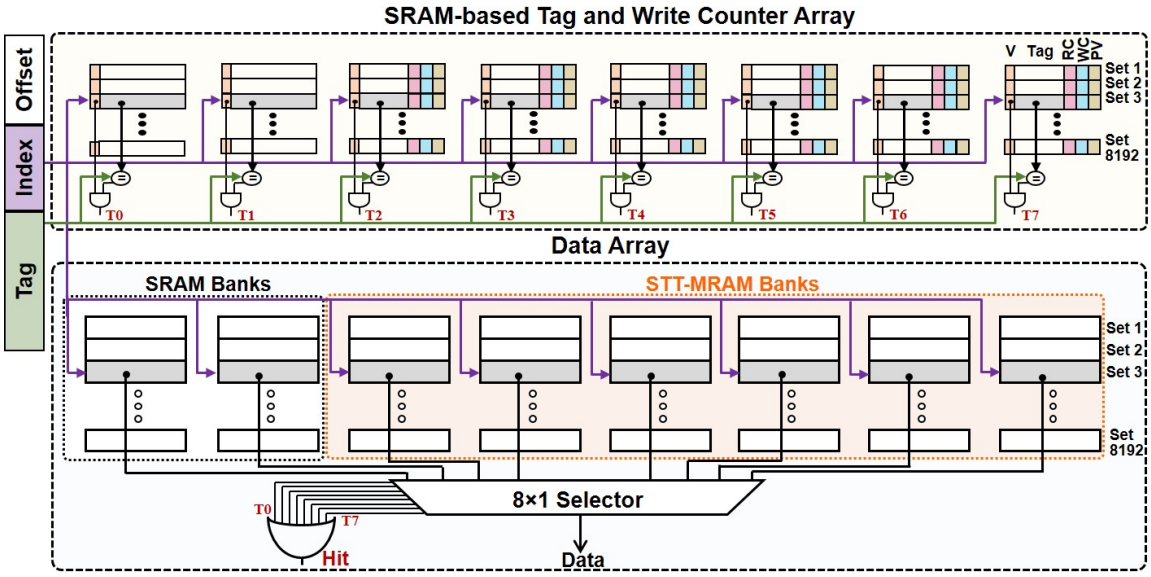


Figure 3.14: The scheme of hybrid 8-way set associative SRAM and STT-MRAM cache design, whereby each bank stores a way. In the above configuration, two SRAM-based banks and six STT-MRAM based banks are illustrated. [2]

Based on our observation presented in [210], a portion of a workload might be re-executed several times, indicating that the read-intensive cache blocks which were brought to LLC once, transferred to low-PV impacted region of a set, and finally evicted need to be re-allocated from low-PV impacted STT-MRAM banks while being re-referenced again. In order to keep track of read-intensive blocks, even after eviction from LLC, we utilize a read-intensive block profiler, which is basically a queue of 16 entries that maintain the address of recent frequently-read blocks. Upon a read miss in LLC, the address of missed data is searched in the profiler. If it is found, a cache block from low-PV impacted STT-MRAM ways based on Least Recently Used (LRU) policy is replaced by fetched data from memory. The dirty victim block is written back into memory while the clean victim block is silently dropped.

3.2.5 Proposed PV/Energy-Aware Cache Migration Policy

Besides considering hybrid SRAM and STT-MRAM designs to accelerate service to write operations and improve bank accessibility, we also propose an efficient block insertion/migration policy to maximize the SOS throughput [2]. The tag store associated with STT-MRAM banks are equipped with three fields, Read Counter (RC), Write Counter (WC), and PV status. The main idea behind using RC is to identify vulnerable read-intensive blocks in the set. If a frequently-read block is allocated to a high-PV impacted STT-MRAM array, the cache block must be relocated to a low-PV impacted region of the set to guarantee reliable read operations.

We conducted an extensive exploration to evaluate the preferred value for the read threshold level, $\mathbf{NR}_{th}$, within our design. We found that if $\mathbf{NR}_{th}$ is small, the ratio of blocks that must be transferred to a low-PV impacted region significantly increases, while if $\mathbf{NR}_{th}$ is large, then SOS utilization significantly decreases because only a few read-intensive cache blocks are selected for migration. Thus, we set $\mathbf{NR}_{th}$ based on extensive study on block access patterns of under test workloads. In addition, the non-access-intensive cache blocks located in low-PV impacted data arrays in STT-MRAM is selected to be replaced by vulnerable read-intensive blocks, if the corresponding RC of one of the high-PV impacted blocks reaches $\mathbf{NR}_{th}$.

Additionally, WC is a saturating counter to keep track of write access patterns to a cache block. If WC reaches its write threshold level, $\mathbf{NW}_{th}$, it is considered as a write-intensive block. We propose to transfer these blocks to SRAM data arrays in order to amortize the latency and high dynamic energy consumption associated with incoming write operations. The PV status determines whether a cache block is located in low-PV or high-PV impacted data array regions. This bit is set based on a consensus decision-making process in the tag store during the POST phase.

Table 3.5: Architecture Simulation and Evaluation Parameters. [2]

Parameter	Values and Description
core	3.3GHz, Fetch/Exec/Commit width 4
L1	Private, 32 KB, I/D separate, 8-way, 64 B, SRAM, WB
L2	Shared, 4 MB, 8 banks, 8-way, 64 B, STT-MRAM, WB
Memory	8 GB, 1 channel, 4 ranks/channel, 8 bank/rank

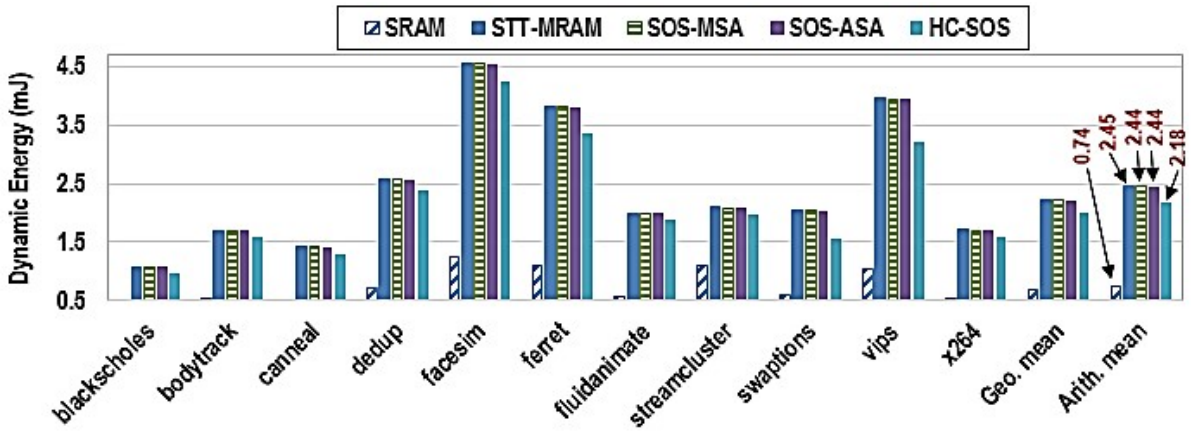
4MB L2 cache bank configuration (32nm, temperature=350K)

L2 Cache Technology	**RL/WL** $(cycles)$	**RE** (nJ)	**WE** (nJ)	**LP** (mW)	**Area** (mm^2)	**Iso-Area**
1MB SRAM	7.43/5.78	0.161	0.156	295.58	1.82	Case 1
4MB STT-MRAM	9.08/25.58	0.216	0.839	18.39	1.86	Case 1
4MB MSA-based SOS	9.08/25.58	PCSA=0.209 SPCSA=0.218	0.839	18.39	2.64	Case 2
4MB ASA-based SOS	9.08/25.58	EASA=0.208 VISA=0.217	0.839	18.39	2.72	Case 2

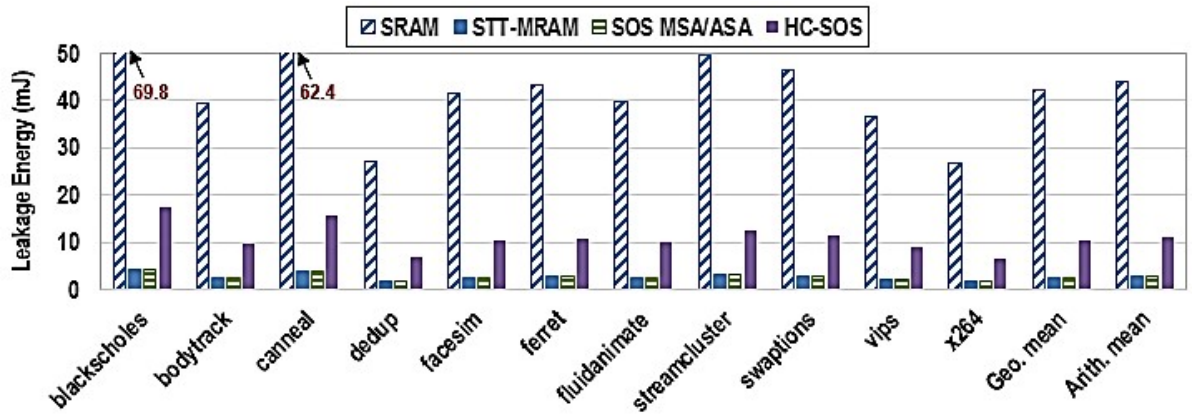
RL: Read Latency, WL: Write Latency, RE: Read Energy, WE: Write Energy, LP: Leakage Power

3.2.6 Architecture-Level Results and Analysis

To comprehensively evaluate the efficacy of SOS, we analyzed SOS on both circuit-level and architecture-level simulators. Architectural experimental results are presented in this Section utilizing the evaluation parameters listed in Table 3.1 and Table 3.5. The latency and energy usage associated with read and write operations for SRAM and conventional SA cache accesses are provided by NVSim [211]. However, we integrate the obtained results from Section 3.1.1 for 1-bit MSA and ASA into NVSim to extract the power and performance parameters for cache accesses in the SOS design. PARSEC 2.1 benchmarks suite [207] is executed on a modified MARSSx86 [212], which supports asymmetric cache read and write from distinct cache banks to extract the evaluation parameters of different cache designs during program execution. We model a Chip Multi-Processor (CMP) with four single-threaded x86 cores. Each core consists of private $L1$ cache, and shared LLC among all the cores. Eleven workloads are executed for 500 million instructions starting at the Region of Interest (RoI) after warming up the cache for 5 million instructions. The **simsmall** input sets are used for all PARSEC workloads [207].



(a)



(b)

Figure 3.15: (a) LLC dynamic energy comparison, and (b) LLC leakage energy comparison for SRAM, STT-MRAM, SOS-MSA, SOS-ASA, and HC-SOS, respectively. [2]

The experimental results indicate that HC-SOS can save up to 10.6%, on average, of dynamic energy consumption compared to STT-MRAM-based LLC. Although SRAM exhibits lower dynamic energy consumption, its high leakage power has worsened the overall consumed energy compared to other designs, as shown in Figure 3.15. Both STT-MRAM and SOS-MSA/ASA can conserve 88% on average of the total consumed energy. HC-SOS incurs higher leakage energy compared to STT-MRAM and SOS-MSA/ASA due to leveraging two SRAM-based banks in the design, incurring relatively more leakage energy to the entire cache subsystem.

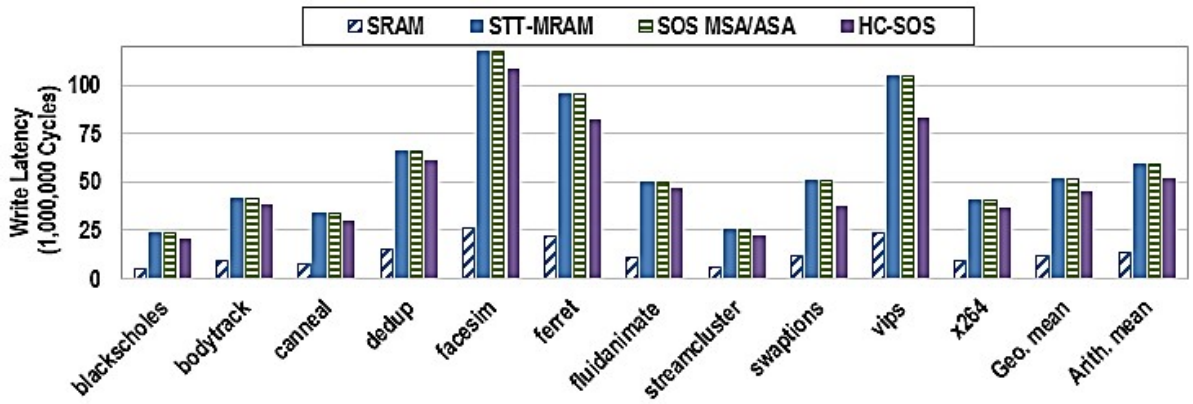


Figure 3.16: Write performance comparison for SRAM, STT-MRAM, SOS MSA/ASA, and HC-SOS. [2]

HC-SOS improves the write performance by 12.4%, on average, compared to STT-MRAM. The results indicate that the workloads, such as $vips$, $swaptions$, and $ferret$, leverage the full potential of HC-SOS to further diminish the high write latency, which adversely impacts the entire cache sub-system throughput and accessibility.

The proposed PV-/Energy-Aware cache block migration policy further improves the SOS throughput by relocating read/write intensive blocks, which results in enhanced TDS, write performance, and bank service time. Namely, the VFDS in the HC-SOS-ASA is reduced by 89% on average compared to LLC with STT-MRAM, thus improving the mean TDS from 72.5% to 97% across all workloads.

The energy consumption of PV/energy-aware migration policy is shown in Figure 3.18, which demonstrates the dynamic energy consumption breakdown associated with swapping high-PV impacted read-intensive blocks within STT-MRAM-based banks and migrating write-intensive blocks to SRAM-based banks. The corresponding energy overhead for PV/energy-aware migration policy is around $14\mu J$ which is less than 0.7% of total LLC dynamic energy consumption. This implies that the migration energy overhead is insignificant and incurs a minor energy overhead to the entire system. A comparison with previous works is listed in Table 3.6.

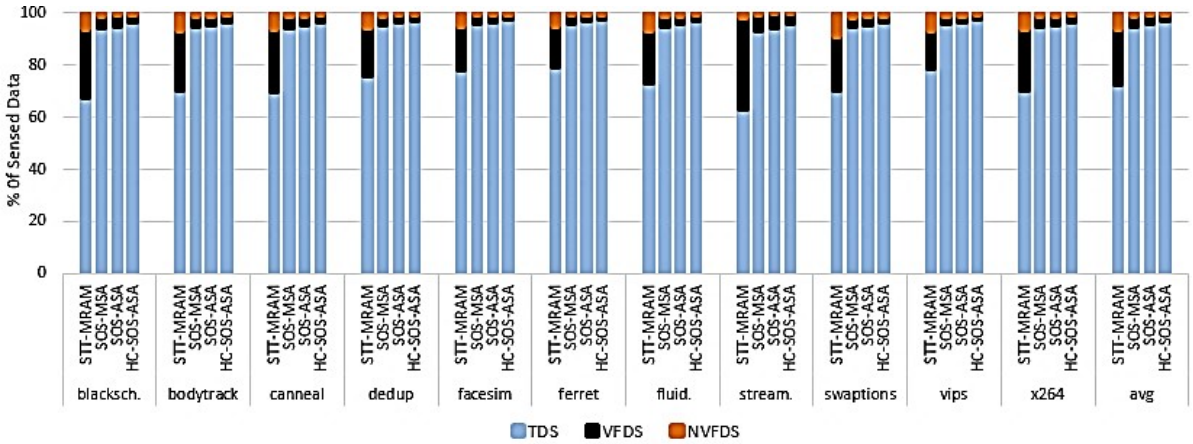


Figure 3.17: Distribution of sensed data. SOS is equipped with MSA, ASA, and migration policy for ASA design. [2]

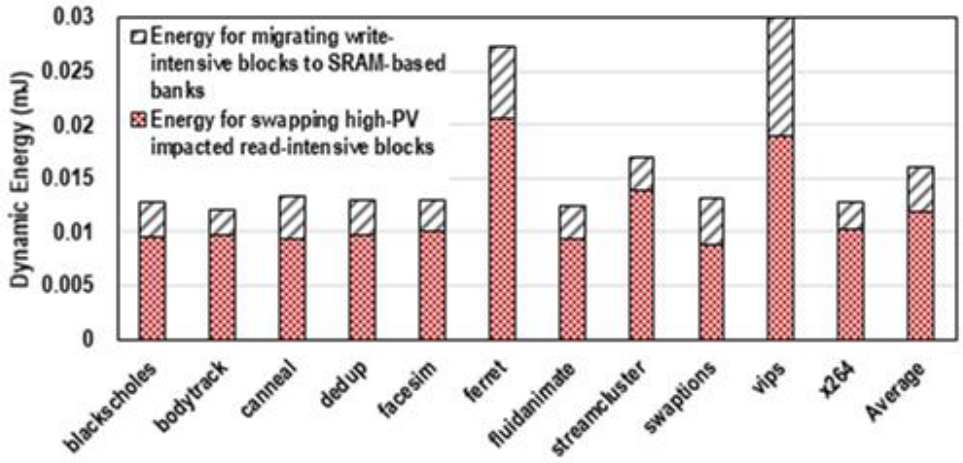


Figure 3.18: The dynamic energy consumption associated with PV/energy-aware migration policy. [2]

3.3 Conclusion

Spin-Transfer Torque Magnetic Random Access Memory (STT-MRAM) has been explored as a post-CMOS technology for embedded and data storage applications seeking non-volatility, near-zero standby energy, and high density. Towards attaining these objectives for practical implementations, various techniques to mitigate the specific reliability challenges associated with STT-MRAM elements are surveyed, classified, and assessed herein.

Table 3.6: Related Work Comparison Table. [2]

Design	Circuit-Level/ Architecture-Level	Read Enhancement		Write Enhancement		Contribution
		PV Reliability	Performance	PV Reliability	Performance	
RWHCA [213]	Architecture-Level	✗	✓	✗	✓	RWHCA reduces power dissipation by 55% on average, while achieving 5% improvement IPC compared to the baseline SRAM cache across 30 workloads.
APM [169]	Architecture-Level	✗	✗	✗	✓	Provides 18.9% and 19.3% reduction in power dissipation for single-thread and multi-thread workloads, respectively.
PHC [170]	Architecture-Level	✗	✗	✗	✓	Offers 28% and 31% reduction in energy consumption compared to existing hybrid architectures in single-core and multi-core systems, respectively.
PVA-NUCA [37]	Architecture-Level	✗	✓	✗	✓	Offers 26.4% reduced energy consumption and provide 25.29% IPC performance improvement while incurring less than 1% area overhead.
Relaxed-Retention [214]	Circuit- and Architecture-Level	✓	✓	✗	✓	Increases read and write performance of STT-MRAM and reliability of read at the cost of decreasing retention time and thus requiring periodic refresh.
HC-SOS (This Work)	Circuit- and Architecture-Level	✓	✓	✓	✓	SOS-enabled Hybrid Cache improves write performance by 12.4% on average compared to STT-MRAM baseline cache design, improves the mean TDS from 72.5% to 97%, and reduces VFDS by 89% on average across all workloads.

Some solutions to the reliability issues identified are addressed for reliable STT-MRAM designs. In an attempt to further improve the Process Variation (PV) immunity of the Sense Amplifiers (SAs), two new SAs have been introduced: Energy Aware Sense Amplifier (EASA) and Variation Immune Sense Amplifier (VISA). Results have shown that EASA and VISA achieve superior performance in most cases compared to two of the most common SAs, namely PCSA and SPCSA respectively, while reducing Bit Error Rate (BER) and increasing reliability.

While inclusion of emerging technology-based Non-Volatile Memory (NVM) devices in on-chip memory subsystems offers excellent potential for energy savings and scalability, their sensing vulnerability creates PV challenges. In this dissertation, I propose a circuit-architecture cross-layer solution to realize a radically-different approach to leveraging as-built variations via specific SA design and use. This novel approach, referred to as a Self-Organized Sub-bank (SOS) design, assigns the preferred SA to each Sub-Bank (SB) based on a PV assessment, resulting in energy consumption reduction and increased read access reliability. To improve the PV immunity of SAs, two reliable and power efficient SAs, called the Merged SA (MSA) and the Adaptive SA (ASA) are introduced herein for use in the SOS scheme. Furthermore, I propose a dynamic PV and

energy-aware cache block migration policy that utilizes mixed SRAM and STT-MRAM banks in Last Level Cache (LLC) to maximize the SOS bandwidth. The experimental results indicate that SOS can alleviate the sensing vulnerability by 89% on average, which significantly reduces the risk of application contamination by fault propagation. Furthermore, in the light of the proposed block migration policy, write performance is improved by 12.4% on average compared to the STT-MRAM-only design.

CHAPTER 4: SELF-ORGANIZED SUB-BANK SHE-MRAM-BASED LLC: AN ENERGY-EFFICIENT AND VARIATION-IMMUNE READ AND WRITE ARCHITECTURE[1]

In this Chapter, we focus on increasing energy efficiency and reliability of write operations in STT-MRAM and is motivated by the observation that the STT switching technique suffers from high dynamic energy consumption [40]. SHE-MTJ has been recently studied as an energy-efficient alternative for STT-MTJ due to its improved performance. Several write circuits have been studied in recent years in order to achieve optimum energy while maintaining high reliability. Herein, we explore SHE-MTJ write circuits and compared those with conventional STT-MTJ write circuits in terms of performance and reliability. Furthermore, a high-resilient write circuit as well as an energy-efficient write circuit are selected in order to be utilized in the SOS approach for further performance and reliability improvements of SHE-MRAM. In particular, the SOS approach is implemented once with the high-resilient write circuit and once with the energy-efficient write circuit. Our results indicate that the energy-efficient write circuit provides significant energy and delay improvements over the conventional STT-MTJ write circuit and high-resilient SHE-MTJ write circuit. On the other hand, the high-resilient write circuit for SHE-MTJ offers reliability improvement over the energy-efficient SHE-MTJ write circuit.

4.1 Write Circuit Design and Analysis

In this section, various write schemes are investigated for switching the states of the STT-MTJ and SHE-MTJ devices. Herein, we have simulated the write circuits using SPICE circuit simulator in

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

22nm PTM library [202] using 1.0V nominal voltage. To provide a fair comparison, the size of the write transistors are enlarged 6-fold to produce a write current greater than the critical current for all of the investigated bit cells. Herein, we have utilized a chain of four inverters to drive Bit Line (BL), Source Line (SL), and Word Line (WL). Each successive inverter is twice as large as the previous one.

4.1.1 STT-MRAM Write Schemes

Figures 4.1(a) and 4.1(b) show energy-aware STT-MRAM bit cell circuits inspired by the designs proposed by Ben-Romdhane *et al.* [11] and Zand *et al.* [109], respectively. The transmission gate (TG)-based write circuit leverages the near-optimal full-swing switching behavior of TGs to provide a high amplitude write current, which leads to a high speed switching. The simulation results listed in Table 4.1 indicate the advantage of TG-based STT-MRAM bit cell circuit (1T-1R) compared to conventional 1T-1R circuit and the write scheme proposed by Ben-Romdhane *et al.* [11]. According to the results listed in Table 4.1, 1TG-1R design provides roughly 1.7-fold improved EDP compared to 1T-1R design and 1.5-fold improved EDP compared to the design proposed in [11].

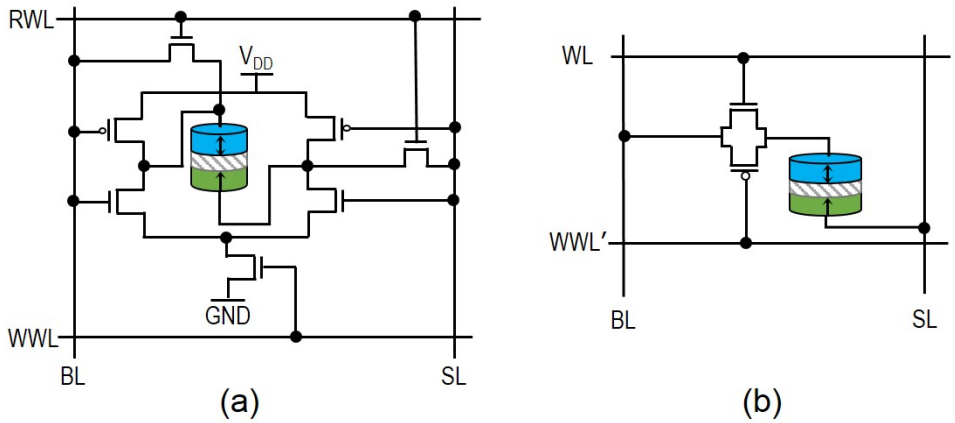


Figure 4.1: (a) 7T-1R [11] STT-MRAM Bit-Cell, (b) 1TG-1R STT-MRAM Bit-Cell. [3]

Table 4.1: Write Characteristics for Various STT-MRAM Bit-Cells. [3]

Features		1T-1R	7T-1R[11]	**1TG-1R**
Parallel (P) to Anti-Parallel (AP)	Current (μA)	136.3	118.1	**172.3**
	Delay (ns)	3.67	4.5	**2.7**
	Power (μW)	137	119.1	**181.5**
Anti-Parallel (AP) to Parallel (P)	Current (μA)	81.36	110.2	**134.3**
	Delay (ns)	5.73	3.95	**3.1**
	Power (μW)	82.13	111.2	**143.46**
Average Energy (fJ)		486.7	487.6	**467.4**
Energy Delay Product (EDP) (fJ$\times$ns)		2270.9	2073.4	**1350.9**
Average EDP Improvement	7T-1R	–	–	**34.8%**
	2T-1R	–	8.7%	**40.5%**

4.1.2 SHE-MRAM Write Schemes

Despite the advantages of conventional STT switching approaches, their main challenge is relatively high switching delay and energy consumption. SHE-assisted STT switching mechanism have been introduced as an alternative for conventional STT switching enabling significantly reduced switching energy. Herein, we have leveraged two bit cells proposed by authors in [109] for switching the SHE-MTJ devices. Figure 4.2(a) shows a 7T-1R bit cell requiring two read transistors and five write transistors. The 7T-1R bit cell has a completed current path from **VDD** to **GND** via the transistors and the **HM**. Since the **BL** and the **SL** are electrically isolated from the current path, the strengths of the **BL** and **SL** drivers do not need to be considered for the write operation. This makes 7T-1R an energy-efficient design, however it incurs significant area overhead. A 1TG-1T-1R bit cell is shown in Figure 4.2(b) which includes one TG for write and one transistor for read operation. The results provided in Table 4.2 shows that 1TG-1T-1R bit cell is the most energy-efficient design with significantly improved EDP values. According to the results, 1TG-1T-1R offers 1.6-fold improved EDP compared to 2T-1R and roughly 1.7-fold improved EDP compared to 7T-1R design.

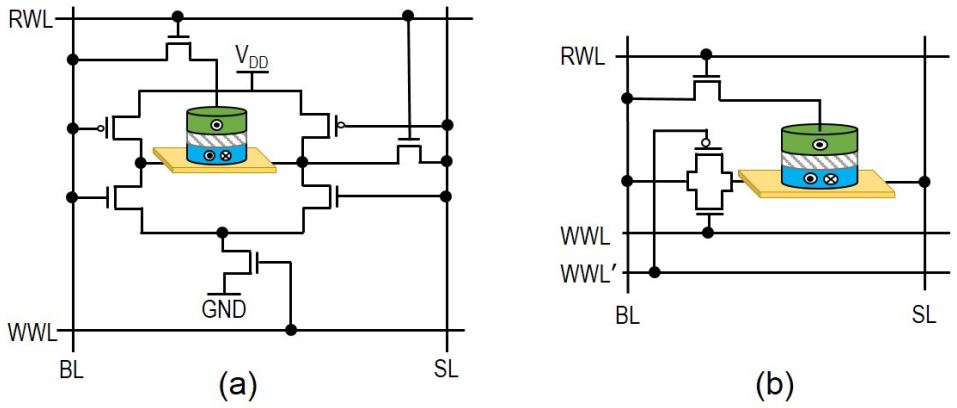


Figure 4.2: (a) 7T-1R SHE-MRAM Bit-Cell, (b) 1TG-1T-1R SHE-MRAM Bit-Cell. [3]

Table 4.2: Write Characteristics for Various SHE-MRAM Bit-Cells. [3]

Features		2T-1R	7T-1R	**1TG-1T-1R**
Parallel (P) to Anti-Parallel (AP)	Current (μA)	138.7	119.8	**181.7**
	Delay (ns)	2.25	2.63	**1.61**
	Power (μW)	139.4	120.9	**190.8**
Anti-Parallel (AP) to Parallel (P)	Current (μA)	112.4	119.8	**176.9**
	Delay (ns)	2.85	2.63	**1.66**
	Power (μW)	113.1	120.9	**186.1**
Average Energy (fJ)		317	318	**308**
Energy Delay Product (EDP) (fJ$\times$ns)		812.18	836.25	**503.7**
Average EDP Improvement	7T-1R	–	–	**34.8%**
	2T-1R	–	8.7%	**40.5%**
Normalized Area Compared to 2T-1R		1	10.1	**2.88**

We have also examined the performance of the introduced SHE-MRAM bit cells in presence of variations in **HM** dimensions (σHM) and transistors' threshold voltage (σV_{th}). These two types of PVs have the most impact on the produced write current. Figure 4.3 shows the produced write current fluctuations versus σHM and σV_{th}. The results exhibit that the 7T-1R bit cell is the most variation-resilient design with less than 8% variation for the worst case scenario investigated herein, i.e. $\sigma HM = 10\%$ and $\sigma V_{th} = 10\%$. In 7T-1R bit cell, the size of the transistors should be tripled to generate a write current greater than switching critical current. Although this

leads to a significant area overhead, it can enhance the PV-tolerance since increasing the size of the transistors is one of the most commonly-used methods to improve the variation resistance [215].

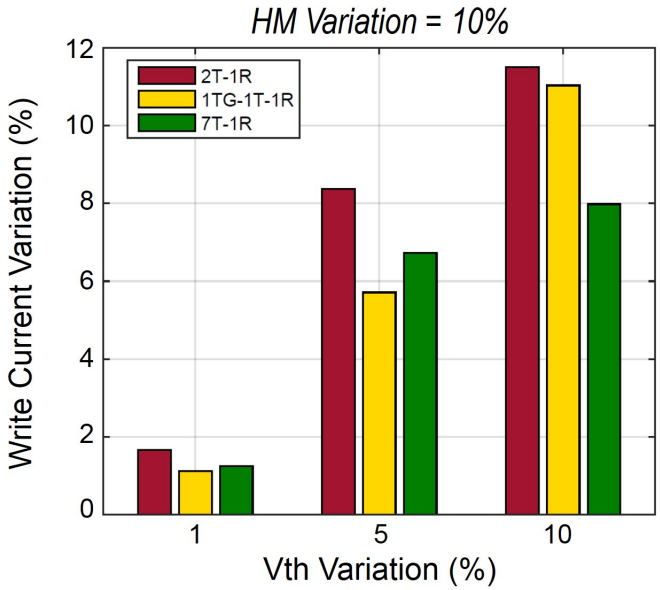


Figure 4.3: Write current variations versus σV_{th} for $\sigma HM = 10\%$. [3]

4.2 Architecture-Level Simulation Results

In order to fully evaluate the SOS approach's efficacy using both the STT and SHE write approaches, architectural-level analysis is necessary. The evaluation parameters used in order to extract the architectural-level simulation results are listed in Table 4.3. In our analysis method, first the circuit-level simulation is performed in order to extract the required parameters for a single bit-cell and then these parameters are forwarded to architecture-level simulators, GEM5 [216] and NVSim [211], to extract system-level results.

In this Section, the EDP is calculated for read and write memory operations in each cache access based on the circuit-level results from Section 4.1 and [2]. PARSEC benchmarks suite [207] executed on modified MARSSx86 [212] which supports asymmetric cache read and write from

distinct cache banks to extract the evaluation parameters of different cache designs during program execution. We model a Chip Multi-Processor (CMP) with four single-threaded x86 cores. Each core consists of private $L1$ cache, and shared LLC among all the cores. Eleven workloads are executed for 500 million instructions starting at the Region Of Interest (ROI) after warming up the cache for 5 million instructions. The **simsmall** input sets are used for all PARSEC workloads [39].

Table 4.3: Architecture Parameters. [3]

Parameter	**Values and Description**
core	3.3GHz, Fetch/Exec/Commit width 4
L1	private, 32 KB, I/D separate, 8-way, 64 B, SRAM, WB
L2	shared, 4 MB, 8 banks, 8-way, 64 B, STT-MRAM, WB
	shared, 4 MB, 8 banks, 8-way, 64 B, SHE-MRAM, WB
memory	8 GB, 1 channel, 4 ranks/channel, 8 bank/rank

4MB L2 cache bank configuration (32nm, temperature=350K)

L2 Cache Technology	**RL/WL** $(cycles)$	**RE** (nJ)	**WE** (nJ)	**LP** (mW)	**Area Overhead***
STT-based SOS	9.08/25.58	PCSA=0.209 SPCSA=0.218	0.839	18.39	7.44**
SHE-based SOS	9.08/13.30	PCSA=0.209 SPCSA=0.218	0.553	18.39	7.44**

RL: Read Latency, WL: Write Latency, RE: Read Energy,
WE: Write Energy, LP: Leakage Power
$*$: Area overhead reflects the overhead of SA and calculated as:
$\frac{(Area\ using\ the\ MSA)}{(Area\ using\ the\ PCSA)}$
$**$: Since STT-MTJ and SHE-MTJ are fabricated on top of the CMOS circuitry, the area of the STT-based and SHE-based SOS are identical.

4.2.1 Energy Delay Product (EDP)

In order to clarify the advantage of using SOS equipped with SHE devices for read and write operations, four LLC designs are compared in terms of EDP. The conventional STT-MRAM based LLC utilizes PCSA [83] in its organization to maintain the consumed power as low as possible. On the

other hand, SOS improves the sense margin of high-PV impacted regions of LLC through SPCSA [117] employment. Even though SPCSA sacrifices energy efficiency to offer higher reliability, the amount of additional consumed energy due to the utilization of SPCSA can be alleviated by SOS technique. To be specific, SOS neutralizes the high energy consumption of SPCSA via assigning low-power PCSA array to the LLC regions that are impacted by PV negligibly. The effect of this compensation has been shown in Figure 4.4 whereby the EDP of SOS technique and the design that benefit exclusive SPCA is approximately even across all benchmarks.

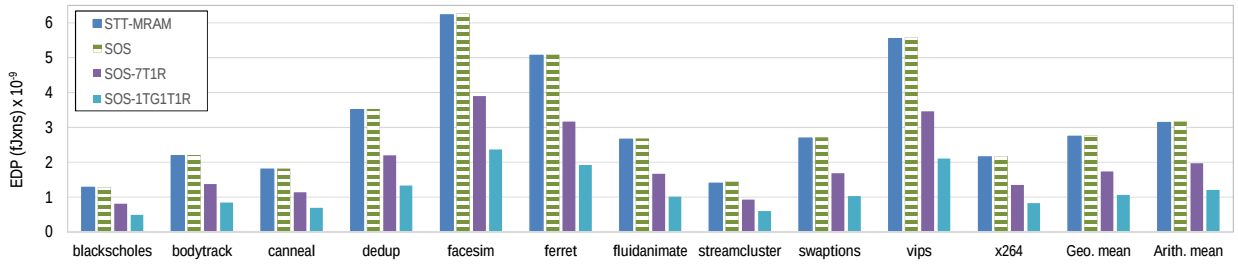


Figure 4.4: EDP comparison for STT-MRAM, SOS, SOS-7T1R, and SOS-1TG1T1R. [3]

The high write energy overhead for storing a value into STT-MRAM cell incurs significant energy overhead in both conventional STT-MRAM and SOS while SHE devices utilized in SOS-7T1R and SOS-1TG1T1R significantly reduce the required write energy. In particular, the EDP of each memory cell in SOS-1TG1T1R is less than SOS-7T1R according to the basis presented in Section 4.1. Thus, the EDP of LLC entailing 1TG1T1R in its arrangement is significantly less than other designs. This incident is conspicuous for write-intensive workloads such as **facesim**, **ferret**, and **vips** where the ratio of write accesses to the LLC is significantly more than read accesses. On average, SOS-1TG1T1R decreases the EDP by 39% compared to SOS-7T1R, leading to the considerable performance improvement and energy consumption reduction.

4.2.2 Empirical Analysis of Fault Model Associated with Sensed Data

The memory accesses to the LLC blocks has unprecedented pattern for each class of workloads. This means that some cache lines in the banks may experience a large number of read operations while others may be accessed by frequent write operations. This non-uniform access can be problematic when the read-intensive cache lines are placed into a high PV-impacted region of a bank, which results in increasing the ratio of VFDS for that particular workload. SOS reduces the ratio of VFDS by assigning high-reliable SA to the high-PV impacted sub-banks while substantially reducing the energy consumption overhead associated with read operations through low-power SA assignment to low-PV impacted sub-banks.

Figure 4.5 illustrates the comparison of distribution of read operation reliability between LLC equipped with conventional STT-MRAM, SOS circuit strategy, and different SHE devices. We assume that the PV map for each cache bank is similar to the floorplan shown in Figure 3.13. Based on the position of accessed sub-bank in the floorplan, different PV ratios are applied during fault analysis of the workloads. To be specific, if a sub-bank experiences a high amount of PV, the probability that the data will be sensed incorrectly is high. Our experimental results indicate that the PV effect may incur around 27.5% of the sensed data to be read incorrectly from which 21.5% are extremely vulnerable which implies that about one fifth of the overall sensing operations have the potential to contaminate the application's data structure. If this rate of sensed data is not accommodated, it may induce application crashes or prolong the program execution. Based on the PV map and the access pattern shown in different class of benchmarks, we observed that the proportion of read operations and dirty victim blocks residing in LLC in **blackscholes** and **canneal** workloads, are more than write operations which results in the increased VFDS. Furthermore, the **streamcluster** workload is a read-intensive application in which more than 85% of memory operations are read accesses which increases the chance for enduring higher VFDS.

The probability of sensing incorrect data is addressed through leveraging PV-resilient SAs array (SPCSA) in the SOS approach whenever the sub-bank's PV ratio is more than a predefined threshold. Namely, the VFDS in the SOS design is reduced by 82% on average compared to LLC with conventional STT-MRAM. The VFDS of write-intensive benchmarks such as **ferret** and **vips** is 15.87% and 14.35%, respectively which is substantially less than the VFDS of read-intensive benchmarks such as **streamcluster** (34.93%).

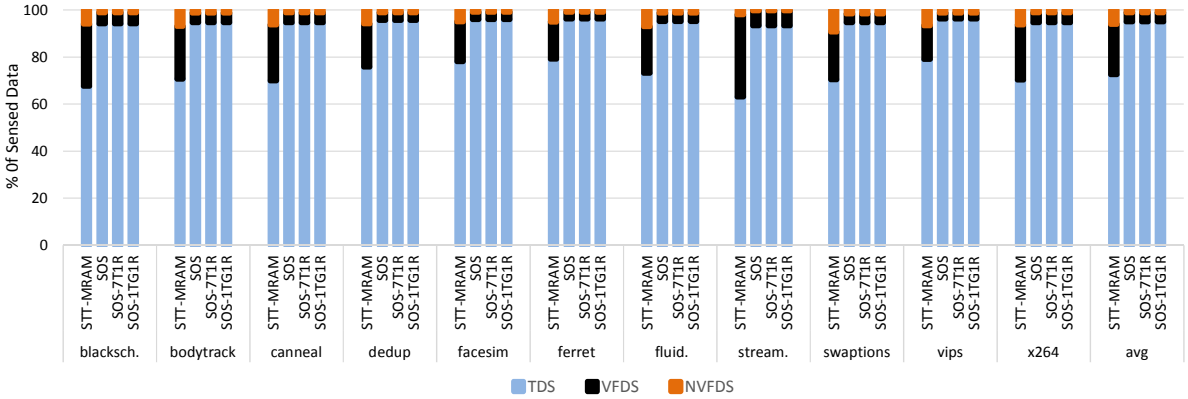


Figure 4.5: Distribution of read operation reliability. The rightmost bars for each workload show the SOS equipped with SHE (7T1R) and SHE (1TG1T1R), respectively. [3]

Additionally, SOS considerably improves the TDS of read-intensive benchmarks. In particular, we observed that the TDS of **streamcluster** can improve by 32.43% using SOS. However, this improvement is not significant for write-intensive benchmarks. For example, SOS improves the TDS of **vips** workload by 17.87% which is almost half of the improvement observed in read-intensive workloads. In overall, the mean TDS is improved from 72.5% to 95% across all workloads via SOS utilization.

4.3 Conclusion

In order to reduce static energy consumption, emerging NVM technologies such as STT-MRAM, Spin-Hall Effect Magnetic RAM (SHE-MRAM), Phase Change Memory (PCM), and Resistive

RAM (RRAM) are under intense research. Additionally, there is a demand for more reliable circuits as the technology scales due to increased error rates caused by the increased impact of PV. In order to combat PV-induced reliability problems, a novel approach is proposed herein that improves the reliability of read and write operations in emerging NVMs. In the proposed SOS design, two SAs have been adopted, one with improved reliability and one with improved energy efficiency profiles, in order to increase the performance of the read operation. In particular, based on the result of a Power-On Self-Test (POST), which detects PV-impact on sub-banks, SOS chooses between a reliable and an energy-efficient SA and assigns a preferred SA to each sub-bank. Based on the preliminary observation in our case study, 21.5% of read operations are extremely vulnerable to PV impacts. Our results indicate that the proposed SOS approach reduces the vulnerability of the read operation by 40% on average, hence reducing the fault propagation. In particular, the SOS alleviates Vulnerable False Data Sensing (VFDS) by 82% on average, while enhancing True Data Sensing (TDS) from 72.5% to 95% across all workloads studied herein compared to LLC with conventional STT-MRAM.

Furthermore, in order to increase the performance of the write operation, SHE-MRAM is replaced with STT-MRAM to provide better write energy profile. Additionally, SOS design is once implemented with a reliable write scheme and once with an energy-efficient write scheme and results are compared and analyzed. Additionally, SOS using the reliable write circuit provides 161% improved Energy Delay Product (EDP) on average compared to SOS with conventional STT-MRAM, while providing less than 8% write current variation. On the other hand, SOS using energy-efficient write circuit offers 39% improved EDP on average compared to the SOS using reliable write circuit and 62% EDP improvement over conventional STT-MRAM.

CHAPTER 5: BGIM: BIT-GRAINED INSTANT-ON MEMORY CELL FOR SLEEP POWER CRITICAL MOBILE APPLICATIONS[1]

In this Chapter, we devise an energy-efficient and fast Non-Volatile Static Random Access Memory (NV-SRAM) design utilizing emerging spin-based devices. Herein, we propose a Bit-Grained Instant-on Memory (BGIM) cell. The proposed BGIM, utilizes Differential SHE-MRAM (DSH-MRAM) devices to provide fast back-up and restore operations in a novel energy-efficient fashion. By leveraging non-volatility and zero leakage power dissipation, BGIM can reduce stand-by energy consumption via instant off/on operation without the use of a separate Non-Volatile Memory (NVM) macro, such as FLASH. This design takes advantage of the SRAM cell's speed during normal operation and uses the corresponding DSH-MRAM cell for nonvolatile storage of the memory's state with very low time and energy costs for each store/restore operation.

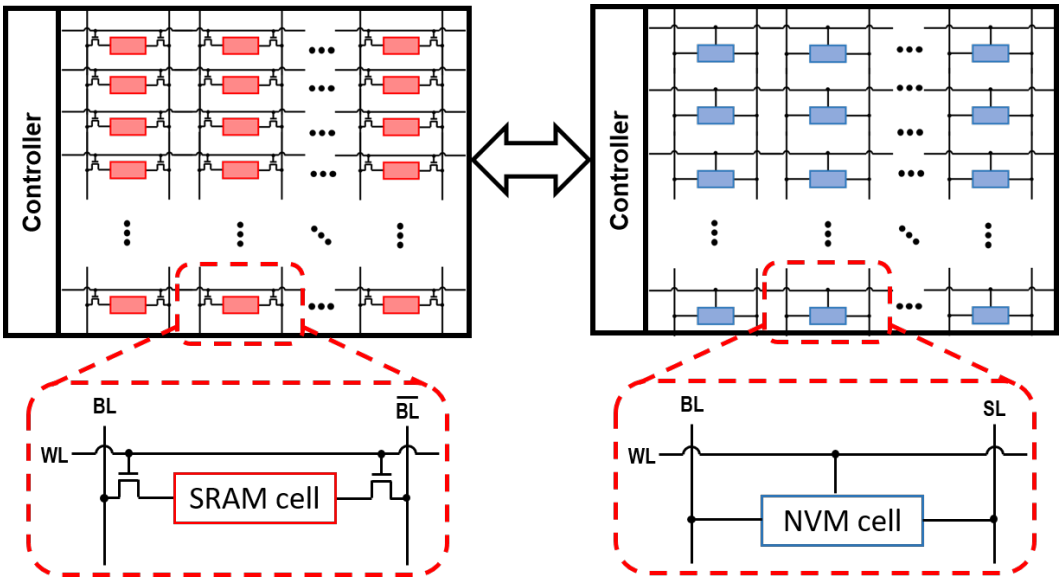
Internet of Things (IoT) and mobile devices that operate under significant energy constraints but require fast memory performance, especially those that undergo frequent transitions to and from stand-by mode, could particularly benefit from the proposed BGIM design. The BGIM in-situ intra-cell retention mechanism proposed herein offers two advantages in that regard. First, design regularity is increased while long-wire and busing complexity are decreased, especially during verification and validation, as compared to a checkpointing-and-restore strategy. Second, the long break-even sleep period required due to the energy overhead of a backing store can be partitioned and reduced using a bit-cell resolution only where needed throughout any datapath or storage module.

[1] ©IEEE. Part of this chapter is reprinted, with permission, from [4]

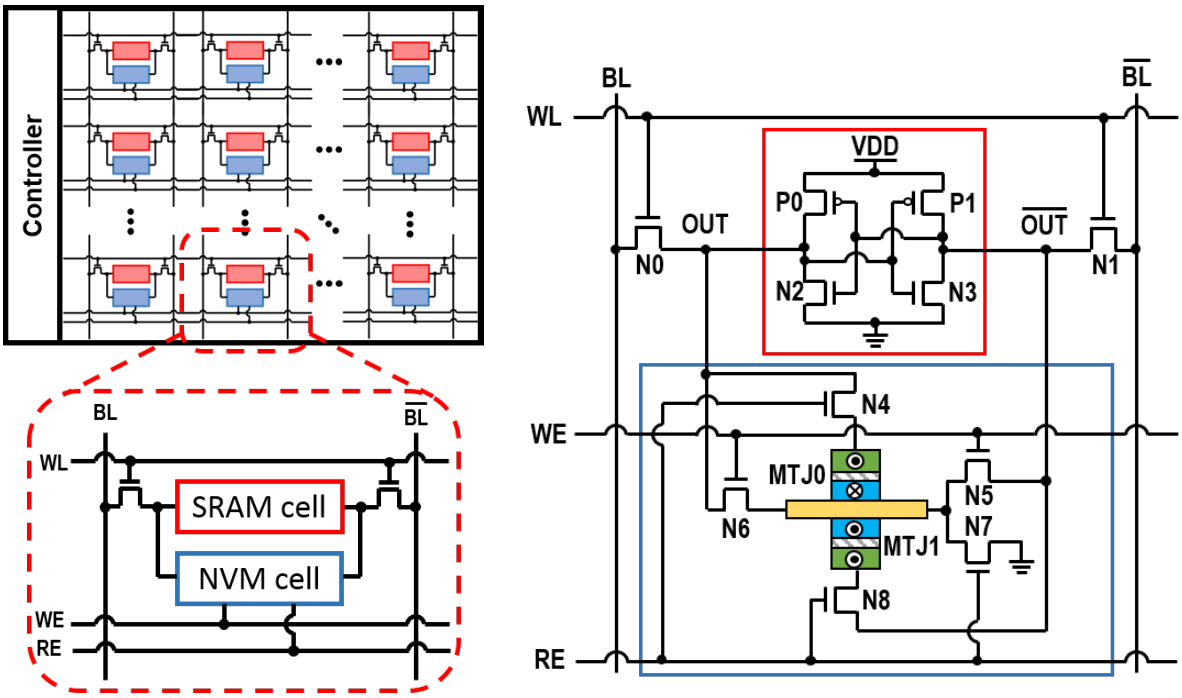
5.1 Proposed Bit-Grained Instant-on Memory (BGIM) Cell

Our proposed BGIM one-macro architecture is shown in Figure 5.1(b) and the circuit view of the BGIM cell leveraging DSH-MRAM devices is shown in Figure 5.1(c). As observed from Figure 5.1(a) and 5.1(b), our proposed BGIM one-macro architecture is capable of energy-efficient and rapid back-up and restore operations compared to the conventional two-macro architecture due to the elimination of data transfer between the SRAM and NVM macros. As shown in Figure 5.1(b), two control signals, namely Write Enable (**WE**) and Read Enable (**RE**), are included in our bit-cell to control the back-up and restore operations. The proposed BGIM cell consists of a 6-Transistor (6T) SRAM cell accompanied with a NVM device. As shown in Figure 5.1(c), the DSH-MRAM device consists of 5 access transistors, **N4-N8**, to control back-up, stand-by, and restore operations and 2 MTJ devices, **MTJ0** and **MTJ1**, used for holding the SRAM data. The combination of the control transistors and MTJ devices, which comprises the NVM part of the BGIM cell, is referred to as $5T2R$ herein.

Additionally, control signals for different operating modes of the proposed BGIM cell are listed in Table 5.1. One of the major benefits of the proposed BGIM cell is that unlike STT-MRAM design, it does not require high current densities for write operations. Another major benefit of the proposed BGIM cell is that unlike other NV-SRAM cells, it does not require an additional sensing step to read the value store in the SRAM cell before back-up operation. Furthermore, as shown in Figure 5.1(c), the proposed BGIM cell requires only five additional transistors (**N4-N8**) and two additional control signals (**WE** and **RE**) to perform back-up and restore operations compared to other conventional NV-SRAM approaches using STT-MRAM and RRAM, where more additional peripheral circuitry is required for back-up and restore operations [17].



(a)



(b) (c)

Figure 5.1: (a) Conventional two-macro architecture, (b) the proposed one-macro BGIM architecture, and (c) The proposed one-macro BGIM bit-cell circuit view using DSH-MRAM. [4]

Table 5.1: The Signaling of the BGIM cell for different operations. [4]

Operation	BL	$\overline{BL}$	WL	WE	RE
Normal	Original Data to store in SRAM	Differential Data to store in SRAM	1	0	0
Back-up	0	0	0	1	0
Stand-by	0	0	0	0	0
Restore	Pre-charge to VDD	Pre-charge to VDD	1	0	1

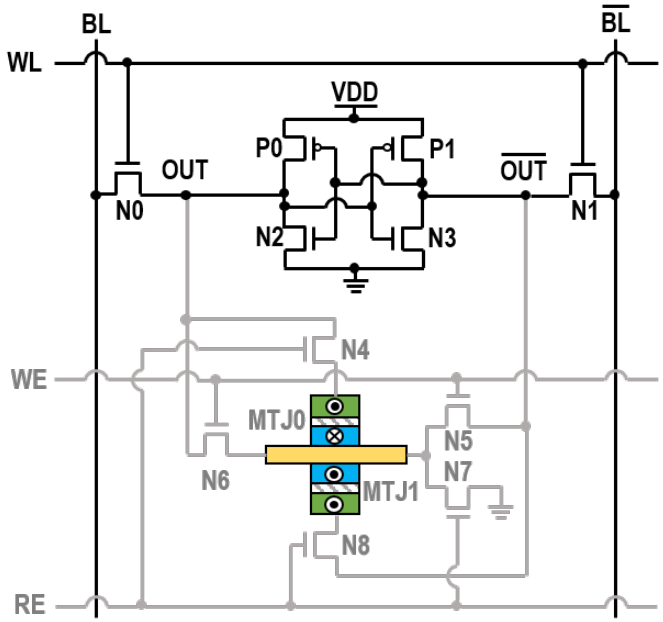


Figure 5.2: The proposed one-macro BGIM bit-cell in normal operation mode. [4]

5.1.1 Normal Operation

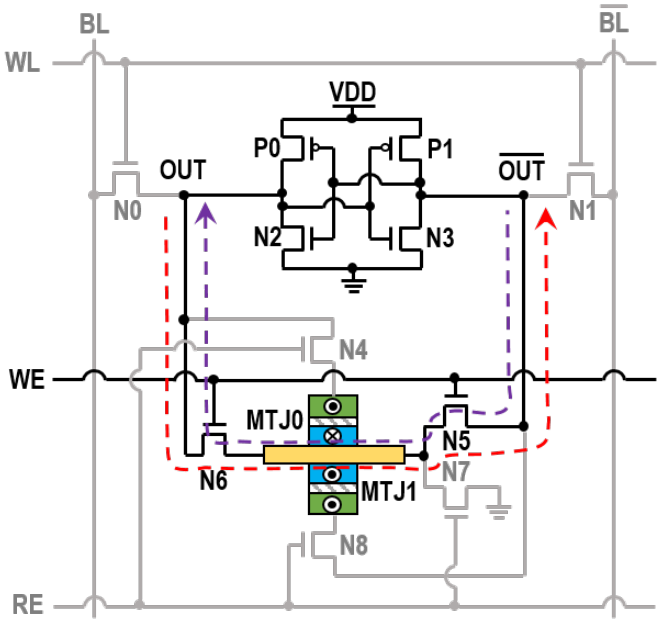
During the normal operation of the BGIM cell, **WE** and **RE** signals are set to 0 and the SRAM cell is separated from the NVM cell, as shown in Figure 5.2. When the SRAM cell is in the normal operation mode, **N4**, **N5**, **N6**, **N7**, and **N8** transistors are turned off and **P0**, **P1**, **N2**, and **N3** transistors are turned on to hold the value stored in the SRAM cell. In the normal operation mode, if a data is ready to be written in the SRAM cell, **WL** signal will turn **N0** and **N1** transistors on and

connects the **BL** and $\overline{\texttt{BL}}$ to **OUT** and $\overline{\textbf{OUT}}$, respectively, and the data will be stored in the SRAM cell as a result.

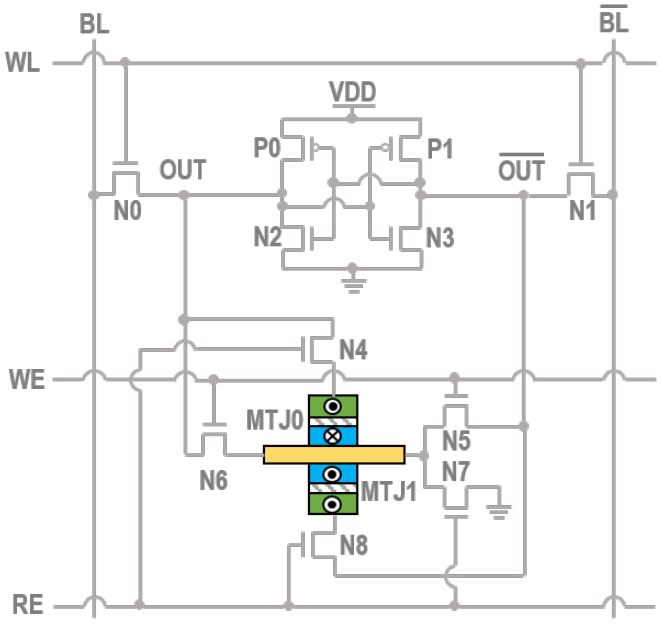
5.1.2 Back-up and Stand-by Operations

When the stand-by mode is activated, the device will go into back-up and PG state as shown in Figure 5.3. During the back-up operation, **WE** is set to 1 while **RE** is set to 0, which causes **N5** and **N6** transistors to turn on and store the SRAM data into the NVM device, as shown in Figure 5.3(a). According to the data stored in the SRAM cell, if **OUT** holds a value of 1, and $\overline{\textbf{OUT}}$ holds a value of 0, a charge current, $\mathbf{I}_{SHE}$, will be applied in the positive direction of x-axis of the Cartesian coordinate system, shown as a red dashed line in Figure 5.3(a). In this case, $\mathbf{I}_{SHE}$ will lead to two spin currents, $\mathbf{I}_{Spin-P}$ in the positive direction of the z-axis of the Cartesian coordinate system, and $\mathbf{I}_{Spin-N}$ in the negative direction of the z-axis of the Cartesian coordinate system, which will change the magnetic orientation of the **MTJ0** and **MTJ1** free-layers simultaneously and this results in the storage of 1 in **MTJ0** and 0 in **MTJ1**.

On the other hand, if **OUT** holds a value of 0, and $\overline{\textbf{OUT}}$ holds a value of 1, a charge current, $\mathbf{I}_{SHE}$, will be applied in the negative direction of x-axis of the Cartesian coordinate system, shown as a purple dashed line in Figure 5.3(a). Similarly in this case, $\mathbf{I}_{SHE}$ will lead to two spin currents, $\mathbf{I}_{Spin-P}$ in the negative direction of the z-axis of the Cartesian coordinate system, and $\mathbf{I}_{Spin-N}$ in the positive direction of the z-axis of the Cartesian coordinate system. As a result, the value of 0 will be stored in **MTJ0** and the value of 1 will be stored in **MTJ1**. Furthermore, as soon as the back-up operation terminates, the PG state will turn off the SRAM cell as well as **N4**, **N5**, **N6**, **N7**, and **N8** transistors to reduce the leakage and static power dissipation as shown in Figure 5.3(b). Since there is no need to read the data from the SRAM cell before storing it in the NVM cell, this will result in a significant reduction in the energy and delay of the back-up operation.



(a)



(b)

Figure 5.3: The proposed one-macro BGIM bit-cell in (a) back-up and (b) stand-by operation modes. [4]

5.1.3 Restore Operation

When the stand-by mode is deactivated, the device will go to restore and power-on mode. During the restore operation, **WE** is set to 0, which causes **N5** and **N6** transistors to turn off, as shown in Figure 5.4. During the restore operation, in order to read the values stored in **MTJ0** and **MTJ1**, first **WL**, **BL**, and $\overline{\textbf{BL}}$ are set to 1. As a result, **N0** and **N1** are turned on to pre-charge the output nodes **OUT** and $\overline{\textbf{OUT}}$ to **VDD**. Then, **RE** is set to 1, which causes **N4**, **N7**, and **N8** transistors to turn on to restore the SRAM data stored in **MTJ0** and **MTJ1**.

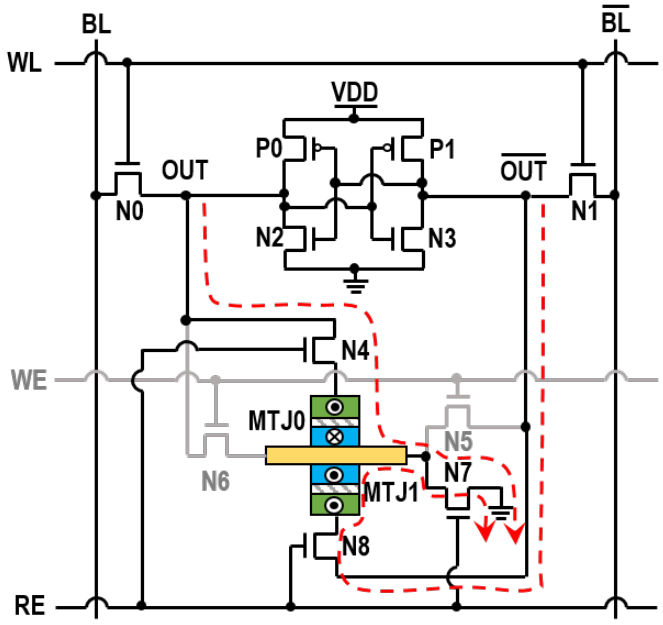


Figure 5.4: The proposed one-macro BGIM bit-cell in restore operation mode. [4]

When **N7** transistor turns on, it will provide discharging paths from **OUT** and $\overline{\textbf{OUT}}$ to the **GND**. As a result, based on the difference between **MTJ0** and **MTJ1** resistances, which are determined by the magnetization orientation of their free-layer compared to their fixed-layer, one of the two output nodes, **OUT** and $\overline{\textbf{OUT}}$, begins to discharge more rapidly to the **GND**, leading either **P0** to turn on and charge **OUT** to **VDD**, or **P1** to turn on and charge $\overline{\textbf{OUT}}$ to **VDD**. As a result, if **OUT** node is charged to **VDD**, this will cause the **N3** transistor to turn on more rapidly than the

N2 transistor and discharge $\overline{\textbf{OUT}}$ to **GND**. Additionally, if $\overline{\textbf{OUT}}$ node is charged to **VDD**, this will cause the **N2** transistor to turn on more rapidly than the **N3** transistor and discharge **OUT** to **GND**. After completion of the data restoration phase, SRAM will return to its normal operation and NVM cell will turn off until the next back-up and stand-by operations.

5.2 Simulation Results and Analysis

In order to verify and analyze the behavior of the DSH-MRAM BGIM cell proposed herein, SPICE simulation is conducted utilizing the parameters listed in Table 5.2 as well as 22nm Predictive Technology Model (PTM) [202]. To accurately model the behavior of the DSH-MRAM BGIM devices proposed herein, the modeling approach introduced in [15] and [16] is utilized.

Table 5.2: Circuit parameters and constants with their corresponding values for the DSHE-MRAM device model. [4] (Parameters are taken from [15, 16])

Parameter	Description	Default Value
M_s	Saturation Magnetization	$6.8 \times 10^5 A/m$
α	Gilbert Damping Factor	0.007
t_{ox}	Oxide-layer Thickness	$1.2nm$
RA	MTJ Resistance Area Product	$10.6\Omega\mu m^2$
$(L \times W \times t)_{FL}$	MTJ Free-Layer Dimensions	$40 \times 20 \times 2nm^3$
$(L \times W \times t)_{SHM}$	Spin Hall heavy Metal (SHM) dimensions	$100 \times 40 \times 2.8nm^3$
ρ_{SHM}	Resistivity of SHM (W)	$200\mu\Omega cm^2$
θ_{SHM}	Initial Spin-Hall angle	0.3
TMR_{AP}	Tunnel Magneto Resistance	172%
λ_{sf}	Spin Flip Length	$1.5nm$
P	Electron Polarization Percentage	0.52
λ_{sf}	Spin Flip Length	$1.5nm$
H_k	Anisotropy Field	$80Oe$
$\mu 0$	Permeability of Free Space	$1.25663 \times 10^6 T.m/A$
e	Electric charge	$1.602 \times 10^{-19} C$
h	Reduced Planck's Constant	$6.626 \times 10^{-34}/2\pi J.s$
γ	Gyromagnetic Ratio	$1.76 \times 10^7 (Oe.s)^{-1}$
ϕ	Potential Barrier Height	$0.4V$

Table 5.3: Comparison of the Proposed BGIM cell with other NV-SRAM designs. [4] (* The values are taken from [17])

NV-SRAM Design	NVM Technology	VDD (V)	Back-up Energy	Back-up Delay
[42]	RRAM	1.8	836.2 fJ	10.0 ns
[43]*	STT-MRAM	2.4	10.5 pJ	32.7 ns
[44]*	STT-MRAM	1.6	7.71 pJ	30.0 ns
[45]*	STT-MRAM	1.0	4.78 pJ	35.7 ns
[46]*	STT-MRAM	1.7	8.43 pJ	36.0 ns
[47]*	STT-MRAM	2.5	12.5 pJ	5.0 ns
[48]*	STT-MRAM	1.8	12.7 pJ	25.0 ns
[17]*	SHE-MRAM	1.2	189.7 fJ	2.0 ns
[49]	SHE-MRAM	N/A	492.8 fJ	6.43 ns
BGIM	DSH-MRAM	1.2	121.51 fJ	1.0 ns

Based on the simulation results of the proposed BGIM cell using DSH-MRAM device, the energy consumption of each back-up operation is 121.51fJ, and the energy consumption of each restore operation is 1.56fJ. Furthermore, each back-up operation only requires 1ns and the restore operation can be done in 13.2ps. Circuit operation waveforms of the proposed BGIM cell using DSH-MRAM device is shown in Figure 5.5.

A comparison of the proposed BGIM cell design and other NV-SRAM cells using various NVM technologies is provided in Table 5.3. As it can be observed, among the most energy-efficient NV-SRAM designs listed in Table 5.3, the proposed BGIM cell using DSH-MRAM devices provides $\sim 36\%$ reduction in the energy consumption compared to the lowest energy consuming design, which utilizes SHE-MRAM devices [17]. Furthermore, the proposed BGIM cell using DSH-MRAM devices can perform the back-up or store operation ~ 2-fold faster than the fastest design listed in Table 5.3, which uses SHE-MRAM devices [17].

The energy-efficiency and high performance of the proposed BGIM are due to the fact that it can perform the back-up and store operation with a single write operation on both MTJ devices utilizing a single Spin Hall heavy Metal (SHM). However, the proposed SHE-MRAM-based designs in [17]

and [49] require separate write operations on two MTJ devices using two different SHMs, which can incur extra energy consumption. Moreover, the restore operation is highly reliable because **MTJ0** and **MTJ1** hold differential values, the sense margin is large and as a result, the data stored in the NVM cell can be restored rapidly, reliably, and with high energy-efficiency.

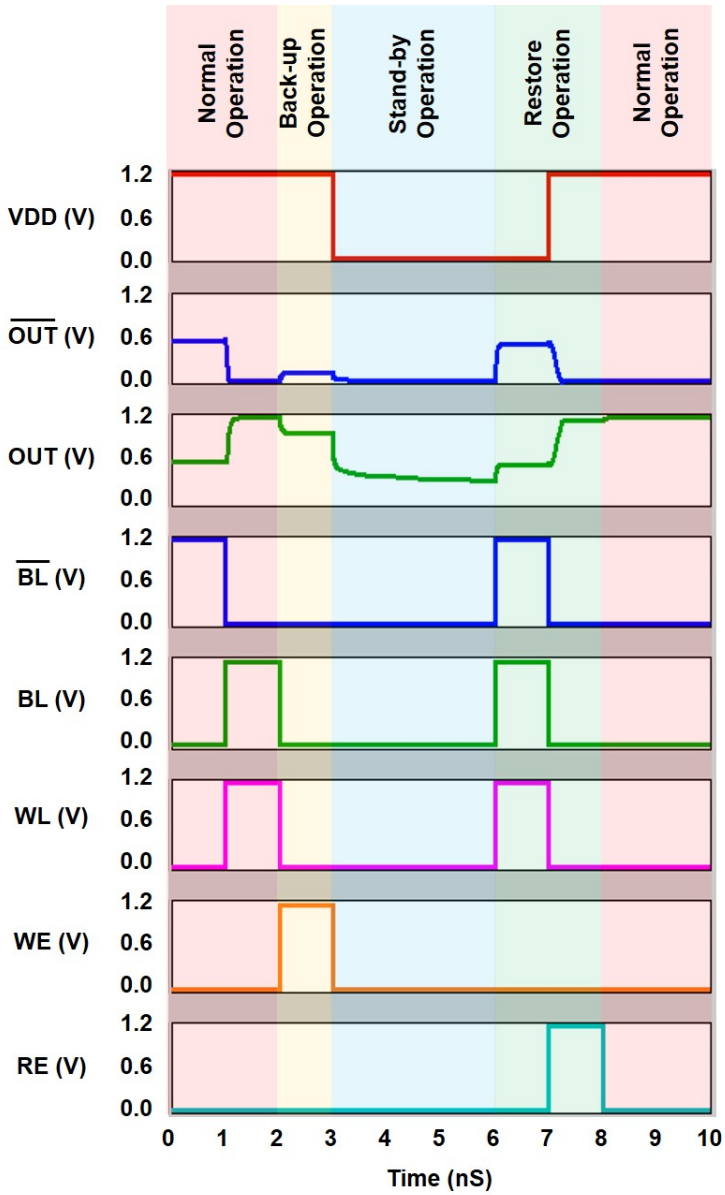
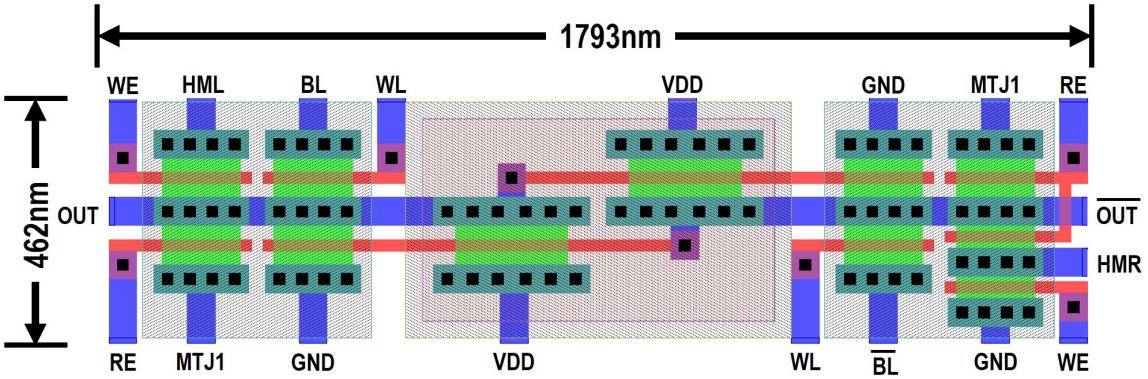


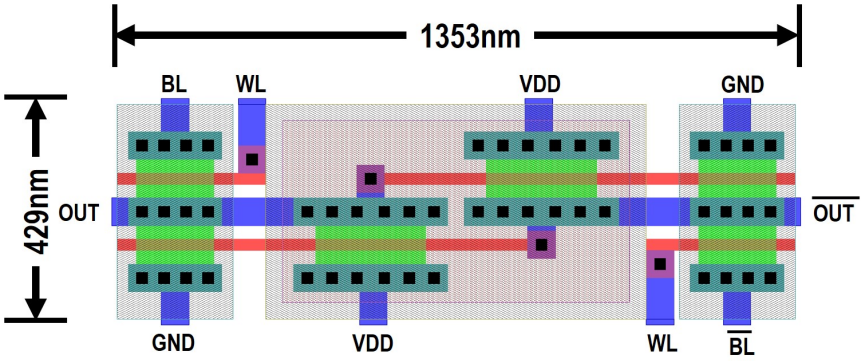
Figure 5.5: Sample simulation waveforms for the proposed BGIM cell using DSH-MRAM device in the presence of parasitic capacitances. [4]

Additionally, the proposed BGIM only incurs $\sim 0.4\mu m^2$ area overhead compared to conventional 6T SRAM cell, due to the addition of 5 access transistors for realizing the one-macro NV-SRAM cell, as shown in the Figure 5.6. The ***HML*** and ***HMR*** terminals shown in Figure 5.6(a) will be connected to the left and right terminals of the SHM terminals, respectively. Additionally, the ***MTJ0*** and ***MTJ1*** terminals shown in Figure 5.6(a) will be connected to the **MTJ0** and **MTJ1**, respectively. Since the DSH-MRAM device can be fabricated on top of the baseline CMOS process, it won't affect the area of the proposed BGIM cell and it is not shown in Figure 5.6(a). It is worth noting that the area overhead can be considered negligible since the need for an extra NVM macro, such as FLASH, which incurs additional energy consumption and delay due to the data movements for each back-up and restore operation, is eliminated.

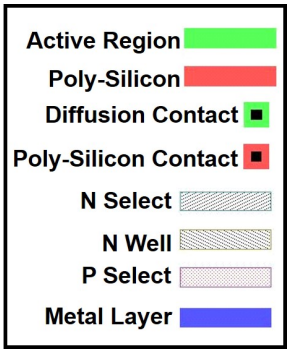
Furthermore, in order to analyze the reliability of the back-up and restore operations of the proposed BGIM cell, Monte Carlo simulation is performed to cover a wide range of Process Variation (PV) cases that may occur in the fabricated device. The MC simulation is performed considering the effects of PV on CMOS peripheral circuit, the SHM, and the MTJ devices. In particular, maximum variation of 10% for the MTJs' resistance levels, which is mainly due to the oxide thickness fluctuations during the fabrication process, along with 10% variation on the threshold voltage and 1% variation on width and length of the CMOS transistors are assessed via MC simulation in agreement with [2]. According to the MC simulation results, the proposed BGIM device provides reliable performance by only incurring 0.14% back-up failure errors. Additionally, since the states of the MTJ devices are differential, they provide a large sense margin and as a result, there are no restore errors. Figure 5.7a depicts the distribution of the back-up time and Figure 5.7b illustrates the distribution of MTJ resistances in P and AP states for the 10,000 MC instances.



(a)

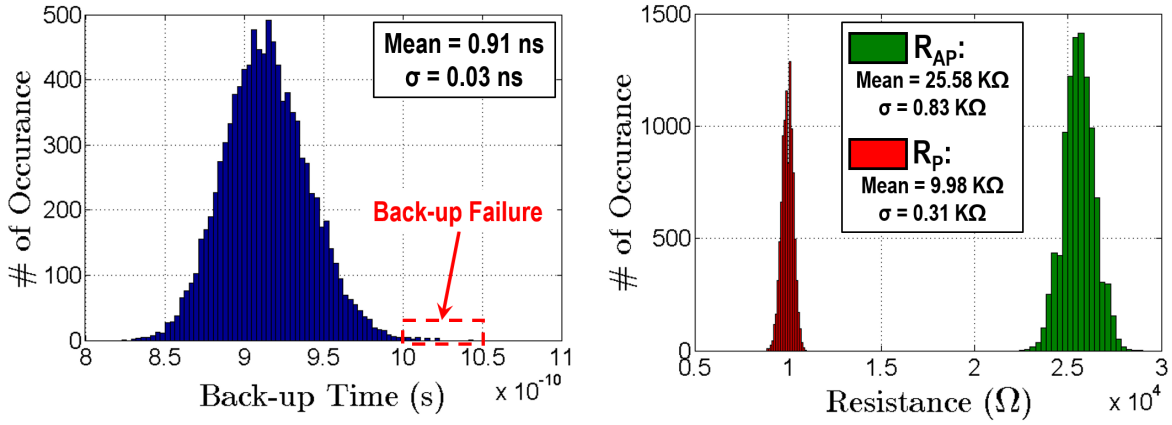


(b)



(c)

Figure 5.6: (a) Layout of the proposed BGIM cell, (b) Layout of a traditional 6T SRAM cell, and (c) Layout legend. [4]



(a) (b)

Figure 5.7: Simulation Results of $10,000$ MC instances for (a) Back-up Time and (b) $\mathbf{R}_{AP}$ and $\mathbf{R}_{P}$ states of the DSH-MRAM. [4]

5.3 Conclusion

A novel energy-aware Non-Volatile Static Random Access Memory (NV-SRAM) framework for sleep power critical mobile applications is devised. The beyond-Complementary Metal Oxide Semiconductor (CMOS) hardware architecture has been designed to minimize the overall static and leakage energy consumption while providing fast back-up and restore operations. Differential Spin-Hall Effect Magnetic Random Access Memory (DSH-MRAM) devices are utilized to realize the proposed framework called Bit-Grained Instant-on Memory Cell (BGIM). Our results indicate that the proposed BGIM consumes 121.51fJ on average for each back-up operation and 1.56fJ on average for each restore operation. Furthermore, the proposed BGIM can perform rapid back-up operations in 1ns and fast restore operations in 13.2ps. Moreover, the proposed BGIM cell incurs $< 1\mu m^2$ area overhead compared to the traditional 6T SRAM cell, however it eliminates the need for data transmission and a separate non-volatile memory macro.

CHAPTER 6: CLOCKLESS SPIN-BASED LOOK-UP TABLES WITH WIDE READ MARGIN[1]

In this chapter, the goal is to address the challenges of previous spin-based LUTs proposed in the literature such as narrow sense margin and low reliability while incurring significant area and power dissipation overheads [56, 57, 58, 59, 60, 61]. Herein, in order to design a spin-based LUT for combinational logic operation without the need for a clock, we develop a clockless 6-input fracturable non-volatile Combinational LUT (C-LUT) with wide read margin using spin Hall effect (SHE)-based Magnetic Tunnel Junction (MTJ) and provide a detailed comparison between the SHE-MRAM and Spin Transfer Torque (STT)-MRAM C-LUTs. Additionally, we provide detailed analysis on the reliability of our proposed C-LUT in the presence of Process Variation (PV).

6.1 Proposed Fracturable 6-Input Clockless LUT

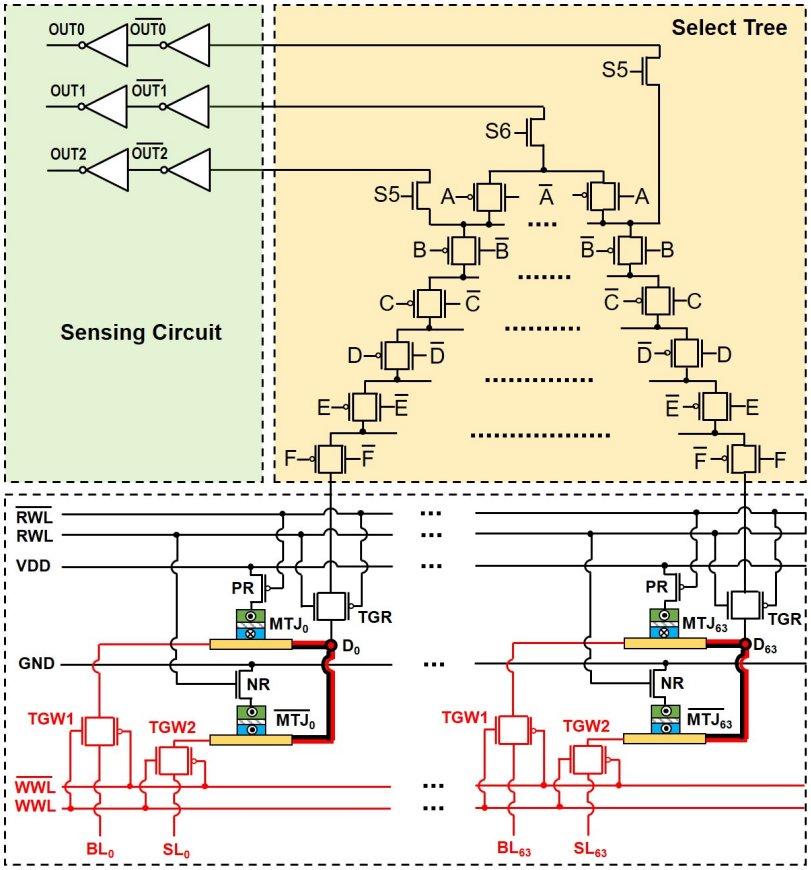
The primary goal of using LUTs in the reconfigurable fabrics is for implementing combinational logic. Generally, M-input Boolean functions are implemented using LUTs that are considered a memory that has 2^M memory cells. The inputs are assigned using a select tree which is constructed with Pass Transistors and Transmission Gates (TGs) [30]. Most contemporary FPGAs, utilize fracturable 6-input LUTs in their design in order to be able to implement one 6-input boolean function or two 5-input boolean functions [217]. Figure 6.1(a) depicts our proposed 6-input fracturable SHE-MRAM C-LUT and Figure 6.1(b) illustrates the 6-input fracturable STT-MRAM C-LUT. In Figure 6.1(a) and Figure 6.1(b), where red color indicates the write path and black color indicates the read path. When the **WWL** and $\overline{\textbf{WWL}}$ signals are asserted, the Write TGs of each memory

[1] ©IEEE. Part of this chapter is reprinted, with permission, from [12]

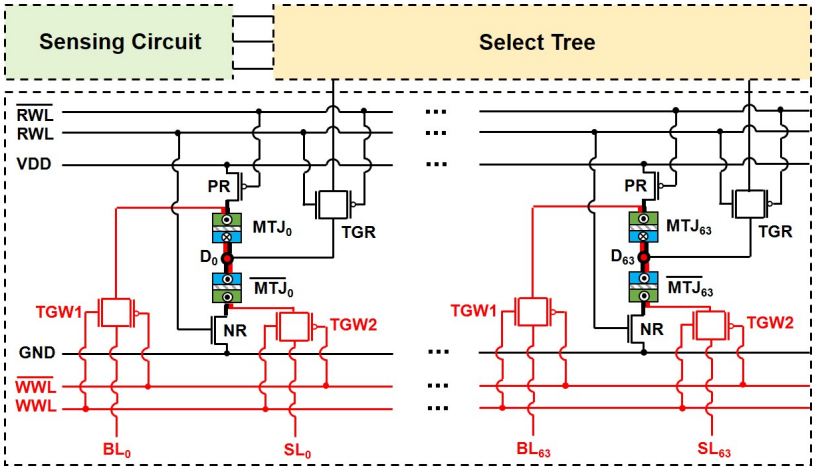
cell, **TGW1** and **TGW2**, will turn on and using Bit Lines, $\mathbf{BL}_i$, and Source Lines, $\mathbf{SL}_i$, we write into both MTJs in each memory cell, $\mathbf{MTJ}_i$ and $\overline{\mathbf{MTJ}_i}$, so that they hold complementary values. If $\mathbf{MTJ}_i$ is in the P state then $\overline{\mathbf{MTJ}_i}$ will be in the AP state and vice versa. This will result in a wide read margin during the read operation.

After the termination of the write operation, in order to read the data stored in the MTJs, **RWL** and $\overline{\mathbf{RWL}}$ signals will be enabled, which results in activation of Read TGs of each memory cell, **TGR**. During the read operation, **PR** and **NR** transistors are turned on when **RWL** and $\overline{\mathbf{RWL}}$ are asserted, which provides the read path from **VDD** to **GND**. The source of **PR**, which is a PMOS transistor, is connected to **VDD** to provide strong one and the source of **NR**, which is an NMOS transistor, is connected to **GND** to provide strong zero. A voltage divider circuit is designed as a result of resistance difference between the $\mathbf{MTJ}_i$ and $\overline{\mathbf{MTJ}_i}$, and the divided voltage can be observed at the $\mathbf{D}_i$ nodes shown in Figure 6.1(a) and Figure 6.1(b). According to the select tree input signals, shown as **A**, **B**, **C**, **D**, **E**, and **F** in Figure 6.1, using two inverters, the voltage on $\mathbf{D}_i$ nodes will be amplified to generate the required output. Since the values stored in the $\mathbf{MTJ}_i$ and $\overline{\mathbf{MTJ}_i}$ devices are complementary, using one MTJ device to retain the data value and the other as the reference value will result in a wide read margin from AP to P [4], which we leverage herein to increase the reliability of the read operation.

In the proposed C-LUT design, there is no need for an external clock or a large sense amplifier circuit. Furthermore, the proposed fracturable C-LUT can perform as a single 6-input LUT or two 5-input LUTs. The Operation mode of the proposed LUT is controlled using **S5** and **S6** signals. If **S5** signal is enabled and **S6** is disabled, then the C-LUT will be operating as two 5-input LUTs and the outputs of the C-LUT will be **OUT0** and **OUT2**. On the other hand, if **S5** signal is disabled and **S6** signal is enabled, then the C-LUT will be operating as a 6-input LUT and **OUT1** will be the C-LUT's output. The proposed fracturable C-LUT provides significantly higher functional flexibility at the expense of slightly more power consumption as studied in Section 6.2.



(a)



(b)

Figure 6.1: The circuit-level diagram of the proposed 6-input fracturable Combinational Look-Up Table (C-LUT) using (a) SHE-MTJ devices and (b) STT-MTJ devices. [12]

6.2 Simulation Framework, Results, and Analysis

Herein, we use the HSPICE circuit simulator to validate the functionality of proposed C-LUT using 45nm CMOS technology and the STT-MRAM model developed by Kim *et al.* in [218]. Figure 6.2(a) and 6.2(b) show the transient response of the C-LUT implementing a 6-input OR operation for $ABCDEF =$ "000000" and $ABCDEF =$ "111111" input signals, respectively. In order to generate the current required for a write delay of less than 2ns, the write transistors are required to be enlarged 4-fold. As shown, the HSPICE simulations verify the correct functionality of our proposed C-LUT. Table 6.1 lists comparison results between the SRAM-LUT and proposed C-LUT in terms of power consumption and delay.

Table 6.1: Comparison between SRAM-LUT and MRAM-LUT. [12]

		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 C-LUT	Logic "0"	14.38	81.16	0.31	20 ps	2 ns
	Logic "1"	19.91	81.25	0.31	60 ps	2 ns
	Average	17.15	81.18	0.31	40 ps	2 ns

The results show more than 80% standby power reduction at the cost of increased write power which can be tolerated due to its infrequent occurrence of write operations in LUTs. There are three energy profiles in the FPGA LUT circuits: (1) Read energy consumption during the FPGA normal operation, (2) Standby energy for the LUTs that are not on the active datapath, which can constitute a significant portion of the FPGA fabric, and (3) write energy that is consumed during the LUTs' configuration operation which occurs rarely.

Table 6.2 provides an area and energy consumption comparison between SRAM-LUT and C-LUT. As listed, the structure of a 6-input MRAM-based C-LUT requires $1,547$ MOS transistors plus

128 MTJs, which can be fabricated on top of the CMOS transistors incurring low area overhead, while the conventional 6-input SRAM-LUT includes 1,029 MOS transistors. This results in an area overhead of roughly 50% for C-LUT compared to SRAM-LUT, which is primarily induced by the write circuits. Thus, innovations are sought to reduce the area and energy consumption of the MRAM cell's write circuit to mitigate these issues.

Table 6.2: Area and Energy Consumption comparison between SRAM LUT and MRAM C-LUT. [12]

	Features	SRAM LUT	MRAM C-LUT
Device Count	Storage Cells	384 MOS	128MTJ
	Write/Control	384 MOS	256×4 + 256 MOS[(1)]
	Read	261 MOS	267 MOS
	Total	1029 MOS	1547 MOS + 128 MTJ
Average Energy Consumption	Read	2.53 fJ	8.58 fJ
	Write	14 fJ	162.36 fJ

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

Recently, SHE-MRAM cells have attracted considerable attentions as an alternative for the conventional STT-MRAMs. Herein, we have used the SHE-MRAM device model proposed by Camsari *et al.* [219] to realize a circuit-level simulation of our SHE-MRAM C-LUT. The results obtained exhibit that a TG-based write circuit with minimum-sized MOS transistors can produce the sufficient write current amplitude required for switching the SHE-MRAM's state in less than 2ns. Thus, table 6.3 provides an iso-delay comparison between STT-MRAM and SHE-MRAM C-LUT in terms of device count and write energy. As listed, the SHE-MRAM C-LUT can achieve more than 49% area reduction, while realizing comparable write energy consumption. Moreover, the SHE-MRAM C-LUT achieves at least 24% device count reduction compared to SRAM-LUT.

Furthermore, to analyze the reliability of the read and write operations of the proposed C-LUT, Monte Carlo (MC) simulation is performed to cover a wide range of PV scenarios that may occur in the fabricated device. The MC simulation is performed with 1,000 instances considering the effects of PV on CMOS peripheral circuit and the MTJs. In particular, variation of 10% for the

MTJs' dimensions along with 10% variation on the threshold voltage and 1% variation on transistors dimentions are assessed.

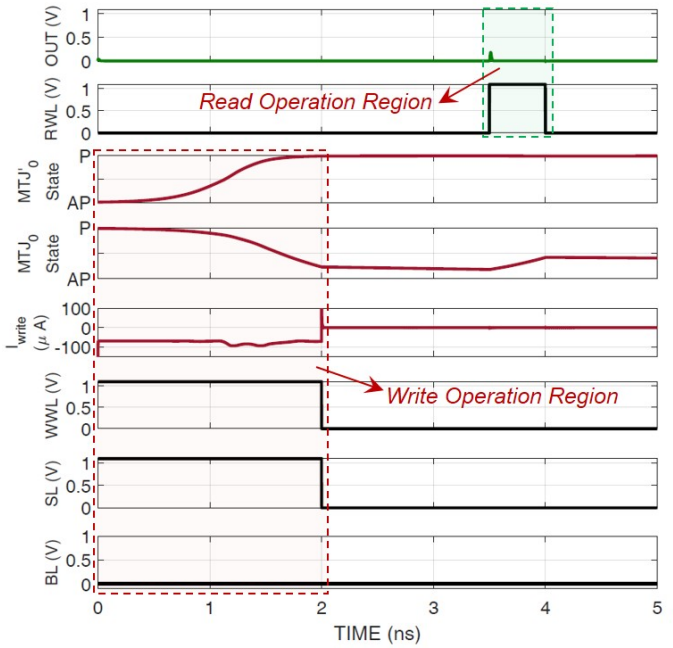
Table 6.3: Iso-Delay Area and Write Energy Consumption comparison between STT-MRAM and SHE-MRAM C-LUTs. [12]

	Features	C-LUT	
		STT-MRAM	SHE-MRAM
Device Count	Storage Cells	128MTJ	128MTJ
	Write/Control	(256×4)+256MOS [(1)]	256+256MOS [(2)]
	Read	267MOS	267MOS
	Total	1547MOS+128MTJ	779MOS+128MTJ
Average Write Energy per Cell		162.3 fJ	175.5 fJ

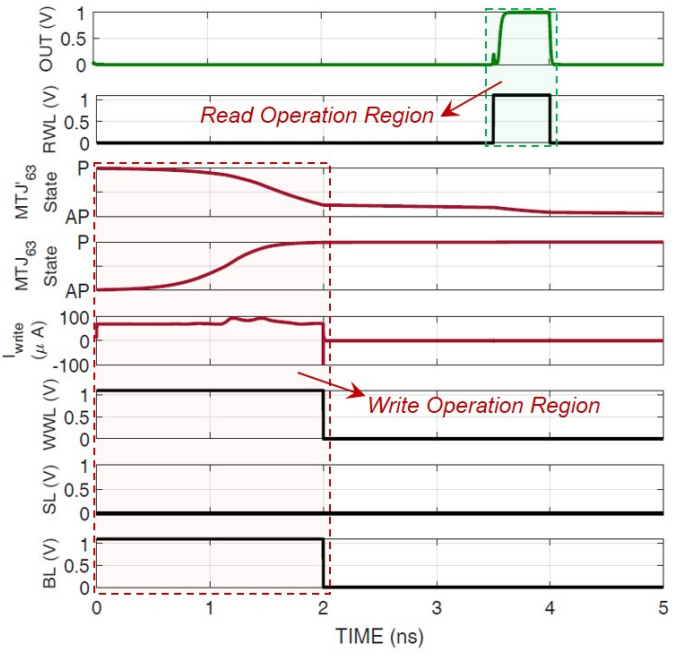
[(1)] Write transistors are $4\times$ larger than minimum feature size.
[(2)] Write transistors with minimum feature size are used.

Figure 6.3(a) depicts the distribution of the switching times for T_{P-AP} and T_{AP-P}, Figure 6.3(b) illustrates the distribution of MTJ resistances in R_{AP} and R_P states, and Figure 6.3(c) shows the distribution of read, I_{READ}, and write, I_{Write} currents for the $1,000$ MC instances. According to the MC simulation results, C-LUT provides reliable write performance resulting in less than 0.001% write errors in $1,000$ error-free MC instances. In particular, results of the MC simulation show that the switching time for $P-AP$ is 1.63ns on average and the switching time for $AP-P$ is 1.13ns on average, which both fall under the 2ns duration of the write operation, as depicted in Figure 6.3(a).

Additionally, since the states of the MTJs are differential, they provide a wide read margin and as a result, there are less than 0.001% read errors caused by PV based on the $1,000$ error-free MC simulation results. Furthermore, our proposed C-LUT does not suffer from read disturbance due to the small read current compared to the write current as shown in Figure 6.3(c). According to our MC simulation results, the read current is 38.21μA on average, which is significantly lower than the write current that is 71.13μA on average.



(a)



(b)

Figure 6.2: Transient response of C-LUT implementing 6-input OR operation for (a) $ABCDEF$ = "000000" input signal, and (b) $ABCDEF$ = "111111" input signal. [12]

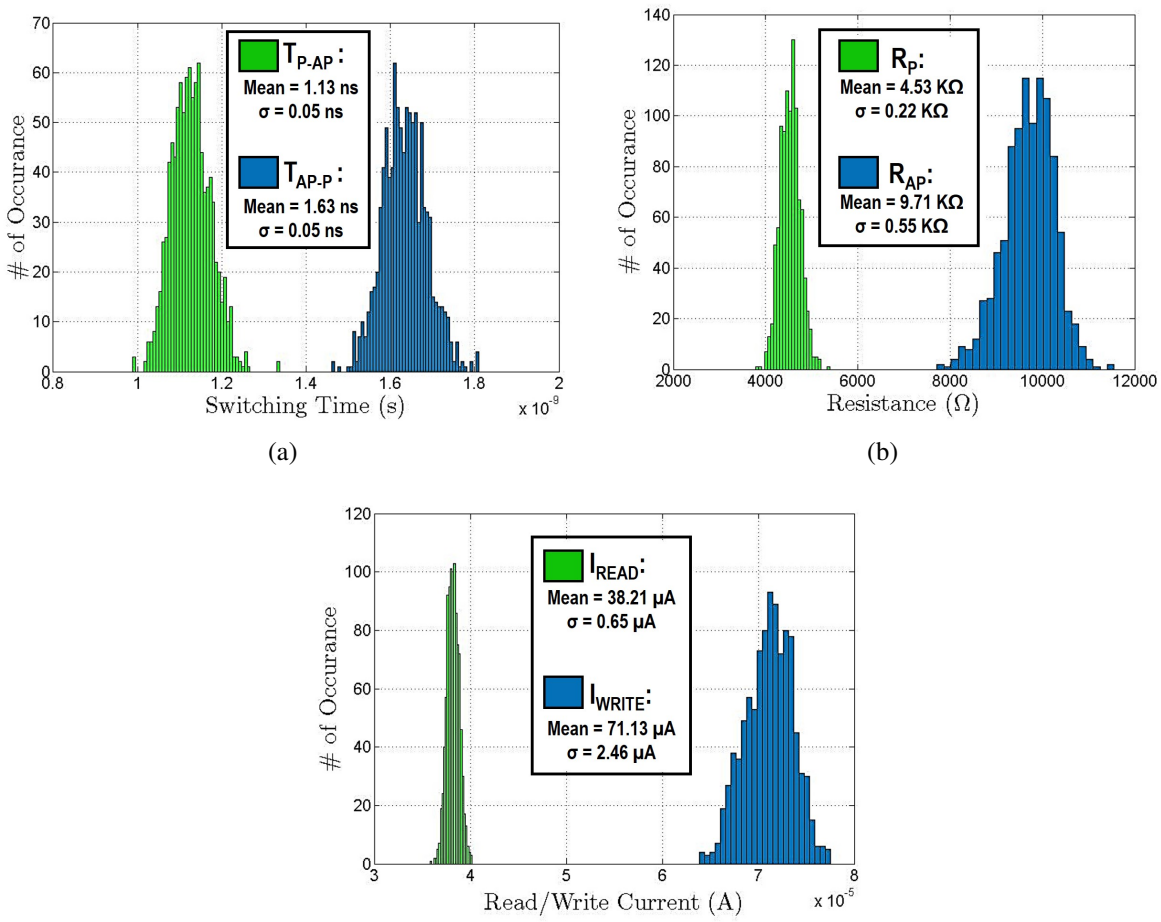


Figure 6.3: Simulation Results of $1,000$ MC instances for (a) T_{P-AP} and T_{AP-P} Switching Times, (b) R_{AP} and R_P resistance states, and (c) read, I_{READ}, and write, I_{Write} currents. [12]

6.3 Conclusion

A 6-input fracturable non-volatile Clockless LUT (C-LUT) using spin Hall effect (SHE)-based Magnetic Tunnel Junctions (MTJs) is developed and a detailed comparison between the SHE-MTJ-based C-LUT and Spin Transfer Torque (STT)-MTJ-based C-LUT is provided. The proposed C-LUT offers an attractive alternative for implementing combinational logic as well as sequential logic versus previous spin-based LUT designs in the literature. Foremost, C-LUT eliminates the

sense amplifier typically employed by using a differential polarity dual MTJ design, as opposed to a static reference resistance MTJ. This realizes a much wider read margin and the Monte Carlo simulation of the proposed fracturable C-LUT indicates no read and write errors in the presence of a variety of process variations scenarios involving MOS transistors as well as MTJs. Additionally, simulation results indicate that the proposed C-LUT reduces the standby power dissipation by 5.4-fold compared to the SRAM-based LUT. Furthermore, the proposed SHE-MTJ-based C-LUT reduces the area by 1.3-fold and 2-fold compared to the SRAM-based LUT and the STT-MTJ-based C-LUT, respectively.

CHAPTER 7: ENERGY-AWARE ADAPTIVE RATE AND RESOLUTION SAMPLING OF SPECTRALLY SPARSE SIGNALS LEVERAGING VCMA-MTJ DEVICES[1]

In this Chapter, we devise an adaptive framework for efficient acquisition of spectrally sparse signals utilizing emerging spin-based devices. In the first contribution herein, we propose a Spin-based Adaptive Intermittent Quantizer (AIQ) to perform adaptive signal sampling and quantization. AIQ utilizes Voltage-Controlled Magnetic Anisotropy Magnetic Tunnel Junction (VCMA-MTJ) devices to provide fast SR and adaptive QR in a novel energy-efficient fashion. By leveraging non-volatility, a spin-based AIQ can reduce energy consumption via instant off/on operation without use of a backing store.

The second contribution herein focuses on investigating the trade-offs between SR and QR under power and bandwidth constraints using dynamic optimization of SR and QR. The energy consumption, hardware limitations, and specifics of the underlying sampler and quantizer become central to system optimization.

The proposed beyond-CMOS hardware architecture and corresponding adaptive quantized CS techniques are considered in synergy with each other. Together, these are used to minimize the *overall cost of signal acquisition* which is later formulated as a combination of the amount of dynamic energy consumed in hardware for acquisition (energy constraint) and the number of bits acquired for each frame (bandwidth constraint) within the reconstruction error (MSE) for a spectrally-sparse input signal.

[1] ©IEEE. Part of this chapter is reprinted, with permission, from [5]

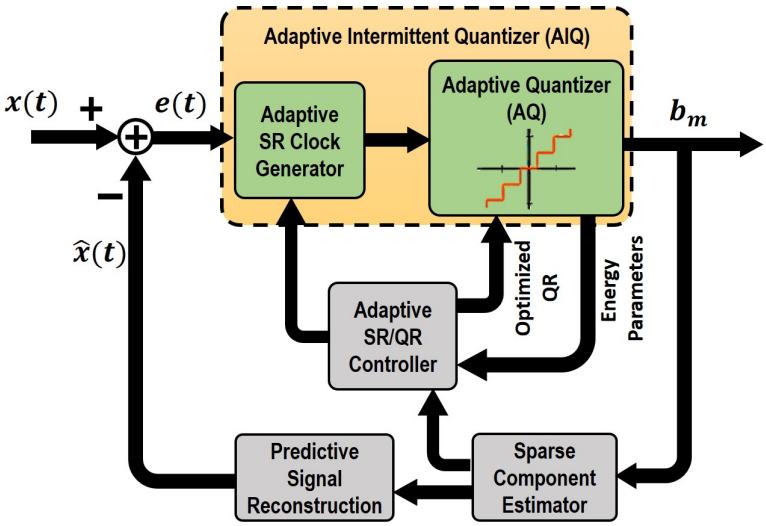


Figure 7.1: The system-level block diagram of the proposed signal acquisition [5].

7.1 Proposed Cross-Layer Approach

In this Chapter, we develop a novel cross-layer device/circuit/architecture design for adaptive signal sampling, reconstruction, and the enabling hardware for energy-efficient acquisition of wideband spectrally sparse signals. First, a framework for smart and adaptive determination of the sampling rate and quantization resolution based on the instantaneous signal and hardware constraints is introduced. Second, we develop a spin-based *Adaptive Intermittent Quantizer* (AIQ) to facilitate the realization of the adaptive sampling proposed herein. Figure 7.1 shows the system-level diagram for our proposed design. In this figure, the input signal $x(t)$ is compared with the estimate signal $\widehat{x}(t)$. The error signal $e(t)$ then goes through our proposed *AIQ* which samples each frame of the input at a specific sample rate, i.e. frame n_f is sampled at $t = m\tau^{(n_f)}$, quantized to symbols c_m and subsequently to the corresponding bit stream b_m. Note that $\tau^{(n_f)}$ is the sampling interval, which is adaptively determined for frame n_f of the signal.

The *adaptive Sample-Rate (SR) / Quantization-Resolution (QR) controller* is a key innovation of our approach. This block optimizes SR ($\frac{1}{\tau^{(n_f)}}$Hz) and the number of digital bits used to quantize

each sample (QR) for each frame of the input signal. For that, this block utilizes the signal parameters (e.g., sparsity, noise level) estimated at the previous frame and hardware-level constraints (e.g., energy, bandwidth). This block provides the optimized clock period and bit depth for the next frame of the signal. The same block is present at the receiver to extract bit-depth resolution and the sampling rate from the received sequence of bits. Components of our design are described below.

Spin-based devices have been extensively researched as promising companions to CMOS devices. As CMOS scaling trends continue, the need to identify viable approaches for reducing leakage power increases. With attributes of non-volatility, near-zero standby energy, and high density, Magnetic Tunneling Junction (MTJ) has emerged as a promising alternative post-CMOS technology for embedded memory and logic applications [1, 10, 30, 73]. The basic concept of spin-based Non-Volatile Memory (NVM) devices is to control the intrinsic spin of electrons in a ferromagnetic solid-state nano-device. Recent research studies have shown that the use of Voltage-Controlled Magnetic Anisotropy (VCMA) effect facilitates the use of an electric field to ease or eliminate the demand of charge current for switching the state of MTJ devices. As a result of using VCMA-MTJ devices, the majority of the dynamic power dissipation caused by ohmic losses and joule heating during the switching of the spin-based devices can be significantly reduced [73, 74, 75, 76, 77, 78].

Adaptive Intermittent Quantizer (AIQ): Herein, to implement the adaptive rate/resolution sampling, a recently-developed type of spin-based device, namely the VCMA-MTJ, is utilized to provide faster and more energy-efficient signal sampling and quantization. Previously, emerging spin-based technologies have been explored as an alternative to CMOS technology for embedded and data storage applications due to their non-volatility, near-zero standby energy, and high density. These emerging devices, such as Spin Transfer Torque Magnetic RAM (STT-MRAM) and Spin-Hall Effect Magnetic RAM (SHE-MRAM), have been the focus of the research in recent years [1, 10, 30, 73, 79, 80, 81]. Using spin-based devices can increase energy efficiency via a signif-

icant reduction in leakage energy. Furthermore, these devices offer small area footprint and can be fabricated in 3D stacks on top of baseline CMOS design using the same backend fabrication process. A detailed explanation of this block is provided in Section 7.2.

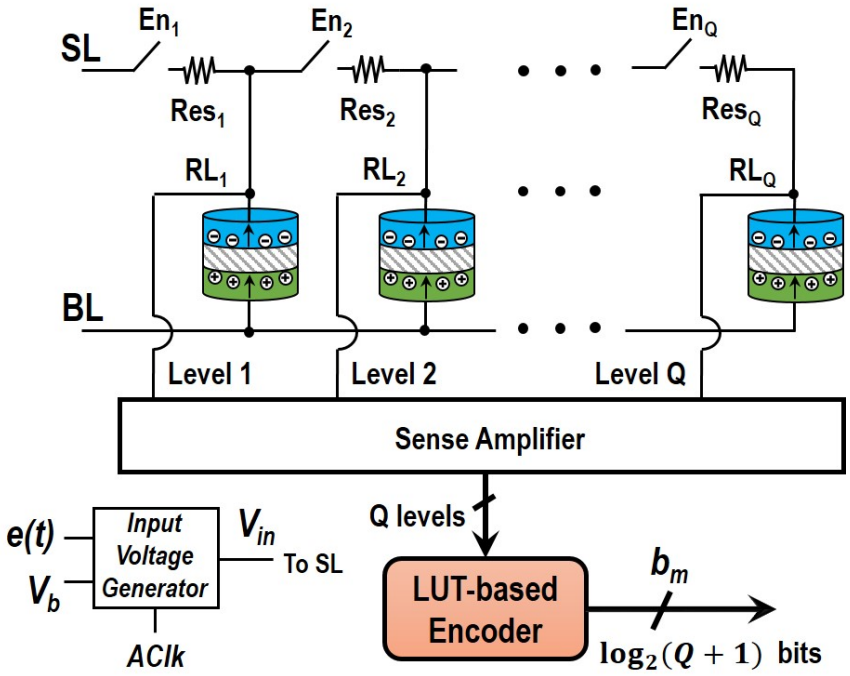


Figure 7.2: The Proposed AIQ Architecture [5].

Adaptive SR/QR Optimization: The concept of energy-aware SR/QR optimization is motivated by the fact that, in any practical scenario, sensing operations need to be able to satisfy the power and bandwidth constraints. Under a bandwidth constraint, the sensing device is constrained by a bit-rate when transmitting or storing the signal. On the other hand, the energy supply might impose strict constraints on the SR and/or QR. Thus, it is desirable to have a system that adapts SR and QR to maximize the sensing performance in the long run, while considering the power and bandwidth constraints.

7.2 Intermittent Spin-based Adaptive Quantizer Using VCMA-MTJ Devices

In recent studies, researchers have exploited the use of emerging devices for signal processing applications. In particular, they have explored designing ADCs using emerging devices such as SHE-MTJ [79], Domain Wall Motion (DWM) [82], and Racetrack Memory [112]. Herein, we propose an *Adaptive Intermittent Quantizer (AIQ)* to perform signal sampling and quantization. AIQ uses VCMA-MTJ devices to provide fast SR and adaptive QR, along with energy-efficient sampling and quantization operations. Use of VCMA-MTJs enables AIQ to provide various quantization levels by changing the energy barrier of MTJ devices. An example of Q-level AIQ architecture is shown in Figure 7.2, where Q is the number of QR levels determined by the optimization algorithm. The operation of AIQ has three main steps:

- First, during the Reset step, all active VCMA-MTJ devices will be reset to zero representing a Parallel state,
- Second, during the Sampling step, based on the determined SR and QR , first a bias voltage, V_b, will be applied across the active VCMA-MTJ devices' terminals to modify and set their energy barrier followed by the analog input, $e(t)$, to write into the active VCMA-MTJ devices, as shown in Figure 7.1, and
- Third and final step is the Read (or Sensing) step to sense the data stored in each device using a sense amplifier in a conventional fashion.

Based on the architecture shown in Figure 7.2, during the Reset step, Source Line (SL) is set to zero, Bit Line (BL) is set to one, and Read Lines (RLs) are high impedance, which causes all devices to go to the P state. During the Sampling step, SL is set to input voltage (V_{in}), BL is set to zero, and RLs are high impedance. In this state, an Input Voltage Generator circuit is used to

allow the VCMA bias voltage, V_b, followed by the analog input, $e(t)$, to be applied through V_{in} to adjust the threshold of MTJ devices and write into the MTJs. A signal called Adaptive Clock (AClk), which is set based on the $\tau^{(n_f)}$ as described in Section 7.1, will control the sampling rate of the input signal. During the Read (or Sensing) step, SL is set to high impedance, BL is set to zero, and RLs are sent to sense amplifiers to read the value stored in each MTJ. The design of the sense amplifiers for an MTJ read operation is discussed broadly in the literature [1].

The combination of switches and resistors included in our proposed architecture is used to realize the *adaptive quantization resolution levels*. The switch ladder is used to adaptively set the resolution and the resistance ladder is used to provide different VCMA bias voltages, V_b, for different MTJs. By providing different bias voltage levels for different MTJs, some MTJs turn on with lower input voltages while some require higher input voltages to switch state. Furthermore, the switches, which are realized using transmission gates in order to provide reliable switching [10], enable the Adaptive SR/QR Controller shown in Figure 7.1 to optimize the QR by turning unused MTJs off. As demonstrated subsequently, this results in significant energy savings.

A Pre-Charge Sense Amplifier (PCSA) [83] is used to read the value stored in the SHE-MTJ devices. The AIQ circuit provides different QR levels. A Look-up Table (LUT)-based encoder is used to encode the values for different levels into bits. For instance, the example shown in Figure 7.2 can provide 1 bit with 1 level, 2 bits with 3 levels, 3 bits with 7 levels, and so forth. Since the number of active components of the LUT-based encoder depends on the number of active levels, spin-based devices have also been utilized within the encoder structure. Correspondingly, depending on how many QR levels our algorithm is using, we can adaptively disable the parts of the LUT-based encoder that are not being used in the encoding process. This will lead to significant energy savings and improved performance as shown in [30] compared to conventional CMOS encoders since spin-based devices offer zero leakage energy consumption.

The behavior of a single VCMA-MTJ device is demonstrated in Figure 2.4. As it is observed, different values of V_b results in different energy barrier heights. Different energy barrier heights result in different switching behavior for the VCMA-MTJ devices. In our proposed AIQ design, we have utilized an example of 255 VCMA-MTJ devices to realize a wide range of quantization resolutions from 1-bit to 8-bit ADC operation. Additionally, different V_b values will be applied to the active VCMA-MTJ devices to realize discriminable quantization resolutions. Moreover, for 1-bit resolution, one level is used, which is set to 650mV considering a signal range that is normalized between $[0 - 1.3]$V. Additionally, the levels are spaced by 542mV, 201mV, 90mV, 43mV, 21mV, 10mV, and 5mV for 2-bit, 3-bit, 4-bit, 5-bit, 6-bit, 7-bit, and 8-bit resolution levels, respectively.

7.3 Simulation Results and Analysis

7.3.1 AIQ Sampling Results and Performance Analysis

In order to evaluate and validate the behavior and functionality of the proposed AIQ design, SPICE and MATLAB simulations were performed. We have utilized the 22nm Predictive Technology Model (PTM) [202] as well as VCMA-MTJ model represented in [73] along with other circuit parameters and constants listed in Table 7.1 in our simulations to implement and evaluate the proposed AIQ design.

To examine the performance and potential of the VCMA-MTJ devices in circuit designs and applications, the circuit behavior of VCMA-MTJ devices maintaining resistance in P ($\theta = 0^\circ$) and AP ($\theta = 180^\circ$) states as well as the voltage-dependent TMR effect are modeled by Kang, *et al.* [73] and expressed using the following equations [10, 73]:

$$R_P = \frac{t_{ox}}{Factor \times Area \cdot \sqrt{\phi}} \exp\left(\frac{2\sqrt{2me}}{\hbar} \times t_{ox} \cdot \sqrt{\phi}\right), \tag{7.1}$$

$$TMR(V_b) = \frac{TMR(0)}{1 + (\frac{V_b}{V_h})^2}, \tag{7.2}$$

$$R_{MTJ}(V_b) = R_P \frac{1 + (\frac{V_b}{V_h})^2 + TMR(0)}{1 + (\frac{V_b}{V_h})^2 + TMR(0)[0.5(1 + \cos(\theta))]}, \tag{7.3}$$

where V_b is the bias voltage, $TMR(V_b)$ is the Tunnel Magneto-Resistance (TMR) ratio, $V_h = 0.5V$ is the bias voltage when TMR ratio is half of the $TMR(0)$, t_{ox} is the oxide thickness of MTJ, $Factor$ is obtained from the resistance-area product value of the MTJ that relies on the material composition of its layers, $Area$ is the surface area of the MTJ, and ϕ is the oxide layer energy barrier height. The switching of the perpendicular magnetization of the VCMA-MTJ's free-layer is determined by θ is the polar angle of the magnetization vector of the free-layer, $\vec{m}$. In other words, $m_z = cos(\theta)$ provides the component of the magnetization vector, $\vec{m}$, along the z-axis of the Cartesian coordinate system. The parameters and constants used in the VCMA-MTJ model for the simulation results are provided in Table 7.1 [73].

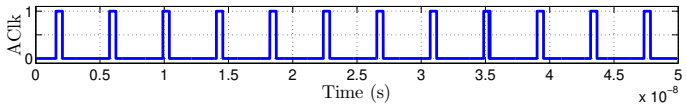
As depicted in Figure 7.3(b), a growing sinusoidal signal is sampled by 3 levels to 2 bits based on the AClk signal, shown in Figure7.3(a), with 12 sampling intervals resulting in the bit budget of $\mathcal{B}^{(n_f)} = 24$. Additionally, Figure 7.3(c) illustrates the switching of each of the 3 VCMA-MTJ devices with different switching energy barriers resulting in different levels. According to our results, the energy consumption of this sampling configuration equals 596.31fJ, which consists of the reset, sample, and read operations as well as the peripheral circuitry energy consumption during the 50ns signal duration. The corresponding quantized CS reconstruction algorithm achieves a Mean Square Error (MSE) of 4.7×10^{-5} on this signal which proves efficient reconstruction capability of the proposed design. Furthermore, Figure 7.4(b) depicts sampling of the same growing sinusoidal using the AClk signal with 8 sampling intervals, as shown in Figure 7.4(a), to achieve 3 bits resolution while maintaining the same bit budget of $\mathcal{B}^{(n_f)} = 24$. Moreover, Figure 7.4(c)

demonstrates the switching of each of the 7 VCMA-MTJ devices. The energy consumption of this configuration is 906.39fJ during the 50ns signal duration. The proposed reconstruction algorithm achieves an MSE of 1.2×10^{-4} in this case.

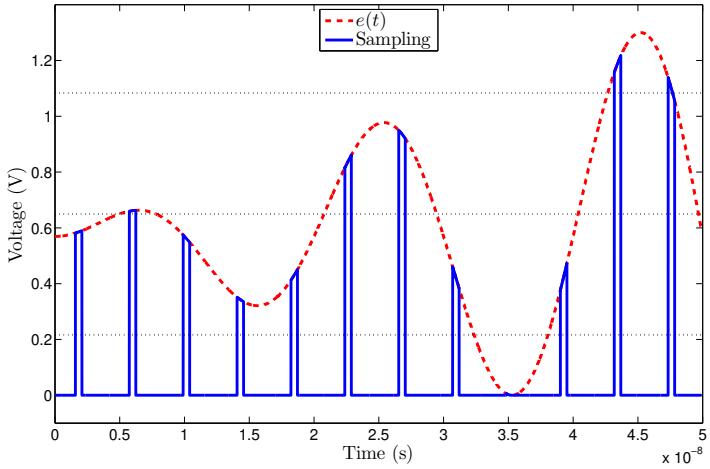
Table 7.1: Circuit parameters and constants values for the VCMA-MTJ model [5].

Parameter	**Description**	**Default Value**
M_s	Saturation magnetization	$0.625 \times 10^6 A/m$
$K_i(0)$	Initial interfacial PMA energy	$0.32 mJ/m^2$
t_f	Free-layer thickness	$1.1nm$
α	Gilbert damping factor	0.05
$\Delta(0)$	Thermal stability factor at $V_b = 0$	40
T	Temperature	$300K$
ξ	VCMA coefficient	$60 fJ/V \cdot m$
t_{ox}	Oxide-layer thickness	$1.4nm$
H_x	External Magnetic Field	$4.8 \times 10^4\ ^\circ/m$
P	STT polarization factor	0.58
d	MTJ diameter	$50nm$
ϕ	Potential barrier of MgO	$0.4V$
$TMR(0)$	TMR ratio at $V_b = 0$	200%
V_h	Bias Voltage at TMR2	$0.5V$
Constants	**Description**	**Default Value**
γ	Gyromagnetic ratio	$2.21276 \times 10^5 m/(A \cdot s)$
k_B	Boltzmann constant	$1.38 \times 10^{-23} J/K$
μ_0	Vacuum permeability	$1.2566 \times 10^{-6} H/m$
m	Electron mass	$9.11 \times 10^{-31} kg$
e	Elementary charge	$1.6 \times 10^{-19} C$
$\hbar$	Reduced Planck constant	$1.054 \times 10^{34} Js$

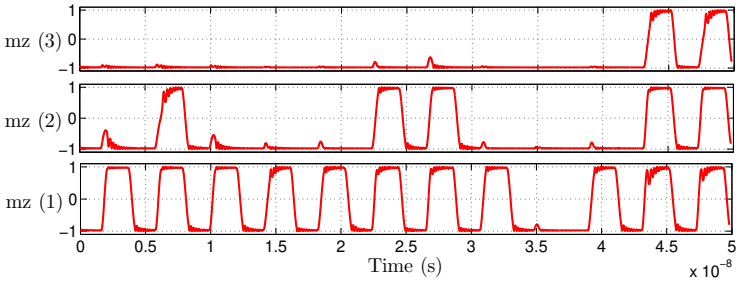
It is observed that as the bit budget is fixed in the experimental scenarios of Figure 7.3 and Figure 7.4, an increase in the number of QR results in a decreased SR considering the SR / QR trade-off.This is observed in Figure 7.3(b) and Figure 7.4(b) when the number of samples is decreased from 12 to 8 in the provided snapshot of the signal. The MSE values achieved show that for the experiment parameters (noise, power, bit budget, etc.), an increased number of coarsely quantized samples, as shown in Figure 7.3, perform better than accurately quantized samples acquired at a decreased rate, as shown in Figure 7.4.



(a)



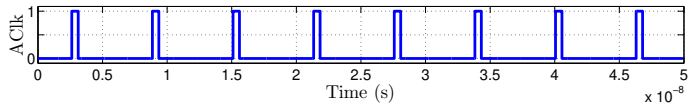
(b)



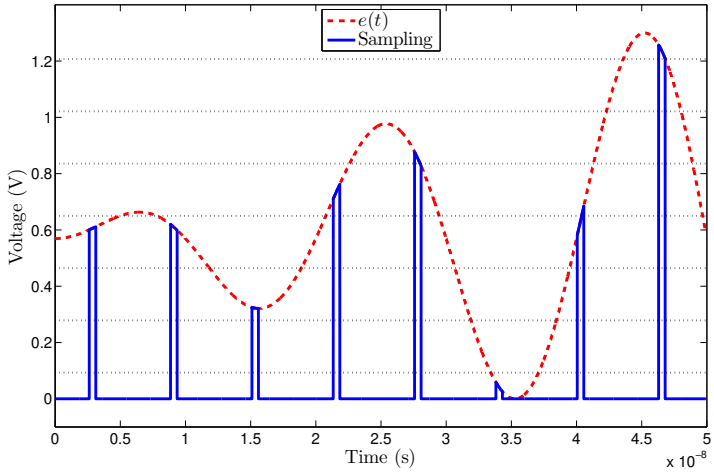
(c)

Figure 7.3: (a) shows the AClk signal over time, (b) depicts the $e(t)$ signal being sampled with 2 bits (3 levels) with 12 sampling intervals, and (c) illustrates the switching of the 3 VCMA-MTJ devices in the sampling intervals [5].

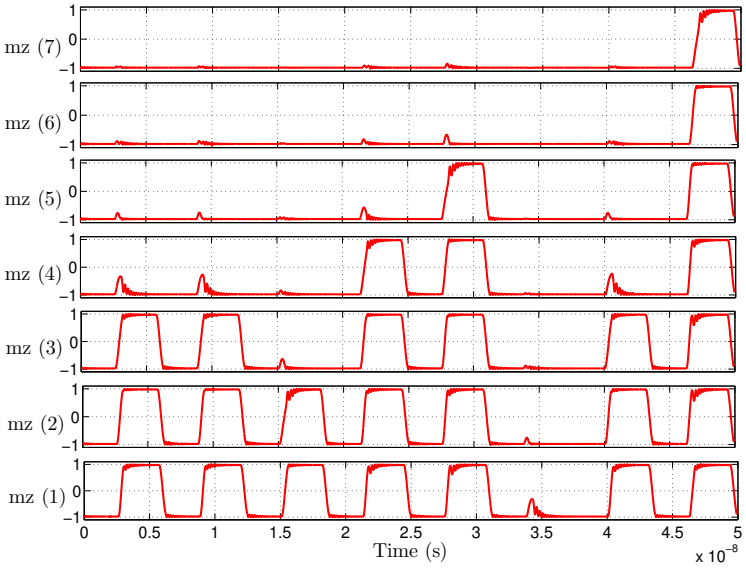
According to our results, the energy consumption of each VCMA device is $\sim$ 17fJ, which consists of a reset and a sample operation for a single VCMA-MTJ device. Meanwhile, the energy consumption of the peripheral circuit that sets the VCMA bias voltages and performs the read operation is $\sim$ 2fJ. Figure 7.5 illustrates the energy consumption versus QR for 22nm technology node, considering two different sampling rates of 5 samples and 10 samples within the same sampling duration of 50ns.



(a)



(b)



(c)

Figure 7.4: (a) shows the AClk signal over time, (b) depicts the $e(t)$ signal being sampled with 3 bits (7 levels) with 8 sampling intervals, and (c) illustrates the switching of the 7 VCMA-MTJ devices in the sampling intervals [5].

It can be observed that for every extra resolution bit, the number of VCMA-MTJs added to the design to provide required QR levels grows exponentially. As a result, the number of reset, sample, and read operations will increase based on the number of active levels. It is known that the lower bound for power of ADCs grows exponentially for every bit of resolution [220, 221]. Thus, QR plays a crucial role in the energy cost of the device. The energy consumed by the proposed MTJ devices can be simplified to a formula to calculate and estimate the amount of dynamic energy consumption for each frame as $E_L \times 2^{\beta} M$, where E_L is the dynamic energy per QR level that is a technology dependent value. According to our simulation results, the E_L value equals 16.63fJ. Hence, the energy per frame is given by $16.63 \times 2^{\beta} M$, where β is the number of bits and M is the number of samples.

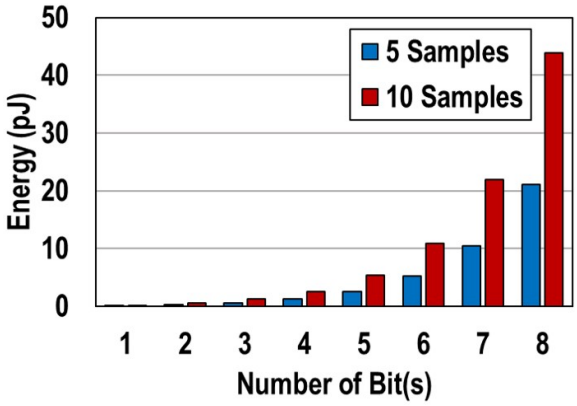


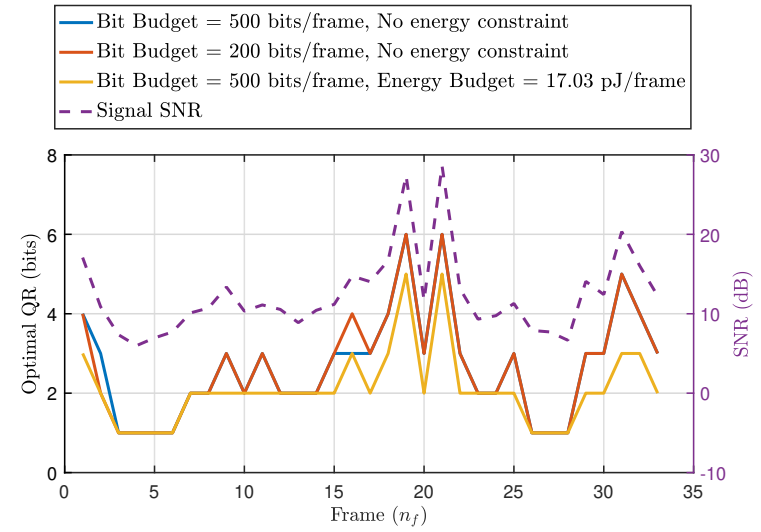
Figure 7.5: Energy consumption versus Quantization Resolution (QR) [5].

Accordingly, for every VCMA-MTJ read, write, and reset operations, approximately 16fJ is required using a 22nm technology node library. The aforementioned energy equation can be employed in the SR/QR optimization process. According to our results, it can be estimated that using VCMA-MTJ devices, overall reset, sample, and read operations would require about 1ns in 22nm technology node library to provide a reliable outcome. As the results show, increasing the QR can increase the energy consumption due to the increase in the number of active MTJ devices. However, by decreasing SR if possible, in cases where increased QRs are required, energy consumption

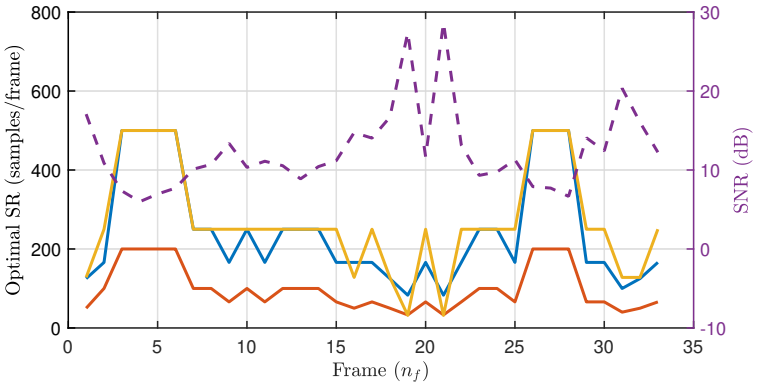
can be decreased. Additionally, an increase in the SR can result in an increase in the energy consumption of AIQ. This is because an increase in SR requires fast reset, sampling, and read steps. Hence, the MTJ devices require to be demagnetized at a faster pace, which can incur extra energy cost. This would be exacerbated if an increase in QR is required, since additional devices will need to be rapidly demagnetized. Overall, energy consumption in the hardware is not simply a function of the bit budget, i.e., $\mathcal{B}^{(n_f)}$. Rather, it is a complex function of its components SR and QR as well as circuit elements and peripherals that are added for every additional quantization level. Herein, we investigated and formulated the hardware energy cost and trade-offs as a function of SR and QR and utilized the results in the proposed cross-layer energy-aware SR/QR optimization.

7.3.2 SR and QR Optimization

Furthermore, to illustrate the necessity of adapting SR and QR during acquisition, we plot the optimal QR and SR values versus the frame number in Figure 7.6. Note that in this simulation scenario, the variance for the $e(t)$, which is the input to our proposed AIQ block, is decreasing with the frame number. This is because $l(t)$ becomes increasingly accurate estimate of $x(t)$ along iterations. However, the input noise is considered random. The resulting SNR along with the QR and SR values that minimize the performance upper bound and the energy bound are provided in Figure 7.6. As the SNR of the signal varies over time, the controller needs to tune the SR and QR to minimize the error metric. It is also worthwhile to point out the fact that, due to exponential growth of the energy with QR, the energy constraint prevents us from sampling the signal with high QR. Thus, adding an energy budget to the simulation places a limit on the QR. These results further encourage adaptive and energy-aware adjustments of SR and QR to improve the performance of the proposed signal acquisition process.



(a)



(b)

Figure 7.6: (a) Optimal QR and (b) optimal SR for different frames. The dashed line shows the SNR of the signal [5].

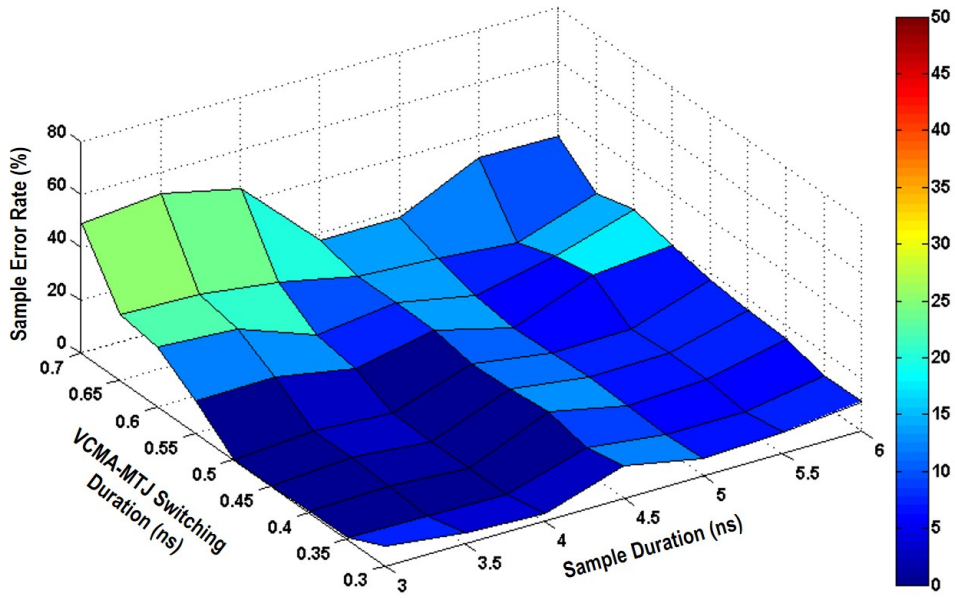
7.3.3 Reliability Analysis

In order to evaluate the functionality of our proposed AIQ design in the presence of Process Variation (PV), we have conducted a series of Monte Carlo (MC) simulations with $10,000$ instances for the sample operation and $10,000$ instances for the read operation. During the MC simulation, we have considered 10% variation for the components of the peripheral circuitry such as threshold voltage of the CMOS transistors as well as 1% variation for the MTJ devices in agreement with

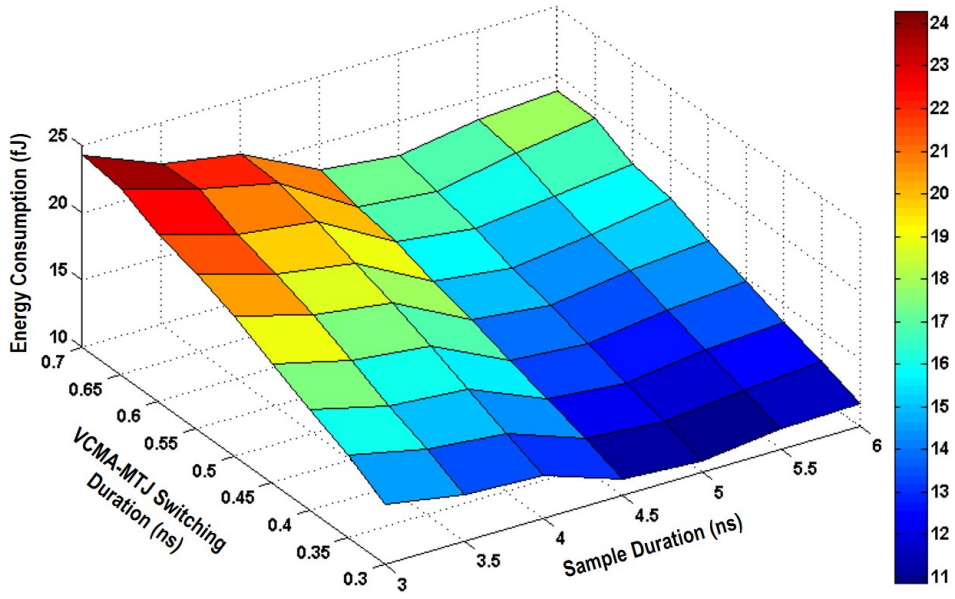
[2]. This can cover a wide range of possible variations enabling a comprehensive PV analysis.

We have analyzed our circuit separately for sample operation, as well as read operation. The results of the reliability analysis for the sample operation are shown in Figure 7.7. As it can be observed from Figure 7.7(a), for sampling duration within the range of 3ns to 3.5ns with VCMA-MTJ switching duration within the range of 0.4ns to 0.5ns, depicted as dark blue region in Figure 7.7(a), the sample error rate is near 0.0%. However, in order to minimize the energy consumption of the sample operation, sample duration should be within the range of 5ns to 5.5ns with VCMA-MTJ switching duration within the range of 0.3ns to 0.35ns according to Figure 7.7(b). Hence, there is a trade-off between sample error rate and energy consumption. Thus, the dark blue region in Figure 7.7(b) reflects a reduced energy consumption at the expense of the corresponding sample error rate indicated in Figure 7.7(a).

Furthermore, the results of the reliability of the read circuit are provided on Figure 7.8. Herein, we have conducted the reliability analysis for four of the most commonly-used approaches for sensing according to the study presented in [2]. As shown in Figure 7.8, the Variation Immune Sense Amplifier (VISA) proposed in [10] and the Separated Pre-Charge Sense Amplifier (SPCSA) proposed in [117], provide highly-reliable outputs considering Tunnel Magnetoresistance Ratio (TMR) of 200% by only incurring 0.05% and 0.07% error rate during the read operation, respectively. However, VISA and SPCSA incur large area and energy consumption overheads compared to the Energy Aware Sense Amplifier (EASA) proposed in [10] and the Pre-Charge Sense Amplifier (PCSA) proposed in [83]. As shown in [2], EASA and PCSA provide area- and energy-efficient sensing circuits, while VISA and SPCSA provide more reliable sensing circuits at the cost of increased energy consumption and area footprint.



(a)



(b)

Figure 7.7: (a) The sample operation error rate trade-off with sample duration and VCMA-MTJ switching duration, and (b) The energy consumption trade-off with sample duration and VCMA-MTJ switching duration [5].

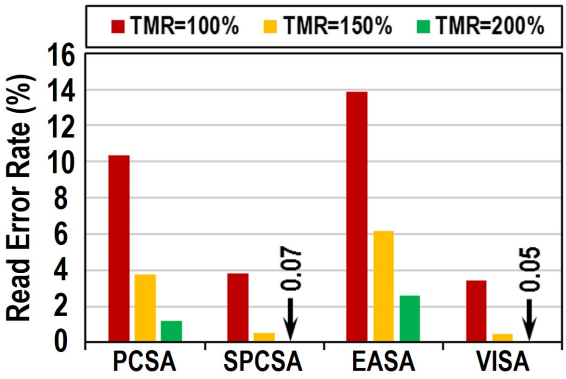


Figure 7.8: The read operation error rate trade-off with different TMR values for PCSA, EASA, SPCSA, and VISA [5].

7.3.4 Comparisons

In Table 7.2, we compare the performance of the developed adaptive acquisition framework with prior non-uniform ADC architectures. The proposed AIQ exhibits power dissipation of 0.32uW for 1-bit resolution, 1uW for 2-bit resolution, 2.33uW for 3-bit resolution, 5.02uW for 4-bit resolution, 10.37uW for 5-bit resolution, 21.08uW for 6-bit resolution, 42.48uW for 7-bit resolution, and 85.3uW for 8-bit resolution. Furthermore, our results indicate that the energy consumption per sample for 1-bit, 2-bit, 3-bit, 4-bit, 5-bit, 6-bit, 7-bit, and 8-bit quantization resolutions are 16.68fJ, 51.79fJ, 120.46fJ, 258.54fJ, 533.98fJ, 1.09pJ, 2.19pJ, and 4.39pJ, respectively.

Since our proposed design benefits from intermittent operation, which enables the proposed AIQ to turn off parts of the circuit that are not being utilized during the sampling process and turn them on whenever appropriate, its dynamic power dissipation is averaged from 1-bit to 8-bit resolutions for one sample. It should be noted that the amount of energy required for our proposed AIQ depends on the SR and QR values adapted for each frame. Thus, we report the power consumed averaged for different number of bits per frame. According to our results, the proposed AIQ incurs only 20.98μW power dissipation on average, while providing uniform digital output of 1 to 8 bits. Furthermore, our results indicate that the AIQ consumes ~ 1pJ energy per sample on average.

Table 7.2: Comparison with prior ADC designs utilizing Non-Uniform Sampling [5].

	Uniform Digital Output	Process (SupplyVoltage)	Adaptive		Power (Average Power)	Maximum Effective Bandwidth	Energy per Sample
			SR	QR (#Bits)			
Bellasi, et al. [63]	No	28nm (1.0V)	Yes ✓	No (4-bit)	7.5mW	2.4 GHz	2.9 pJ
Varshney, et al. [64]	No	45nm (1.2V)	Yes ✓	Yes ✓ (4-6 bit)	80μW-1.15mW (442μW)	120 MHz	3.68 pJ
Wu, et al. [65]	Yes ✓	65nm (1.0V)	Yes ✓	No (4-bit)	30mW	20 MHz	5 pJ
Naraghi, et al. [66]	No	90nm (1.0V)	Yes ✓	No (9-bit)	14μW	300 KHz	98 fJ
Kurchuk, et al. [67]	Yes ✓	65nm (1.2V)	Yes ✓	Yes ✓ (1-3 bit)	1.1mW-10mW (6.2mW)	2.4 GHz	36 fJ
AIQ (herein)	Yes ✓	22nm (1.0V)	Yes ✓	Yes ✓ (1-8 bit)	0.319μW-85.302μW (20.98μW)	500 MHz	1 pJ

As it can be observed in Table 7.2, our proposed AIQ design provides 421μW and 6.18mW power savings on average compared to other adaptive rate and resolution ADC designs proposed in [64] and [67], respectively, while offering a wider range of quantization resolution up to 8 bits. Additionally, our proposed AIQ design consumes ~ 1.34pJ less energy per sample on average compared to other state-of-the-art ADC designs proposed in [63, 64, 65, 66, 67]. Moreover, despite utilizing an adaptive clock for sampling operation, our proposed AIQ design utilizing VCMA-MTJ spin-based devices achieves a performance comparable with other state of the art CMOS-based architecture as shown in Table 7.2 in terms of average power dissipation and energy consumption per sample, while providing adaptive SR and QR.

Moreover, since the MTJ devices are considered as non-volatile memory cells, there is no need for an external FLASH memory or latch to store the data after each sampling operations. The sampled data will remain in the MTJ devices even if the power failure occurs. As a result, an extreme area reduction is achieved. For example, in the 8-bit resolution ADC, 256 comparators are used, and each comparator is connected to a latch for storing the sampled value. However, by utilizing MTJ devices, 256 latches can be eliminated from the circuit, resulting in a significant area reduction. Furthermore, since the MTJ devices can be fabricated on top of the baseline CMOS process, they

need not occupy extra area in lateral space, which further advances an area efficient design.

7.4 Conclusion

A novel adaptive framework for energy-aware acquisition of spectrally-sparse signals is proposed. The adaptive quantized Compressive Sensing (CS) techniques, beyond-Complementary Metal Oxide Semiconductor (CMOS) hardware architecture, and corresponding algorithms which utilize them have been designed concomitantly to minimize the overall cost of signal acquisition. First, a spin-based *Adaptive Intermittent Quantizer (AIQ)* is developed to facilitate the realization of the adaptive sampling technique. Second, a framework for smart and adaptive determination of the sampling rate and quantization resolution based on the instantaneous signal and hardware constraints is introduced. Simulation results indicate that an AIQ architecture using a spin-based quantizer incurs only 20.98μW power dissipation on average using 22nm technology for 1 to 8 bits uniform output. Furthermore, in order to provide 8-bit quantization resolution, 85.302μW maximum power dissipation is attained. Our results indicate that the proposed AIQ design provides up to 6.18mW power savings on average compared to other adaptive rate and resolution CMOS-based CS Analog to Digital Converter (ADC) designs. Additionally, the Mean Square Error (MSE) values achieved by the simulation results confirm efficient reconstruction of the signal based on the proposed approach.

CHAPTER 8: AQURATE: MRAM-BASED STOCHASTIC OSCILLATORS FOR ADAPTIVE QUANTIZATION RATE SAMPLING OF SPARSE SIGNALS[1]

Previous works on adaptive quantization rate and resolution ADCs have been implemented using Complementary Metal Oxide Semiconductor (CMOS) technology and considering a low-pass signal model [65, 97]. In this Chapter, we propose an spin-based Adaptive quantization rate (AQR) generator circuit that considers the signal dependent constraint as well as hardware limitations. The proposed AQR generator circuit utilized Magnetic Random Access Memory (MRAM)-based stochastic oscillator devices, which offer miniaturization and significant energy savings [7].

8.1 Proposed AQR Generator Circuit

To realize an effective hybrid emerging device and CMOS circuit, one useful approach can be to consider stochastic and deterministic attributes separately. For instance, Figure 8.1 depicts the proposed AQR generator circuit wherein a 2-terminal MTJ realizes stochastic behavior to provide the non-uniform clock generation capability.

The quantized Sparsity Rate Estimator (SRE) module shown in Figure 8.1 estimates the sparsity rate of the digital output bit-stream by estimating the sparse spectral components of the digital output using an iterative algorithm. Recently, rapid and optimized sparse component estimation method is proposed in [5]. In the approach proposed in [5], in order to minimize the computational complexity of the sparse component estimation, a sliding window approach is utilized and the algorithm operates only one iteration on each frame of the input by utilizing the previous estimate

[1] ©IEEE. Part of this chapter is reprinted, with permission, from [13]

as an initial value. This will result in gradual convergence of the sparse components to the actual values across iterations. These algorithms can be employed to find the sparsity rate of the signal. In most cases, sparsity rate of analog signals, which can be described as the number of non-zero elements in the sparse representation of the signal divided by the total number of elements, is between 5% to 15% in many applications including those targeted herein.

When the SRE module estimates the sparsity rate of the signal based on the digital output of the previous frame, it will then generate a voltage level according to that sparsity rate of the input analog signal. This voltage, referred to as V_{SR}, will be applied to the gate of the NMOS transistor shown in Figure 8.1 and results in an stochastic bit-stream generated by the MRAM-based stochastic oscillator device. The stochastic bit-stream output generated by the MRAM-based stochastic oscillator device will be forwarded to the D-Flip-Flop (D-FF) as shown in Figure 8.1 and the result of the 2-input NAND gate between the output of the D-FF and the actual clock of the circuit will generate the required quantization rate to be used for the following frame of the signal acquisition, referred to as Asynchronous Clock (A-Clk) in Figure 8.1. Additionally, the SRE module can also be used by the recovery algorithms to efficiently recover the sampled signal [5]. Additionally, the A-Clk will be forwarded to the sparse recovery algorithm to provide necessary information about the samples taken from the signal to assist with the signal reconstruction.

To obtain the relation between the output probability of the stochastic MRAM-based AQR generator and its input voltage, we have applied an input pulse that its amplitude starts from GND and is increased by 200mV every 100ns until it reaches V_{DD}. The output of the building block is sampled with a 1GHz clock frequency using a D-FF circuit, as shown in Figure 8.2.

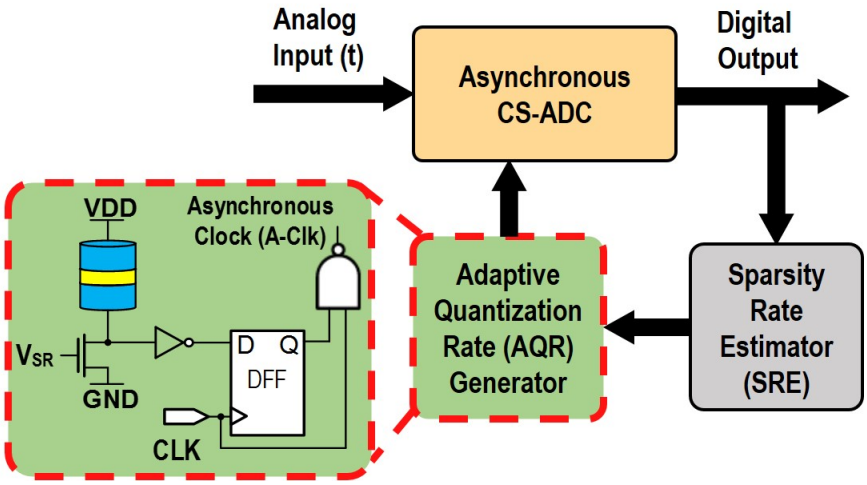


Figure 8.1: Integration of AQR generator circuit within the Compressive Sensing ADC (CS-ADC) system design. [13]

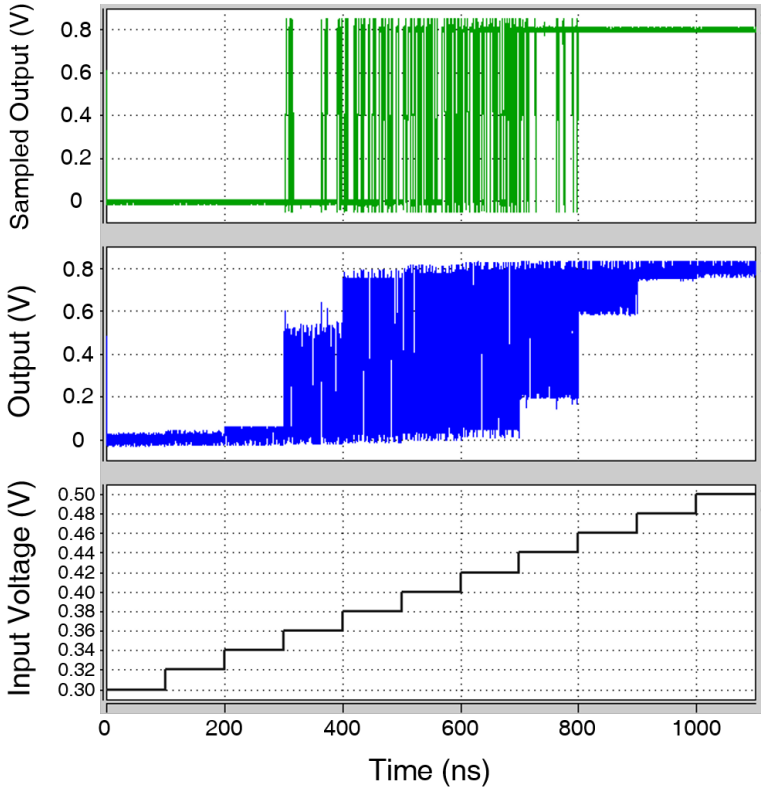


Figure 8.2: The sampled output of the stochastic MRAM-based building block for AQR generator for various input voltages. [13]

8.2 Simulation Results

In order to evaluate and validate the behavior and functionality of the proposed AQR generator circuit, SPICE and MATLAB simulations were performed. We have utilized the 14nm High Performance FinFET Predictive Technology Model (PTM) [222] as well as the MRAM-based stochastic oscillator device model and parameters represented in [7] to implement and evaluate the proposed AQR generator circuit.

According to our results, AQR provides significant power dissipation and area reductions compared to the state-of-the-art non-uniform clock generators listed in Table 8.1 [75, 97, 223, 224]. According to our simulation results, power dissipation of the proposed AQR generator circuit is 22.64μW on average. With respect to area utilization, our proposed AQR design requires only 23 FinFET transistors, which attains a significant reduction in the transistor count and complexity of the non-uniform clock generator circuit present in state-of-the-art designs [75, 97, 223, 224]. Thus, AQR avoids high transistor counts while making it unnecessary to use large LFSR circuits that contain numerous D-FFs as well as several logic gates and multiplexers. For a more equitable comparison in terms of area and power dissipation, we have derived (8.1) and (8.2) considering General Scaling method [225] to normalize the power dissipation and area of the designs listed in Table 8.1. Based on the General Scaling method, voltage and area scale at different rate of U and S, respectively. Thus, the power dissipation is scaled with respect to $1/U^2$ and area per device is scaled according to $1/S^2$ [225]:

$$Power_{norm} = \frac{Power_x}{Power_{AQR}} \times (\frac{1}{U})^2 = \frac{Power_x}{Power_{AQR}} \times (\frac{0.8V}{V_{nominal}})^2, \tag{8.1}$$

$$Area_{norm} = \frac{Area_x}{Area_{AQR}} \times (\frac{1}{S})^2 = \frac{Area_x}{Area_{AQR}} \times (\frac{14nm}{Technology})^2, \tag{8.2}$$

where, $V_{nominal}$ is the nominal voltage of the technology model, $Technology$ refers to the technology node in nanometers, and subscript x refers to the design that we want to scale its power dissipation and area according to the technology models. According to (8.1) and (8.2), AQR provides power dissipation reduction up to one-order-of-magnitude compared to the state-of-the-art nonuniform clock generators as listed in Table 8.1. Additionally, AQR offers up to one-order-of-magnitude area reduction compared to the designs provided in Table 8.1 using the scaling comparison trends accepted in the literature.

Table 8.1: Comparison with recently proposed non-uniform clock generator designs. [13]

Design	**Technology ($V_{nominal}$)**	**Power**$_{norm}$	**Area**$_{norm}$
[75]	65nm (1.1V)	$\sim 1\times$	$\sim 1\times$
[223]	65nm (1.1V)	$\sim 2\times$	$\sim 21\times$
[224]	90nm (1.2V)	$\sim 2\times$	$\sim 51\times$
[97]	28nm (1.0V)	$\sim 18\times$	N/A
This Work	14nm (0.8V)	$1\times$	$1\times$

As described in Section 8.1, sparsity rate of analog signals is usually within the range of $5\% - 15\%$. Moreover, we have embedded our proposed AQR generator within CS recovery algorithms called Orthogonal Matching Pursuit (OMP) and Compressive Sampling Matching Pursuit (CoSaMP) [14] in order to evaluate the architectural simulation results and in order to recover the signal from the samples taken using the AQR generator. According to the results, the mean normalized errors of the reconstruction of the signals with 5%, 10%, and 15% sparsity rates using OMP are 0.0504, 0.0446, and 0.0252, respectively. Moreoever, the mean normalized errors of the reconstruction of the signals with 5%, 10%, and 15% sparsity rates using CoSaMP are 0.0487, 0.0304, and 0.0245, respectively. Figure 8.3 depicts an example signal with 10% sparsity rate and its reconstructed signal using CoSaMP algorithm [14] with MSE=0.0304.

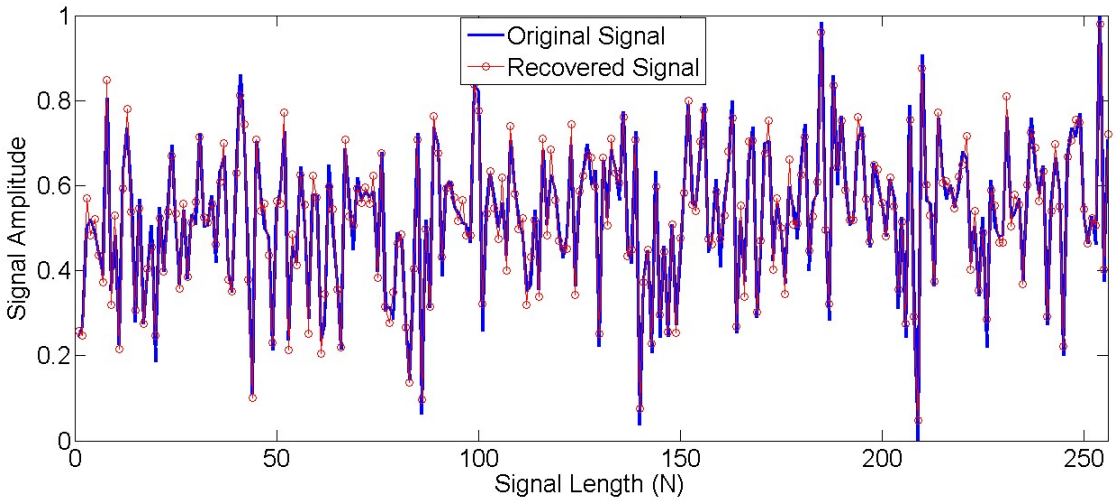


Figure 8.3: Recovery of an sparse signal with sparsity rate of 10% using CoSAMP [14] and samples taken by AQR generator output (MSE=0.0304). [13]

8.3 Conclusions

Recently, the promising aspects of compressive sensing have inspired new circuit-level approaches for their efficient realization within the literature. However, most of these recent advances involving novel sampling techniques have been proposed without considering hardware and signal constraints. Additionally, traditional hardware designs for generating non-uniform sampling clock incur large area overhead and power dissipation. Herein, we propose a novel non-uniform clock generator called *Adaptive Quantization Rate (AQR)* generator using MRAM-based stochastic oscillator devices. Our proposed AQR generator provides $\sim$ 25-fold reduction in area, on average, while offering $\sim$ 6-fold reduced power dissipation, on average, compared to the state-of-the-art non-uniform clock generators.

CHAPTER 9: SLIM-ADC: SPIN-BASED LOGIC-IN-MEMORY ANALOG TO DIGITAL CONVERTER LEVERAGING SHE-ENABLED DOMAIN WALL MOTION DEVICES[1]

Challenges incurred by conventional Von-Neumann computing architectures that are mainly due to interconnection and busing demands [88], increase static energy consumption, causes large access latency, and limited scalability. Furthermore, there is an increasing demand for energy and area efficient Analog to Digital Converters (ADCs) as the need for integrating the signal acquisition and processing as well as rapid parallel data conversion in sensor nodes has increased [90, 91, 92]. Moreover, increased static energy consumption and decreased reliability caused by high process variation have become a major challenge in scaled technology nodes [95]. Thus, we devise a framework for efficient acquisition of analog signals utilizing emerging spin-based devices. In this Chapter, we propose a spin-based intermittent quantizer with logic computation capabilities. The proposed architecture, called Spin-based Logic-In-Memory ADC (SLIM-ADC), utilizes Spin-Hall Effect driven Domain Wall Motion (SHE-DWM) devices to provide fast quantization of analog signals in a novel energy-efficient fashion as well as realizing intrinsic logic operations. By leveraging non-volatility, SLIM-ADC can reduce energy consumption via instant off/on operation without the use of backup storage.

9.1 Proposed Spin-based Logic-In-Memory Analog to Digital Converter (SLIM-ADC)

Our proposed SLIM-ADC design leveraging SHE-DWM devices is shown in Figure 9.1. Our proposed dual-mode device is capable of implementing ADC operations as well as logical operations.

[1] ©IEEE. Part of this chapter is reprinted, with permission, from [6]

The MTJ0, MTJ1, and MTJ2 each can represent a quantization level referred to as L0, L1, and L2 as shown in Figure 9.1, in order to quantize the analog input signal. Additionally, each of these MTJ devices can represent a different function such as 3-input OR gate, 3-input Majority Gate (MG), and 3-input AND gate, shown in Figure 9.1 as F0, F1, and F2, respectively. The proposed write circuit used for the SLIM-ADC device is illustrated in Figure 9.2. Since there are notches [226] in the magnetic domain, in order to move the DW, the input requires an appropriate current magnitude and direction. Furthermore, the signals used to activate the read and write operations of the SLIM-ADC device are listed in Table 9.1. One of the main contributions of the proposed SLIM-ADC devices is their tolerance to intermittency which enables these devices to save energy by going to standby mode when there is no ADC or logic operation requested. Moreover, the instant-on feature of these devices allows them to resume normal operation without loss of data stored in the MTJs.

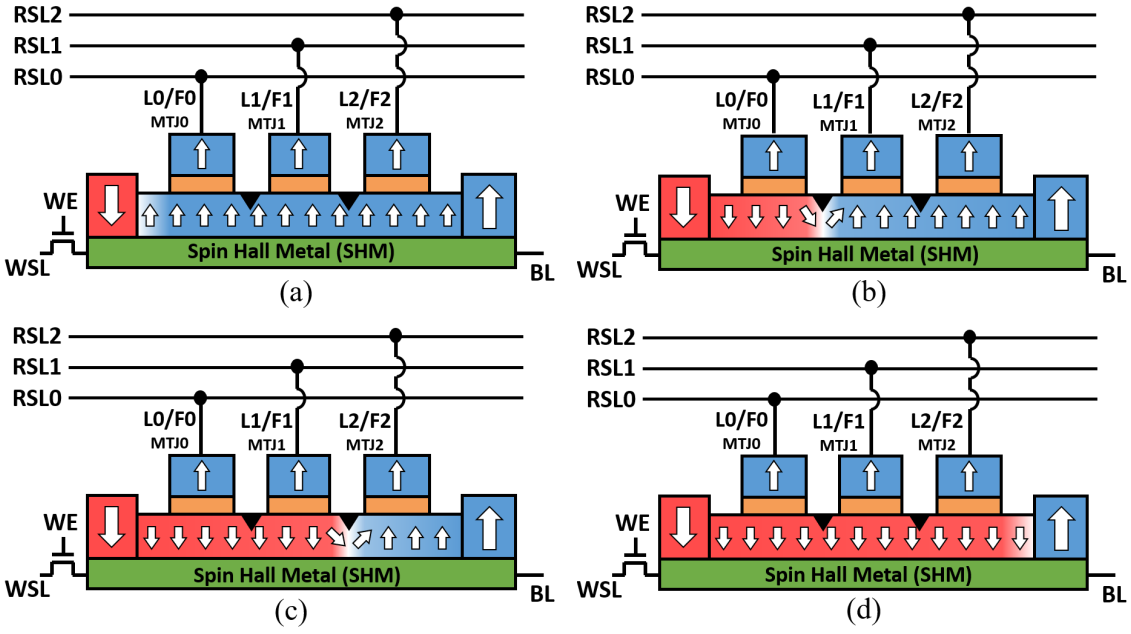


Figure 9.1: The proposed SLIM-ADC device in (a) 000, (b) 100, (c) 110, and (d) 111 modes. [6]

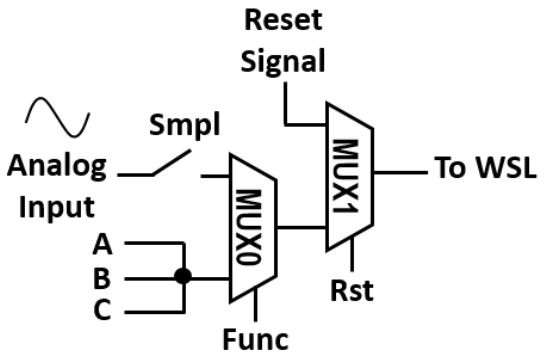


Figure 9.2: The proposed write circuit for the SLIM-ADC device. [6]

Table 9.1: The Signaling of the SLIM-ADC device for read and write operations. [6]

Operation	WSL	RSL	WE	RE	BL
Reset	From write circuit	Hi-Z	1	0	1
Write	From write circuit	Hi-Z	1	0	0
Read	0	From SA	0	1	0

9.1.1 ADC Mode

As depicted in Figure 9.2, when the **`Func`** signal for MUX0 is set to 1, the device will be in ADC mode. ADC mode has three simple steps: 1) reset, 2) conversion, and 3) read-out. During the reset state, the **`Rst`** signal of MUX1 will be set to 1 and the DW will be pushed all the way to the beginning of the magnetic domain (leftmost location) to reset and prepare the device for the conversion state. As shown in Figure 9.2, during the conversion state, the **`Smpl`** signal will be enabled for a short period to sample the analog input signal and the **`Rst`** signal of MUX1 will be set to 0 to allow the sampled analog input signal to move the DW, depending on the input signal's magnitude. During the read-out state, using the read operation and SAs presented in [10], we can read the values stored in all 3 MTJs, and based on their resistance values, find the digital output encoded using 3 levels to realize a 2-bit ADC operation as shown in Figure 9.1, where L0, L1, and L2 refer to Level 0, Level 1, and Level 2, respectively. During the read operation one of the 000, 100, 110, or 111 states will be achieved that can be encoded into two bits as shown in Table 9.2.

Table 9.2: The SLIM-ADC's bit encoding for ADC operation. [6]

MTJ0 / L0	MTJ1 / L1	MTJ2 / L2	Encoded Bits
0	0	0	0 0
1	0	0	0 1
1	1	0	1 0
1	1	1	1 1

9.1.2 Logic-in-Memory Mode

Furthermore, when the **`Func`** signal for MUX0 is set to 0, the device will be in logic operation mode, as illustrated in Figure 9.2. The logic operation has also three steps: 1) reset, 2) computation, and 3) read-out. During the reset state, which is the same as ADC mode, the **`Rst`** signal of MUX1 will be set to 1 and the DW will be pushed all the way to the beginning of the magnetic domain (leftmost location) to reset and prepare the device for the computation state. As depicted in Figure 9.2, after the reset state, the input currents of A, B, and C will be applied during the computation state for logic operation. Based on the current magnitude applied through the inputs A, B, and C, the DW will move. Finally, in the read-out state, the output of each MTJ will provide a different function as shown in Figure 9.1. The proposed device is designed so that F0 provides a 3-input OR gate, OR(A,B,C), F1 provide a 3-input Majority gate, MG(A,B,C), and F2 provides a 3-input AND gate, AND(A,B,C), as listed in Table 9.3. Additionally, one of the inputs can be used as a bias to achieve 2-input OR and AND gates. Herein, if we consider input C as the bias and connect it to logic 0, then MTJ0 will provide OR(A,B) as F0, and MTJ1 will provide AND(A,B) as F1, as listed in Table 9.4.

9.1.3 Sense Amplifier (SA) Circuit for the Read Operation

The Sense Amplifier (SA) circuit shown in Figure 9.3 is used to read the data of the three MTJ devices, namely MTJ0, MTJ1, and MTJ2, simultaneously. The read operation is comprised of two

steps: pre-charge and sensing. During the pre-charge step, the RE signal is connected to the logic 0, which turns on $MP0$ and $MP3$ transistors and causes Transmission Gates (TGs), $TG0$, $TG1$, and $TG2$, to turn off and as a result the output nodes of each SA, referred to as OUT_i and $\overline{OUT_i}$ in Figure 9.3, are pre-charged to VDD. During the sensing step, the RE signal is connected to logic 1, which causes $TG0$, $TG1$, and $TG2$ to turn on and as a result the output nodes of each SA, referred to as OUT_i and $\overline{OUT_i}$ in Figure 9.3, start to discharge to the ground.

Table 9.3: The Truth Table for the 3-input Logic Operations. [6]

A	B	C	F0=OR(A,B,C)	F1=MG(A,B,C)	F2=AND(A,B,C)
0	0	0	0	0	0
0	0	1	1	0	0
0	1	0	1	0	0
0	1	1	1	1	0
1	0	0	1	0	0
1	0	1	1	1	0
1	1	0	1	1	0
1	1	1	1	1	1

Table 9.4: The Truth Table for the 2-input Logic Operations. [6]

A	B	C/Bias	F0=OR(A,B)	F1=AND(A,B)
0	0	0	0	0
0	1	0	1	0
1	0	0	1	0
1	1	0	1	1

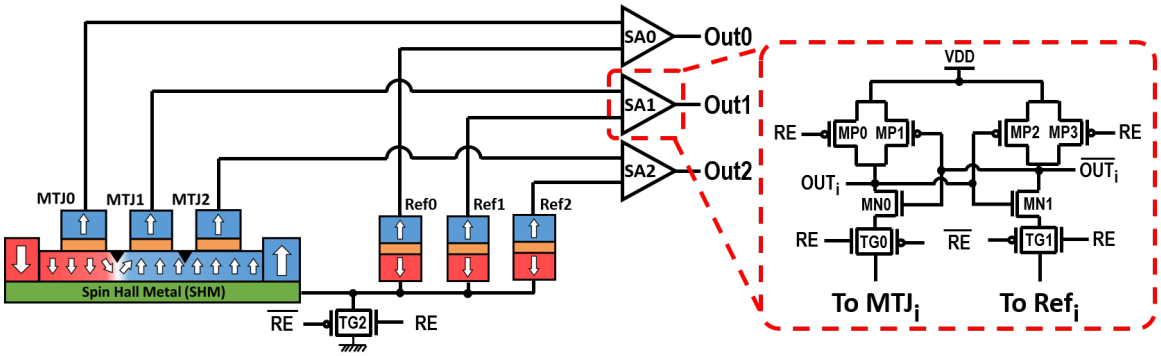


Figure 9.3: The proposed SA circuit for the SLIM-ADC device ($i = \{0, 1, 2\}$). [6]

According to the difference in the resistance states of the MTJ_i and Ref_i in each SA, one of the two output nodes, OUT_i and $\overline{OUT_i}$, discharges more rapidly, leading the other output to charge to VDD [10]. All of the reference cells, Ref_0, Ref_1, and Ref_2, share the same dimensions and resistance values. The dimensions are set so that the resistance values of the reference cells hold a value between the P and AP states of the MTJs used with the SHE-DW. All of the reference cells, Ref_0, Ref_1, and Ref_2, share the same dimensions and resistance values. The dimensions are set so that the resistance values of the reference cells hold a value between the P and AP states of the MTJs used with the SHE-DW in order to achieve a sufficient sensing margin during the read operation.

9.2 Simulation Framework, Results, and Analysis

In order to accurately simulate the behavior of the proposed SLIM-ADC design, we have extracted the values used in [18], [19], and [20]. The Domain Wall Simulator presented in [19] are used along with SPICE simulation with the 22nm Predictive Technology Model (PTM) [202], in order to analyze the behavior of the SHE-DWM devices proposed herein utilizing the parameters listed in Table 9.5.

The modeling of the SHE-DWM can be realized through modifying the Landau-Lifshitz-Gilbert (LLG) equations as shown below [20]:

$$\frac{\mathrm{d}\vec{m}}{\mathrm{d}t} = -\gamma\vec{m} \times \vec{H}_{eff} + \alpha\vec{m} \times \frac{\mathrm{d}\vec{m}}{\mathrm{d}t} + \vec{\tau}_{stt} + \vec{\tau}_{sot}, \tag{9.1}$$

where, $\vec{m}$ is the magnetization vector of the DW's free-layer $\{m_x, m_y, m_z\}$, γ is the gyromagnetic ratio, α is the Gilbert damping factor, $\vec{H}_{eff}$ is the effective magnetic field vector derived from the energy density of the system, τ_{stt} is the Spin-Transfer Torque (STT) factor, and τ_{sot} is the

Spin-Orbit Torque (SOT) factor. Additionally, H_{eff} can be described as [20]:

$$\vec{H}_{eff} = \frac{-1}{\mu_0 M_s} \times \frac{\delta \epsilon_{DM}}{\delta \vec{m}}, \tag{9.2}$$

$$\epsilon_{DM} = -D[m_z \nabla \cdot \vec{m} - (\vec{m} \cdot \nabla) m_z] \quad \text{if } t_{DW} << L_{DW} \,\&\, W_{DW}, \tag{9.3}$$

where, μ_0 is the vacuum permeability, M_s is the Saturation Magnetization, ϵ_{DM} is the Dzyaloshinskii Moriya Interaction (DMI) energy density, and D is the DMI intensity parameter.

Table 9.5: Circuit parameters and constants with their corresponding values for the SHE-DWM device model. The values are taken from [18], [19], and [20]. [6]

Parameter/Constant	**Description**	**Default Value**
M_s	Saturation Magnetization	$6.8 \times 10^5 A/m$
K_u	Initial Interfacial PMA energy	$3.5 \times 10^5 J/m^3$
α	Gilbert Damping Factor	0.03
t_{ox}	Oxide-layer Thickness	$1nm$
A_{ex}	Exchange Stiffness	$1.1 \times 10^{-11} J/m$
ρ	Resistivity of Magnet	$170 \Omega nm$
RA	MTJ Resistance Area Product	$2.38 \Omega \mu m^2$
$(L \times W)_{MTJ}$	MTJ Dimensions	$20 \times 20 nm^2$
$(L \times W \times t)_{DW}$	DW Nano-wire Dimensions	$100 \times 20 \times 2.8 nm^3$
$(L \times W \times t)_{SHM}$	SHM dimensions	$120 \times 20 \times 2.8 nm^3$
MTJ_i	MTJ Resistance in [P, AP] States	$[3.2, 6.4] K\Omega$
Ref_i	Reference Cell Resistance	$4.8 K\Omega$
ρ_{SHM}	Resistivity of SHM (W)	$200 \mu \Omega cm^2$
θ_{SHM}	Initial Spin-Hall angle	0.3
TMR_{AP}	Tunnel Magneto Resistance	100%
P	Spin Polarization	0.6
λ_{sf}	Spin Flip Length	$1.5nm$

As a result, the effective DMI field can be described as below [20]:

$$\vec{H}_{DM} = \frac{-2D}{\mu_0 M_s} \times [\frac{\partial m_z}{\partial x}\vec{u}_x + \frac{\partial m_z}{\partial y}\vec{u}_y - (\frac{\partial m_x}{\partial x}\vec{u}_x + \frac{\partial m_y}{\partial y}\vec{u}_y)\vec{u}_z]. \tag{9.4}$$

Furthermore, assuming that $\hbar$ is the reduced Planck constant, P is the STT polarization factor, j_a is the driving current density, e is the elementary electron charge, μ_B is the Bohr magneton, θ is the initial spin-Hall angle, η is the non-adiabatic Rashba term, ξ is the dimensionless non-adiabatic parameter, α_R is the Rashba parameter, and $\vec{H}_R$ is the effective Rashba field, the STT and SOT factors can be described as below [20]:

$$\vec{\tau}_{stt} = (j_a \frac{\mu_B P}{e M_s}) \times (\vec{u}_x \cdot \nabla)\vec{m} - (j_a \frac{\mu_B P}{e M_s}) \times \xi \vec{m} \times (\vec{u}_x \cdot \nabla)\vec{m}, \tag{9.5}$$

$$\vec{\tau}_{sot} = -\gamma \vec{m} \times \vec{H}_R + \eta\gamma\xi \vec{m} \times (\vec{m} \times \vec{H}_R) - \gamma \vec{m} \times (\vec{m} \times H_{SH}\vec{u}_y), \tag{9.6}$$

$$\vec{H}_R = \frac{\alpha_R P}{\mu_0 \mu_B M_s}(\vec{u}_z \times \vec{j}_a) = \frac{\alpha_R P j_a}{\mu_0 \mu_B M_s}\vec{u}_y, \tag{9.7}$$

$$H_{SH} = \frac{\hbar \theta_{SH} j_a}{\mu_0 2 e M_s t_{DW}} = \frac{\mu_B \theta_{SH} j_a}{\gamma 2 e M_s t_{DW}}. \tag{9.8}$$

Authors in [19] have utilized the modeling approach described in [20] to implement a standalone one-dimensional DWM simulator. This model takes STT, SOT, and DMI fields into account [19]. According to our results, if $j_a \simeq 0.75 \times 10^{12}$A/m^2 is applied, the DW will move to the first notch within 1ns, if $j_a \simeq 1.44 \times 10^{12}$A/m^2 is applied, the DW will move to the second notch within 1ns,

and if $j_a \simeq 2.08 \times 10^{12}$A/m^2 is applied, the DW will move all the way to the end of the magnetic domain within 1ns. Figure 9.4 depicts sample simulation waveforms for the proposed SLIM-ADC device. According to our results, the energy consumption of each ADC or logic operation on average is ~ 201.48fJ, which on average includes ~ 117.94fJ for the reset operation, ~ 79.70fJ for the sampling/computing operation, and ~ 3.84fJ for the read operation. Considering 0.6ns for the reset operation, 1ns for the sample/compute operation, and 0.4ns for the read operation, the overall operation time is 2ns, which means the proposed SLIM-ADC device can perform ADC or logic operations with 500MHz frequency.

Faster ADC and logic operations can be achieved by increasing the input current corresponding to reset and sample/compute operations, however this will elevate the power dissipation. In order to increase the speed of the proposed SLIM-ADC device to be able to perform ADC or logic operations with 1GHz frequency, the reset operation is required to be done in 0.3ns, sample/compute operation is required to be done in 0.5ns, and read operation requires to be done in 0.2ns. Furthermore, according to our results, if $j_a \simeq 1.57 \times 10^{12}$A/m^2 is applied, the DW will move to the first notch within 0.5ns, if $j_a \simeq 2.85 \times 10^{12}$A/m^2 is applied, the DW will move to the second notch within 0.5ns, and if $j_a \simeq 4.1 \times 10^{12}$A/m^2 is applied, the DW will move all the way to the end of the magnetic domain within 0.5ns. In this case, the energy consumption of each ADC or logic operation is equal to ~ 196.65fJ on average, which includes ~ 117.1fJ for the reset operation, ~ 79.52fJ for the sampling/computing operation, and ~ 0.03fJ for the read operation.

In Table 9.6, we compare the performance of the developed SLIM-ADC device with other low-resolution ADC architectures that utilize CMOS or emerging spin-based technologies. It can be observed that the proposed SLIM-ADC device provides fast and energy-efficient analog to digital conversion compared to state of the art ADC designs. In particular, the proposed SLIM-ADC operating in 1GHz frequency, in most cases outperform other designs listed in Table 9.6 in terms of power dissipation by ~ 5.3mW on average.

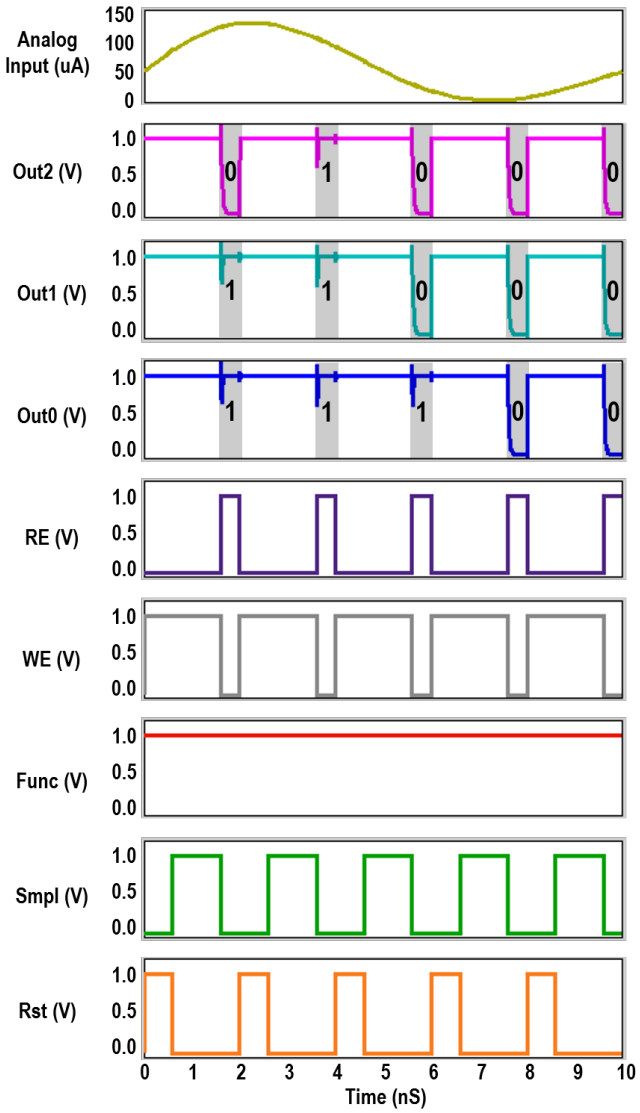


Figure 9.4: Simulation waveforms for the proposed SLIM-ADC device. [6]

Moreover, the proposed SLIM-ADC design improves the power dissipation of the sampling in 500MHz frequency by $\sim$ 5.6mW on average compared to most of the designs provided in Table 9.6. Additionally, a 1-bit MG-based Full-Adder (MG-FA) circuit is implemented utilizing the proposed SLIM-ADC devices using the circuit shown in Figure 9.5. According to our results, the power dissipation of the proposed SLIM-ADC-based MG-FA in $1GHz$ frequency is equal to $589.95\mu W$ and the result of the addition will be ready within $2ns$. Furthermore, the power dissi-

pation of the proposed SLIM-ADC-based MG-FA in $500MHz$ frequency is equal to $302.22\mu W$ and the result of the addition will be ready within $4ns$. The output waveform of the 1-bit MG-FA circuit using the proposed SLIM-ADC devices is shown in Figure 9.6.

Table 9.6: Comparison with prior low-resolution ADC designs. [6] (N/A: Data Not Available in the referenced manuscript.)

Design	Technology	Resolution in Bits	Power	Maximum Bandwidth Frequency	Energy per Sample
[65]	CMOS	4-bit	30mW	20MHz	5pJ
[227]	CMOS	3-bit	3.1mW	2GHz	0.27pJ
[228]	SHE-MTJ	3-bit	1.9mW	500MHz	0.48pJ
[82]	DWM	5-bit	3.4mW	500MHz	3.5pJ
[111]	DWM	3-bit	0.22mW	200MHz	N/A
			1.44mW	500MHz	N/A
			6.56mW	1GHz	N/A
[112]	Racetrack DWM	8-bit	96.5μW	20MHz	21fJ
SLIM-ADC	**SHE-DWM**	**2-bit**	**285.87μW**	**500MHz**	**79.71fJ**
			549.51μW	**1GHz**	**79.52fJ**

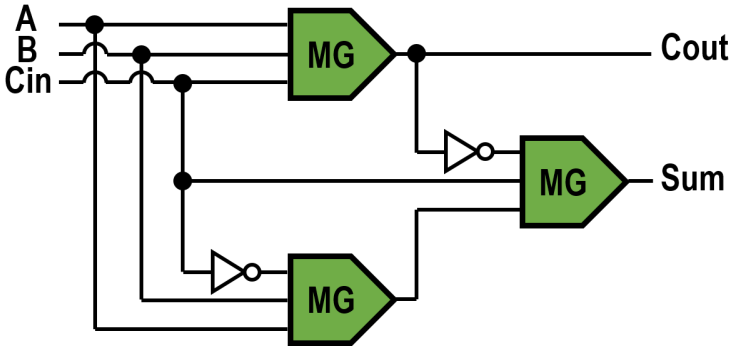


Figure 9.5: Proposed 1-bit MG-FA circuit implemented utilizing the SLIM-ADC devices. [6]

Table 9.7 compares the performance of the developed SLIM-ADC-based MG-FA with other FA designs that utilize CMOS or emerging spin-based technologies. It can be observed that the proposed SLIM-ADC-based MG-FA operating in 1GHz frequency outperforms the other FA designs listed in Table 9.7 in terms of power dissipation by 2.7-fold on average. Additionally, the proposed SLIM-ADC MG-FA offers faster FA operation by 3.2-fold on average compared to other

FA designs listed in Table 9.7. Furthermore, the proposed SLIM-ADC MG-FA offers $\sim$ 2-fold and $\sim$ 3.8-fold reduced power dissipation on average in 1GHz and 500MHz operating speeds, respectively, and provides $\sim$ 2.3-fold and $\sim$ 1.13-fold delay improvement on average in 1GHz and 500MHz operating speeds, respectively, compared to other emerging spin-based FA designs listed in Table 9.7.

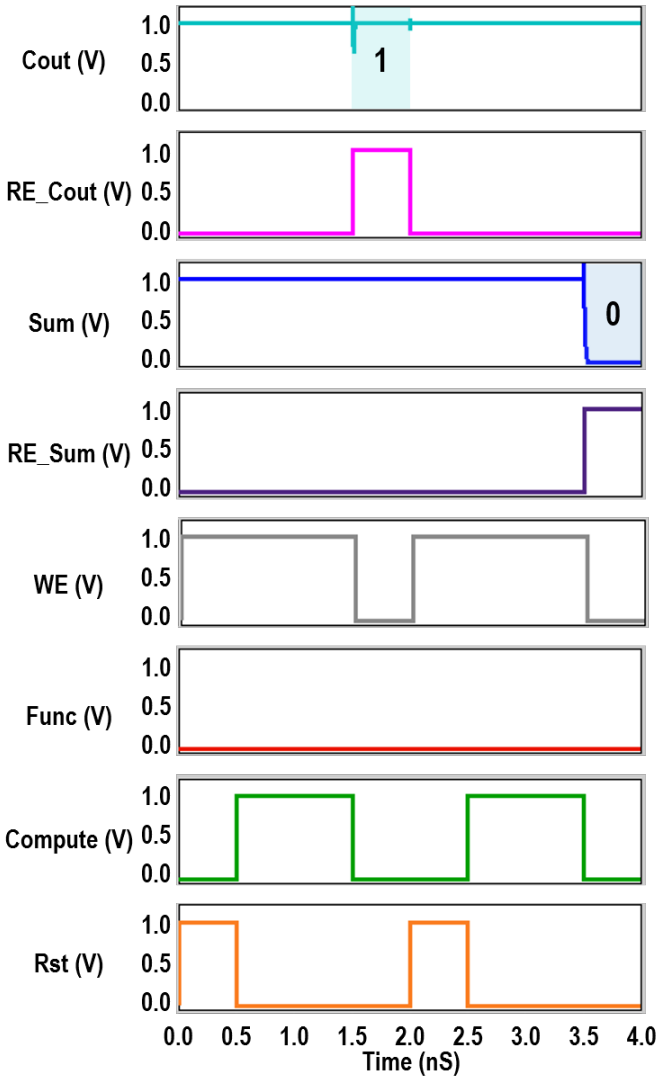


Figure 9.6: Simulation waveforms for the proposed 1-bit MG-FA circuit implemented utilizing the SLIM-ADC devices with inputs A=1, B=0, and Cin=1. [6]

Table 9.7: Comparison with prior Full-Adder designs. [6] (*The values are taken from [21].)

Design	Technology	Power	Delay	Energy per bit
[229]*	CMOS	2mW	2.2ns	4pJ
[229]*	STT-MTJ	2.1mW	10.2ns	4.2pJ
[21]*	SHE-MTJ	0.71mW	7ns	4.3pJ
[31]	DWM	1.364mW	2.54ns	1.4pJ
[230]	STT-MTJ	0.315mW	2.1ns	6.3pJ
[231]	Racetrack DWM	0.432mW	3.03ns	1.3pJ
SLIM-ADC	**SHE-DWM**	**589.95μW**	**2ns (1GHz)**	**0.6pJ**
		302.22μW	**4ns (500MHz)**	**0.6pJ**

Previous results indicate that for a conventional 2-bit CMOS ADC requires 49 transistors [232] and 1-bit CMOS FA requires 42 transistors [229], while our proposed SLIM-ADC can perform 2-bit ADC and 1-bit FA operation with 90 total transistors and 6 total MTJs. Thus, the device count for implementing a single ADC and FA circuit is comparable with the conventional CMOS-based approaches. However, compared to the conventional Logic In Memory (LIM) and ADC approaches, the area of the proposed SLIM-ADC is reduced, since an array of SAs is shared among the entire column of SLIM-ADC devices and an array of write circuits is shared among the entire row of SLIM-ADC devices. Hence, there is no need for a distinct SA and write circuit per device. Furthermore, the proposed SLIM-ADC device is capable of both logic and ADC operations while conventional approaches are only capable of performing one operation, either logic or ADC. Additionally, according to our results, the proposed SLIM-ADC consumes $\sim 0.2\mu$W leakage power which is negligible compared to CMOS designs which is around $\sim$ 1nW [229].

9.3 Conclusion

Herein, a novel ADC framework for energy-aware acquisition of analog signals with Logic-in-Memory capabilities is devised. Spin-Hall Effect driven Domain Wall Motion (SHE-DWM) devices are utilized to realize the proposed framework called *Spin-based Logic-In-Memory ADC*

(SLIM-ADC). Our simulation results indicate that the proposed SLIM-ADC offers $\sim$ 200fJ energy consumption on average for each analog conversion or logic operation with up to $1GHz$ speed. Furthermore, our results indicate that the proposed SLIM-ADC outperforms other state of the art spin-based ADC designs by offering $\sim$ 5.45mW improved power dissipation on average. Additionally, a Majority Gate (MG)-based Full-Adder (MG-FA) is implemented using the proposed SLIM-ADC. Our results show that the proposed MG-FA offers $\sim$ 2.9-fold reduced power dissipation on average and $\sim$ 1.7-fold reduced delay on average compared to the state of the art Full-Adder designs reported herein.

CHAPTER 10: MRAM-BASED STOCHASTIC OSCILLATORS FOR ADAPTIVE NON-UNIFORM SAMPLING OF SPARSE SIGNALS IN IOT APPLICATIONS[1]

A novel circuit-algorithm solution called Adaptive Sampling of Sparse IoT signals via STochastic-oscillators (ASSIST) is devised in this Chapter. ASSIST utilizes non-uniform compressive sensing algorithms as well as spin-based hardware circuit to improve energy-efficiency and performance of sampling and reconstruction operations within IoT applications. The proposed ASSIST approach utilizes Spin-based Stochastic Oscillator circuit to generate the CS measurement and then uses Spin Orbit Torque Magnetic Random Access Memory (SOT-MRAM) based resistive devices to store the CS measurement matrix elements.

10.1 Proposed Adaptive Sampling of Sparse IoT signals via STochastic-oscillators (ASSIST)

The proposed MRAM-based stochastic bitstream generator circuit is depicted in Figure 10.1(a), wherein a 2-terminal low energy-barrier thermally unstable MTJ is utilized. As shown in Figure 10.1(a), the output of the MSO is connected to a D-Flip-Flip (D-FF) which is controlled by a Power-Gated Clock (PG-CLK). This will provide control over the number of stochastic outputs provided by the MSO. In other words, by setting the duration of PG-CLK to run for M clock cycles, we would have a stochastic bitstream output, V_M, with the length of M bits, as shown in Figure 10.1(a). Additionally, having control over V_N enables us to adaptively adjust the number of '1's that appear in the output bitstream, V_M.

As shown in Figure 10.1(c), we have utilized a complementary SHE-MRAM array to store the

[1] ©IEEE. Part of this chapter is reprinted, with permission, from [8]

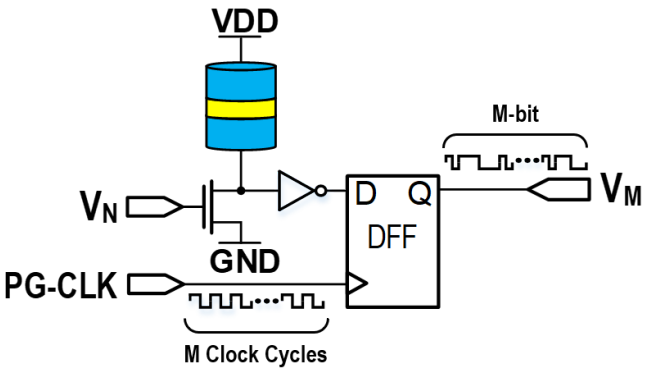
elements of the measurement matrix and for each column of the measurement matrix we have used an MRAM-based stochastic bitstream generator. Thus, in order to adaptively change the number of rows in the measurement matrix to account for increased sparsity rate, we can adjust V_M accordingly to increase the number of measurements. Furthermore, in order to increase accuracy of the signal recovery, we can increase V_N of the MRAM-based stochastic bitstream generators located in the columns corresponding to the RoI to maintain more '1's in the measurement matrix. It is worth noting that in order to use the MRAM-based stochastic bitstream generator output to write into the SHE-MRAM bit-cells, the PG-CLK clock cycle should be long enough for the write current to flow through the HM of the SHE-MTJs.

As mentioned earlier, we utilize the non-volatile complementary SHE-MRAM array, which will result in a wide read margin and increases reliability of the read operation [4]. Additionally, using a non-volatile complementary SHE-MRAM array enables a clockless read operation that is rapid, reliable, and energy-efficient. In order to use the MSO to write into the SHE-MRAM bit-cells, we utilize the circuit shown in Figure 10.1(b). Every column of the SHE-MRAM array shown in Figure 10.1(c) is populated using a separate MSO shown in Figure 10.1(a).

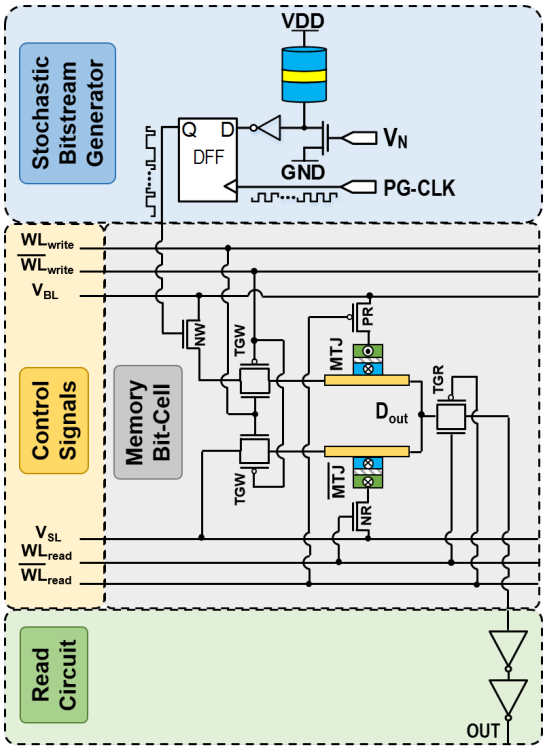
In order to write into each memory cell, **WWL** should be asserted to enable the write Transmission Gates (TGs), **TGW**. Then by setting Bit Line, **BL**, and Source Line, **SL**, we can write complementary data values in **MTJ** and $\overline{\textbf{MTJ}}$. Additionally, in order to use the MSO to write into the SHE-MTJ devices, the output of the D-FF is connected to the write NMOS transistor, **NW**. Thus, if the output of the D-FF is '1', then **NW** is turned on and will result in a current passing through the SHE-MTJs. On the other hand, if the output of the D-FF is '0', then **NW** will not turn on and the contents of the SHE-MTJs will remain untouched.

To read the data stored in the SHE-MTJs, **RWL** is asserted, which turns on the read TG, **TGR**. Additionally, the read transistors, **PR** and **NR**, are enabled. Thus, by applying **VDD** at **BL** and

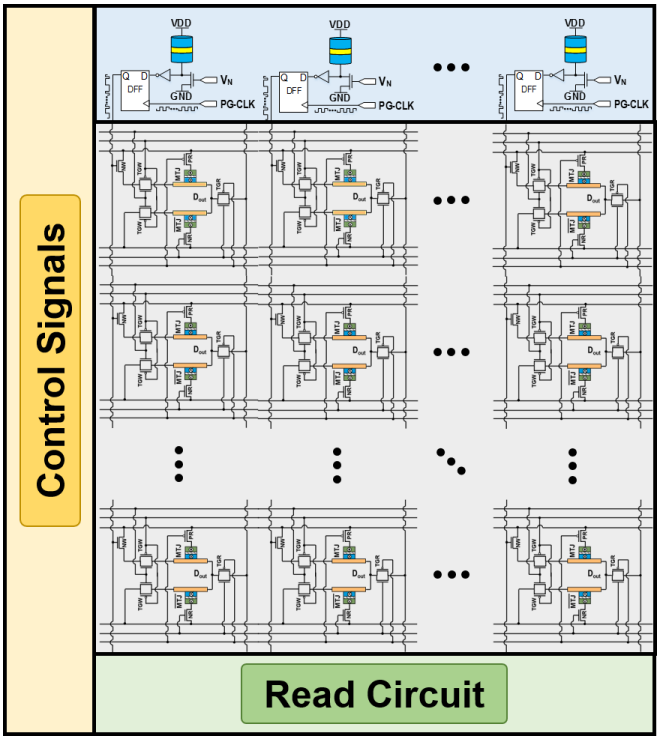
GND at **SL**, a read path from **VDD** to **GND** is formed. This will lead to a voltage divider circuit and by connecting the node between the complementary SHE-MTJs, $\mathbf{D}_{out}$, to two inverter logic gates, the output voltage will be amplified and presented at the output node, **OUT**.



(a)



(b)



(c)

Figure 10.1: The proposed ASSIST approach, where (a) depicts the stochastic bitstream generator circuit, (b) shows a complementary MTJ memory bit-cell connected to the stochastic bitstream generator, and (c) illustrates the architecture view. [8]

Table 10.1: Parameters of the 3-terminal SHE-MTJ device. [8]

Parameter	Description	Value
MTJ_{Area}	$l_{MTJ} \times w_{MTJ} \times \pi/4$	$60nm \times 30nm \times \pi/4$
HM_{Volume}	$l_{HM} \times w_{HM} \times t_{HM}$	$100nm \times 60nm \times 3nm$
t_f	Free Layer thickness	1.3 nm
RA	MTJ resistance-area product	$9\ \Omega \cdot \mu m^2$
T	Temperature	358 K
α	Gilbert Damping factor	0.007
P	Spin Polarization	0.52
θ_{SHE}	Spin Hall Angle	0.4
ρ_{HM}	HM Resistivity	$200\mu\Omega.cm$
λ_{sf}	Spin Flip Length	$1.5nm$

10.2 Simulation Results

In order to evaluate and validate the behavior and functionality of the proposed ASSIST approach, SPICE and MATLAB simulations were performed. We have utilized the 14nm HP-FinFET Predictive Technology Model (PTM) library as well as the MSO device model and parameters represented in [7] along with other circuit parameters and constants listed in Table 10.1 and Table 2.2 in our simulations to implement and evaluate the proposed ASSIST approach.

According to our simulation results, power dissipation of the stochastic bitstream generator circuit is 23μW on average over a period of $100ns$ for generating a 100-bit bitstream composed of equal likelyhood for '0's and '1's. Furthermore, the area estimate of each stochastic bitstream generator circuit in the 14nm technology node according to the transistor count is 0.4μm^2. For a more equitable comparison in terms of area and energy consumption per bit, we have derived (10.1) and (10.2) considering general scaling method [225] to normalize the energy consumption per bit and area of the designs listed in Table 10.2. Based on the general scaling method, voltage and area scale at different rates of U and S, respectively. Thus, the energy consumption is scaled with respect to

$1/SU^2$ and area per device is scaled according to $1/S^2$ [225], as shown below:

$$Energy_{norm} = \frac{Energy_x}{Energy_{MSO}} \times (\frac{1}{S}) \times (\frac{1}{U})^2 = \frac{Energy_x}{Energy_{MSO}} \times (\frac{14nm}{Technology}) \times (\frac{0.8V}{V_{DD}})^2, \tag{10.1}$$

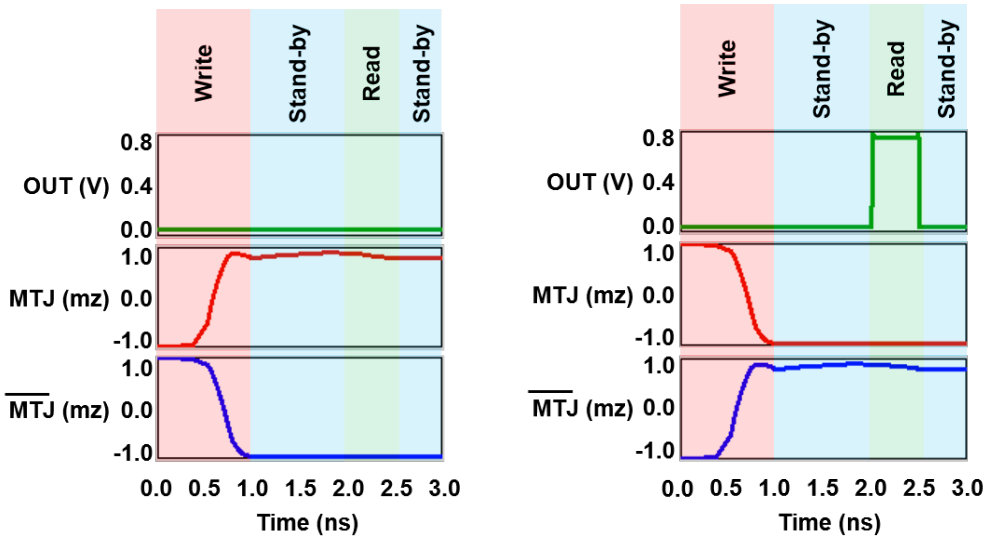
$$Area_{norm} = \frac{Area_x}{Area_{MSO}} \times (\frac{1}{S})^2 = \frac{Area_x}{Area_{MSO}} \times (\frac{14nm}{Technology})^2, \tag{10.2}$$

where, V_{DD} is the nominal voltage of the technology model, $Technology$ refers to the technology node in nanometers, and subscript x refers to the design that we want to scale its power dissipation and area according to the technology models. According to (10.1) and (10.2), MSO reduces energy consumption per bit by $\sim$ 9-fold on average compared to the state-of-the-art TRNGs as listed in Table 5.3. Additionally, MSO offers up to $\sim$ 3-fold area reduction on average compared to the TRNG designs provided in Table 5.3 using the scaling comparison trends accepted in the literature.

Table 10.2: Comparison with recent TRNG designs. [8]

Design	**Technology (V_{DD})**	**Energy**$_{norm}$	**Area**$_{norm}$
[99]	28nm (1.0V)	0.3X	1.25X
[100]	28nm (1.0V)	8.9X	4.8X
[101]	28nm (1.0V)	17.4X	3.7X
This Work	14nm (0.8V)	1X	1X

Furthermore, transient output of a single complementary SHE-MRAM NVM bit-cell shown in Figure 10.1(b) is provided in Figure 10.2. According to our simulation results, writing in a NVM bit-cell requires 155.2fJ on average while reading the content of a NVM bit-cell requires 21.9fJ on average. Additionally, based on our simulation results, the standby energy consumption is 36.4aJ. Moreover, in Figure 10.3, we use the sampling and recovery algorithm discussed in [91, 233] to evaluate the performance of ASSIST for different values of undersampling ratios, $\frac{M}{N}$, for a signal with sparsity level of $\frac{k}{N} = 0.1$ considering $N = 200$ and with RoI that occupies 10% of the entire signal.



(a) (b)

Figure 10.2: Transient output for SHE-MRAM NVM array: writing and reading a (a) '0' bit, and (b) '1' bit. [8]

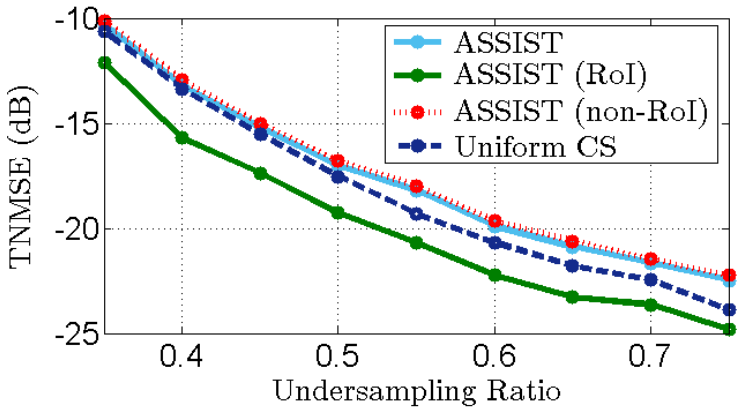


Figure 10.3: TNMSE vs. Undersampling Ratio, $\frac{M}{N}$, for a signal with $\frac{k}{N} = 0.1$, $N = 200$, and RoI occupying 10% of N. [8]

This experiment shows that the proposed ASSIST is able to decrease the Time-Averaged Normalized Mean Squared Error (TNMSE) of RoI coefficients by up to 2dB for various undersampling ratios. This benefit comes at the cost of reduced performance on total recovery error. It is worth noting that for smaller undersampling ratios, ASSIST incurs no additional performance degradation compared to uniform CS for non-RoI entries.

10.3 Conclusion

Recent advances to hardware integration and realization of highly-efficient Compressive Sensing (CS) approaches have inspired novel circuit and architectural-level approaches. These embrace the challenge to design more optimal non-uniform CS solutions that consider device-level constraints for IoT applications wherein lifetime energy, device area, and manufacturing costs are highly-constrained, but meanwhile the sensing environment is rapidly changing. Thus, we develop a novel adaptive hardware-based approach for non-uniform compressive sampling of sparse and time-varying signals. The proposed *Adaptive Sampling of Sparse IoT signals via STochastic-oscillators (ASSIST)* approach intelligently generates the CS measurement matrix by distributing the sensing energy among coefficients by considering the signal characteristics such as sparsity rate and noise level obtained in the previous time step. In our proposed approach, MRAM-based stochastic oscillators are utilized to generate the random bitstreams used in the CS measurement matrix. SPICE and MATLAB circuit-algorithm simulation results indicate that ASSIST efficiently achieves the desired non-uniform recovery of the original signals with varying sparsity rates and noise levels.

CHAPTER 11: CONCLUSION

11.1 Technical Summary

11.1.1 Mitigating Process Variability for Non-Volatile Cache Resilience and Yield

To elevate the reliability and energy-efficiency of emerging NVMs a novel circuit-architecture cross-layer approach called Self-Organized Sub-banks (SOS) is developed and evaluated. SOS organizes the NVM banks into sub-banks and utilizes two different SAs for read operation. An algorithm is designed to assign appropriate SA to each sub-bank based on the PV-based reliability measures of each sub-bank so that a high-resilient SA will be assigned to the sub-banks with high BER and an energy-efficient SA will be assigned to those sub-banks that are adequately reliable based on the results acquired from the POST.

Additionally, SOS-enabled hybrid cache provides a wide-ranging solution to leverage PV in order to improve the performance and reliability of emerging NVM technologies. Our results indicate both STT-MRAM and SOS using MSA or ASA offer up to 88% conservation of the total consumed energy, on average. ASA offers improved reliability and performance, while maintaining a small footprint of $2.5\mu m^2$. Additionally, ASA incurs 0.5-fold, 10.4-fold, 2.3-fold, 3.3-fold, and 1.4-fold area overhead compared to the new MSA, PCSA [10], SPCSA [10], EASA [10], and VISA [10], respectively. Furthermore, our results exhibit that SOS-enabled hybrid cache improves the write performance by 12.4% on average compared to STT-MRAM design. Moreover, the VFDS is reduced by 89% on average in the SOS-enabled hybrid cache using ASA design compared to LLC with STT-MRAM. This improves the mean TDS from 72.5% to 97% across all workloads.

Furthermore, since STT-MRAM suffers from high dynamic energy consumption mostly due to

the write operation, SHE-MRAM is used as a replacement, which offers better write operation performance compared to STT. Moreover, several different write schemes for SHE-MRAM is explored and among those, a high-resilient design, 7T1R, and an energy-efficient design, 1TG1T1R, are chosen to be combined with SOS approach for further improvements of reliability and energy efficiency. In particular, SOS-1TG1T1R outperforms SOS-7T1R in terms of EDP by 1.7-fold, however SOS-7T1R provides increased reliability by providing less than 8% variation in the worst case scenario.

11.1.2 Beyond von Neumann Architectures for Intelligent IoT Edge Processing

A novel framework for sleep power critical mobile applications is proposed to advance energy-sparing and fast NV-SRAM designs. The proposed framework, called Bit-Grained Instant-on Memory (BGIM), is designed to minimize the overall static and leakage energy consumption while providing rapid back-up and instant-on restore operations through the integration of DSH-MRAM devices with SRAM cells. The proposed BGIM cell performs back-up and restore operations within 1ns and 13.2ps, respectively, while consuming 121.51fJ and 1.56fJ, respectively. According to the results, BGIM outperforms similar NV-SRAM cells that utilize emerging devices in their designs. Additionally, BGIM only incurs $0.4\mu m^2$ area overhead compared to the traditional 6T SRAM cell, while eliminating the need for data transmission and a separate NVM macro.

Furthermore, to overcome the conventional SRAM-LUT limitations such as high static power, volatility, and low logic density, we have proposed a novel LUT design using spin-based devices. The proposed Combinational LUT (C-LUT) is a clockless design and a suitable candidate for combinational logic, which can also be combined with a flip-flop circuit to implement sequential logic. According to our simulation results, the standby power dissipation of the proposed C-LUT is 0.31μW, which is reduced by 5.4-fold compared to the SRAM-based LUT. Moreover, the structure

of the proposed SHE-MRAM based C-LUT includes 250 and 768 fewer transistors compared to the SRAM-based LUT and the STT-MRAM based C-LUT, respectively. Additionally, according to the process variation reliability analysis, the C-LUT circuit exhibits $< 0.001\%$ error rate for read and write operations in presence of variations spanning both transistors and MTJs.

Moreover, to advance energy-sparing sampling methods, we propose a spin-based Adaptive Intermittent Quantizer (AIQ) to perform adaptive signal sampling and quantization. The contributions of the developed cross-layer design can be summarized as follows: (1) a novel framework for efficient and intelligent sensing through the integration of resource allocation, quantized compressive sensing, and configurable spin-based devices are introduced using a multilayered approach, (2) the utility of VCMA-MTJ devices within the proposed AIQ architecture are demonstrated to realize rapid and more energy-efficient sampling and signal processing while achieving reduced area footprint compared to conventional CMOS designs is demonstrated, (3) the energy consumption of VCMA-MTJ is formulated and the energy equation that was derived was then utilized for SR/QR optimization, (4) SR and QR trade-off under resource constraints are studied and an energy-aware adaptive SR/QR optimization framework to tune the sampling rate and quantization resolution is demonstrated, and (5) the adaptive SR/QR controller is integrated with the proposed AIQ for energy-efficient signal acquisition. Finally, the novel sampling and reconstruction algorithms, which have been developed in the context of adaptive quantized CS, open the door to broader applications beyond those addressed herein.

Additionally, we have devised a novel non-uniform clock generator called Adaptive Quantization Rate (AQR) generator using MRAM-based stochastic oscillator devices. Our proposed AQR generator considers signal constraints, such as sparsity rate, as well as hardware constraints, such as area and power dissipation, in order to generate the non-uniform clock for the asynchronous CS-ADC. Compared to similar non-uniform clock generators presented in the literature, AQR generator provides significant area reduction of ~ 25-fold on average, while achieving power dissipation

reduction of ~ 6-fold, on average.

Furthermore, a novel framework for efficient and intelligent computing approach through the integration of resource allocation and spin-based devices is introduced to advance energy-sparing sampling methods. The utility of SHE-DWM devices within the proposed Spin-based Logic-In-Memory ADC (SLIM-ADC) architecture is demonstrated to realize rapid and more energy-efficient sampling while achieving reduced area footprint compared to conventional CMOS designs. Moreover, SLIM-ADC takes a step towards the realization of non-Von-Neumann architectures via in-memory computation utilizing SHE-DWM devices. According to our simulation results, the proposed SLIM-ADC offers $\sim$ 200fJ energy consumption on average for each analog conversion or logic operation with up to $1GHz$ speed. Furthermore, our results indicate that the proposed SLIM-ADC outperforms other state of the art spin-based ADC designs by offering $\sim$ 5.5mW improved power dissipation on average. Additionally, a Majority Gate (MG)-based Full-Adder (MG-FA) is implemented using the proposed SLIM-ADC. Our results show that the proposed MG-FA offers $\sim$ 2-fold and $\sim$ 3.8-fold reduced power dissipation on average in 1GHz and 500MHz operating speeds, respectively, compared to the state of the art Full-Adder designs reported herein. Additionally, according to our results, the proposed MG-FA provides $\sim$ 2.3-fold and $\sim$ 1.13-fold reduced delay on average in 1GHz and 500MHz operating speeds, respectively, compared to the state of the art Full-Adder designs reported herein.

Moreover, we have devised a spin-based non-uniform compressive sensing circuit-algorithm solution called Adaptive Sampling of Sparse IoT signals via STochastic-oscillators (ASSIST). High payoff considerations to leverage for device hardware optimization, which are advanced herein, include the signal sparsity and noise levels. According to our simulation result, the MRAM-based Stochastic Oscillator (MSO) used as a TRNG provides significant area improvement of $\sim$ 3-fold while achieving energy consumption per bit reduction of $\sim$ 9-fold, on average, compared to similar TRNGs presented in the literature. Additionally, our circuit-algorithm simulation results indicate

that ASSIST efficiently achieves the desired non-uniform recovery of the original signals with varying sparsity rates and noise levels.

11.2 Technical Insights

A summary of technical insights gained from the proposed research presented in this dissertation is provided below:

- Emerging spin-based circuits offer significant improvements over traditional CMOS-based circuits in terms of static power dissipation and area footprint, which make them a great candidate for IoT applications.
- Recent commercial availability of vertically-integrated spin-based devices provide a foundation for intrinsic device computation.
- Utilization of emerging spin-based devices to realize Compressive Sensing techniques results in significant area and energy consumption improvements within IoT applications.
- Compressive Sensing algorithms are designed to achieve desirable signal acquisition and reconstruction while tolerating certain degree of error. Thus, utilizing emerging spin-based devices within Compressive Sensing approaches can further reduce energy consumption and area footprint while acheiving increase in performance.
- Majority of the energy within Compressive Sensing algorithms is consumed by Vector Matrix Multiplication operations as well as Analog to Digital Conversions. Thus, developing circuits to perform analog computation can results in significant energy consumption and performance improvements.

- Design of low-power Analog to Digital Converters using emerging spin-based devices provides significant energy consumption and area improvements.

11.3 Future Directions

11.3.1 Power Efficient AI Hardware System Design for IoT Edge Sensing and Computing

Recent advances to hardware integration and realization of highly-efficient analog computing approaches have inspired novel circuit and architectural-level innovations that consider device-level constraints for Internet of Things (IoT) applications wherein lifetime energy, device area, and manufacturing costs are highly-constrained. Additionally, recently machine learning approaches have been widely used in IoT applications. However, there is an increasing demand for novel circuits and architectures that can yield several orders of magnitude improvements in energy consumption of machine learning applications while maintaining consistent accuracy. I intend to propose a device-level-to-application-level approach is to integrate front-end signal processing operations within a low-footprint neuromorphic computing array. This cross-cutting beyond-von Neumann view of machine learning is explored within the potential of compressive imaging such as Single-Pixel Camera (SPC) towards the goal of decision-making from data observations rather than reconstruction of the data. This consolidated platform selectively leverages a single memristive post-CMOS device across multiple processing phases to simultaneously reduce the area requirement and energy consumption. Use of a unified platform allows in-the-field adaptation across a continuum of information conversion losses and costs targeted for IoT devices. In this project, I will investigate a novel adaptive hardware-based approach for non-uniform compressive sampling and perform inference on the compressed samples acquired by a SPC without the need for reconstructing the signal utilizing Long Short Term Memory (LSTM) networks on video [234]. Additionally, my proposed approach will eliminate the need for an analog to digital converter and is

able to perform analog processing on the compressed sampled signal. I intend to demonstrate that my proposed approach can achieve orders of magnitude area and speed improvement compared to similar approaches in the literature due to elimination of the bulky analog to digital converter circuit and data storage as well as reduction of the amount of data that requires processing. Regardless of whether or not the hypothesis is validated, the proposed research will advance multiple efforts to produce post-CMOS devices.

11.3.2 Mixed-Signal Reconfigurable Array for Energy-Aware Neuromorphic Processing in IoT

Field Programmable Gate Arrays (FPGAs) are promising candidates for online algorithms requiring dynamic reconfiguration as well as general-purpose computations while minimizing software overheads [235, 236]. However, process variation, soft errors, and hard errors introduces reliability challenges, which results in performance degradation. Thus, fault tolerant FPGAs have been introduced to provide reliable and self-adaptive operations to mitigate these challenges [237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249]. On the other hand, Field Programmable Analog Arrays (FPAAs) are more efficient in terms of computation energy consumption and performance since they enable analog domain computation. As a result, overheads and accuracy loss incurred due to signal conversions from analog to digital domain and vice versa are significantly reduced [250]. Furthermore, the parallelism provided by the reconfigurable fabric can be used for artificial intelligence applications [251].

Furthermore, Neuromorphic computing leveraging analog processing has been shown to be energy, wire-count, and area-efficient [252]. However, the pathways from its software simulation to realizable neuromorphic chips using mixed-signal approaches are underexplored [253]. This project aims to address such need by advancing the research hypothesis that a reprogrammable fabric of a concise palette of analog and digital components can realize an energy-efficient platform for

neuromorphic computing while accommodating adaptable precision, reduced dynamic range, and maintain consistent accuracy despite process variation of emerging devices via inherent reconfigurability features. This research leverages the advantages of mixed-signal processing on a single die to realize neuromorphic architectures yielding orders of magnitude reduction in energy consumption. Thus, I propose investigating a device-level-to-architecture-level approach to integrate front-end signal processing and machine learning operations within a low-footprint reconfigurable fabric that enables mixed-signal processing. This project will advance a new class of chips called Mixed-signal Field Programmable Arrays (MFPAs), which enable high-throughput on-chip learning via established approaches for artificial neural network processing. Mixed-signal techniques combined with in-memory compute geared to the demands of neuromorphic processing will be combined in a field-programmable and run-time adaptable platform.

11.3.3 Intelligent Approaches to Hardware Trojan Detection

In this project, I intend to propose an advanced Intelligent Compressive Sensing Hardware Trojan Detection framework and design methodology to explore the neuromorphic hardware design space in various architecture-to-device granularities to realize an energy-aware, rapid, and accurate hardware trojan detection during design-time and run-time fortified by machine learning methods. Initially, the proposed framework will leverage a top-down approach based on computationally-optimized models of spintronic-based implementation of compressive sensing algorithms. The developed framework will be equipped with intelligent optimization methods such as Genetic Algorithms [254], Bayesian Inference methods such as the ones discussed in [91, 96, 233, 255], and Outlier Detection Algorithms similar to the ones proposed in [256, 257, 258], to realize multi-objective optimizations in terms of accuracy, area, and energy consumption through exploration of the design space and tuning of the algorithm parameters at various granularities. Once the device-level and circuit-level parameters are optimized, a bottom-up modular approach embrac-

ing the essential physics of the spintronic devices will be leveraged to adjust the characteristic of spin-based circuit such that it can realize the desired behavior required for compressive sensing hardware trojan detection. The resulting designs will be evaluated using several benchmarks and the framework and its evaluation results will be disseminated as open source libraries and tools to be used by academia and industry.

APPENDIX A: COPYRIGHT PERMISSIONS

Copyright Clearance Center RightsLink®

Home | Help | Email Support | Sign in | Create Account

IEEE Requesting permission to reuse content from an IEEE publication

Variation-immune resistive Non-Volatile Memory using self-organized sub-bank circuit designs

Conference Proceedings:
2017 18th International Symposium on Quality Electronic Design (ISQED)

Author: Navid Khoshavi

Publisher: IEEE

Date: March 2017

Copyright © 2017, IEEE

Thesis / Dissertation Reuse

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

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

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

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

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

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

BACK CLOSE

Process variation immune and energy aware sense amplifiers for resistive non-volatile memories

Conference Proceedings: 2017 IEEE International Symposium on Circuits and Systems (ISCAS)
Author: Soheil Salehi
Publisher: IEEE
Date: May 2017

Mitigating Process Variability for Non-Volatile Cache Resilience and Yield

Author: Soheil Salehi

Publication: IEEE Transactions on Emerging Topics in Computing

Publisher: IEEE

Date: Dec 31, 1969

Thesis / Dissertation Reuse

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

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

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

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

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

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

BACK CLOSE

Self-Organized Sub-bank SHE-MRAM-based LLC: An energy-efficient and variation-immune read and write architecture

Author: Soheil Salehi,Navid Khoshavi,Ramtin Zand,Ronald F. DeMara

Publication: Integration, the VLSI Journal

Publisher: Elsevier

Date: March 2019

BGIM: Bit-Grained Instant-on Memory Cell for Sleep Power Critical Mobile Applications

Conference Proceedings: 2018 IEEE 36th International Conference on Computer Design (ICCD)

Author: Soheil Salehi

Publisher: IEEE

Date: Oct 2018

Thesis / Dissertation Reuse

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

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

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

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

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

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

BACK CLOSE

Energy-Aware Adaptive Rate and Resolution Sampling of Spectrally Sparse Signals Leveraging VCMA-MTJ Devices

Author: Soheil Salehi

Publication: Emerging and Selected Topics in Circuits and Systems, IEEE Journal on

Publisher: IEEE

Date: Dec. 2018

SLIM-ADC: Spin-based Logic-In-Memory Analog to Digital Converter leveraging SHE-enabled Domain Wall Motion devices

Author: Soheil Salehi,Ronald F. DeMara
Publication: Microelectronics Journal
Publisher: Elsevier
Date: November 2018

MRAM-Based Stochastic Oscillators for Adaptive Non-Uniform Sampling of Sparse Signals in IoT Applications

Conference Proceedings: 2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)
Author: Soheil Salehi
Publisher: IEEE
Date: July 2019

Thesis / Dissertation Reuse

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

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

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

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

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

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

BACK CLOSE

LIST OF REFERENCES

[1] S. Salehi, D. Fan, and R. F. DeMara, "Survey of STT-MRAM Cell Design Strategies: Taxonomy and Sense Amplifier Tradeoffs for Resiliency," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 13, no. 3, pp. 1–16, 2017.

[2] 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.

[3] 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, the VLSI Journal*, vol. 65, pp. 293–307, 3 2018.

[4] S. Salehi and R. F. DeMara, "BGIM: Bit-Grained Instant-on Memory Cell for Sleep Power Critical Mobile Applications," in *Proceedings of the 2018 IEEE 36th International Conference on Computer Design (ICCD)*, (Orlando, FL, USA), pp. 342–345, IEEE, 10 2018.

[5] S. Salehi, M. B. Mashhadi, A. Zaeemzadeh, N. Rahnavard, and R. F. DeMara, "Energy-Aware Adaptive Rate and Resolution Sampling of Spectrally Sparse Signals Leveraging VCMA-MTJ Devices," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, pp. 679–692, 12 2018.

[6] S. Salehi and R. F. DeMara, "SLIM-ADC: Spin-based Logic-In-Memory Analog to Digital Converter leveraging SHE-enabled Domain Wall Motion devices," *Microelectronics Journal*, vol. 81, pp. 137–143, 11 2018.

[7] 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.

[8] S. Salehi, A. Zaeemzadeh, A. Tatulian, N. Rahnavard, and R. F. DeMara, "MRAM-Based Stochastic Oscillators for Adaptive Non-Uniform Sampling of Sparse Signals in IoT Applications," in *Proceedings of the 2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, (Miami, FL, USA), pp. 403–408, IEEE, 9 2019.

[9] S. Salehi and R. F. DeMara, "Adaptive Non-Uniform Compressive Sensing using SOT-MRAM Multibit Crossbar Arrays," *arXiv:1911.08633*, 11 2019.

[10] S. Salehi and R. F. DeMara, "Process Variation Immune and Energy Aware Sense Amplifiers for Resistive Non-Volatile Memories," in *Proceedings of the 2017 IEEE International Symposium on Circuits And Systems (ISCAS)*, (Baltimore, MD, USA), pp. 1–4, IEEE, 2017.

[11] N. Ben-Romdhane, W. S. Zhao, Y. Zhang, J.-O. Klein, Z. H. Wang, and D. Ravelosona, "Design and Analysis of Racetrack Memory Based on Magnetic Domain Wall Motion in

Nanowires," in *Proceedings of the 2014 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, (New York, NY, USA), pp. 71–76, ACM, 2014.

[12] S. Salehi, R. Zand, and R. F. DeMara, "Clockless Spin-based Look-Up Tables with Wide Read Margin," in *Proceedings of the 2019 Great Lakes Symposium on VLSI (GLSVLSI)*, (Tysons Corner, VA, USA), pp. 363–366, ACM, 2019.

[13] S. Salehi, R. Zand, A. Zaeemzadeh, N. Rahnavard, and R. F. DeMara, "AQuRate: MRAM-based Stochastic Oscillator for Adaptive Quantization Rate Sampling of Sparse Signals," in *Proceedings of the 2019 Great Lakes Symposium on VLSI (GLSVLSI)*, (Tysons Corner, VA, USA), pp. 359–362, ACM, 2019.

[14] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, pp. 301–321, 5 2009.

[15] 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, 2017.

[16] Y. Kim, S. H. Choday, and K. Roy, "DSH-MRAM: Differential Spin Hall MRAM for On-Chip Memories," *IEEE Electron Device Letters*, vol. 34, pp. 1259–1261, 10 2013.

[17] W. Kang, W. Lv, Y. Zhang, and W. Zhao, "Low Store Power, High Speed, High Density, Nonvolatile SRAM Design with Spin Hall Effect-Driven Magnetic Tunnel Junctions," *IEEE Transactions on Nanotechnology*, pp. 1–1, 2016.

[18] S. Angizi, Z. He, F. Parveen, and D. Fan, "RIMPA: A New Reconfigurable Dual-Mode In-Memory Processing Architecture with Spin Hall Effect-Driven Domain Wall Motion Device," in *Proceedings of the 2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, (Bochum, Germany), pp. 45–50, IEEE, 7 2017.

[19] H. Kim, S. W. Heo, and C.-Y. You, "Implementation of one-dimensional domain wall dynamics simulator," *AIP Advances*, vol. 7, p. 125231, 12 2017.

[20] E. Martinez, S. Emori, N. Perez, L. Torres, and G. S. D. Beach, "Current-driven dynamics of Dzyaloshinskii domain walls in the presence of in-plane fields: Full micromagnetic and one-dimensional analysis," *Journal of Applied Physics*, vol. 115, p. 213909, 6 2014.

[21] 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, pp. 2134–2138, 12 2017.

[22] S. Salehi Mobarakeh, *Towards Energy-Efficient and Reliable Computing: From Highly-Scaled CMOS Devices to Resistive Memories*. PhD thesis, 1 2016.

[23] N. Khoshavi, R. A. Ashraf, and R. F. DeMara, "Applicability of Power-Gating Strategies for Aging Mitigation of CMOS Logic Paths," in *Proceedings of the 2014 Midwest Symposium on Circuits And Systems (MWSCAS)*, (College Station, TX, USA), pp. 929–932, IEEE, 2014.

[24] R. A. R. 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 *Proceedings of the 2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, (Lisbon, Portugal), pp. 2944–2947, IEEE, 2015.

[25] Z. Sun, H. Li, Y. Chen, and X. Wang, "Voltage Driven Nondestructive Self-Reference Sensing Scheme of Spin-Transfer Torque Memory," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, pp. 2020–2030, 11 2012.

[26] S. Salehi and R. F. DeMara, "Energy and area analysis of a floating-point unit in 15nm CMOS process technology," in *SoutheastCon 2015*, pp. 1–5, IEEE, 2015.

[27] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. B. Tahoori, "Read disturb fault detection in STT-MRAM," in *Proceedings of the 2015 International Test Conference*, pp. 1–7, IEEE, 10 2015.

[28] E. Kultursay, M. Kandemir, A. Sivasubramaniam, and O. Mutlu, "Evaluating STT-RAM as an energy-efficient main memory alternative," in *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pp. 256–267, IEEE, 4 2013.

[29] N. Khoshavi, X. Chen, J. Wang, and R. DeMara, "Bit-Upset Vulnerability Factor for eDRAM Last Level Cache Immunity Analysis," in *Proceedings of 17th International Symposium on Quality Electronic Design (ISQED)*, 2016.

[30] 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, 2016.

[31] 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, pp. 1–7, 8 2016.

[32] S. D. Pyle, D. Fan, and R. F. DeMara, "Compact Spintronic Muller C-Element With Near-Zero Standby Energy," *IEEE Transactions on Magnetics*, vol. 54, pp. 1–7, 2 2018.

[33] A. Roohi, R. F. DeMara, and N. Khoshavi, "Design and evaluation of an ultra-area-efficient fault-tolerant QCA full adder," *Microelectronics Journal*, vol. 46, no. 6, pp. 531–542, 2015.

[34] A. M. Chabi, A. Roohi, R. F. DeMara, S. Angizi, K. Navi, and H. Khademolhosseini, "Cost-efficient QCA reversible combinational circuits based on a new reversible gate," in *18th CSI International Symposium on Computer Architecture and Digital Systems, CADS 2015*, Institute of Electrical and Electronics Engineers Inc., 1 2016.

[35] 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, pp. 450–459, 10 2018.

[36] Y. Zhang, I. Bayram, Y. Wang, H. Li, and Y. Chen, "ADAMS: asymmetric differential STT-RAM cell structure for reliable and high-performance applications," in *Proceedings of the International Conference on Computer-Aided Design*, pp. 9–16, 2013.

[37] 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*, ISLPED '12, (New York, NY, USA), pp. 179–184, ACM, 2012.

[38] E. Eken, Y. Zhang, W. Wen, R. R. R. Joshi, H. H. Li, and Y. Chen, "A Novel Self-Reference Technique for STT-RAM Read and Write Reliability Enhancement," *IEEE Transactions on Magnetics*, vol. 50, pp. 1–4, 11 2014.

[39] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The PARSEC Benchmark Suite: Characterization and Architectural Implications," in *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques*, PACT '08, (New York, NY, USA), pp. 72–81, ACM, 2008.

[40] H. Farkhani, A. Peiravi, and F. Moradi, "Low-Energy Write Operation for 1T-1MTJ STT-RAM Bitcells With Negative Bitline Technique," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 4, pp. 1593–1597, 2016.

[41] T.-K. Chien, L.-Y. Chiou, Y.-S. Tsou, S.-S. Sheu, P.-H. Wang, M.-J. Tsai, and C.-I. Wu, "Write-energy-saving ReRAM-based nonvolatile SRAM with redundant bit-write-aware controller for last-level caches," in *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 1–6, IEEE, 7 2017.

[42] P.-F. Chiu, M.-F. Chang, C.-W. Wu, C.-H. Chuang, S.-S. Sheu, Y.-S. Chen, and M.-J. Tsai, "Low Store Energy, Low VDDmin, 8T2R Nonvolatile Latch and SRAM With Vertical-Stacked Resistive Memory (Memristor) Devices for Low Power Mobile Applications," *IEEE Journal of Solid-State Circuits*, vol. 47, pp. 1483–1496, 6 2012.

[43] 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, pp. 1373–1377, 6 2008.

[44] P. Wang, X. Chen, Y. Chen, H. Li, S. Kang, X. Zhu, and W. Wu, "A 1.0V 45nm nonvolatile magnetic latch design and its robustness analysis," in *2011 IEEE Custom Integrated Circuits Conference (CICC)*, pp. 1–4, IEEE, 9 2011.

[45] K. Huang and Y. Lian, "A Low-Power Low-VDD Nonvolatile Latch Using Spin Transfer Torque MRAM," *IEEE Transactions on Nanotechnology*, vol. 12, pp. 1094–1103, 11 2013.

[46] Y. Shuto, S. Yamamoto, and S. Sugahara, "Nonvolatile static random access memory based on spin-transistor architecture," *Journal of Applied Physics*, vol. 105, p. 07C933, 4 2009.

[47] N. Sakimura, T. Sugibayashi, R. Nebashi, and N. Kasai, "Nonvolatile Magnetic Flip-Flop for Standby-Power-Free SoCs," *IEEE Journal of Solid-State Circuits*, vol. 44, pp. 2244–2250, 8 2009.

[48] T. Endoh, T. Ohsawa, H. Koike, T. Hanyu, and H. Ohno, "Restructuring of memory hierarchy in computing system with spintronics-based technologies," in *2012 Symposium on VLSI Technology (VLSIT)*, pp. 89–90, IEEE, 6 2012.

[49] K.-W. Kwon, S. H. Choday, Y. Kim, X. Fong, S. P. Park, and K. Roy, "SHE-NVFF: spin Hall effect-based nonvolatile flip-flop for power gating architecture," *Electron Device Letters, IEEE*, vol. 35, pp. 488–490, 4 2014.

[50] M. Sadrosadati, A. Mirhosseini, H. Aghilinasab, and H. Sarbazi-Azad, "An efficient DVS scheme for on-chip networks using reconfigurable Virtual Channel allocators," in *2015 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 249–254, IEEE, 7 2015.

[51] S. Tawfik and V. Kursun, "Low Power and High Speed Multi Threshold Voltage Interface Circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, pp. 638–645, 5 2009.

[52] Y. Shuto, S. Yamamoto, and S. Sugahara, "Comparative Study of Power-Gating Architectures for Nonvolatile FinFET-SRAM Using Spintronics-Based Retention Technology," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2015*, (New Jersey), pp. 866–871, IEEE Conference Publications, 2015.

[53] F. Parveen, S. Angizi, Z. He, and D. Fan, "Low power in-memory computing based on dual-mode SOT-MRAM," in *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 1–6, IEEE, 7 2017.

[54] 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*, pp. 23–32, IEEE, 5 2015.

[55] 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.

[56] 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, p. 505002, 12 2017.

[57] X. Tang, G. Kim, P.-E. Gaillardon, and G. De Micheli, "A Study on the Programming Structures for RRAM-Based FPGA Architectures," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, pp. 503–516, 4 2016.

[58] 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, pp. 2605–2613, 9 2014.

[59] A. Attaran, T. D. Sheaves, P. K. Mugula, and H. Mahmoodi, "Static Design of Spin Transfer Torques Magnetic Look Up Tables for ASIC Designs," in *Proceedings of the 2018 on Great Lakes Symposium on VLSI - GLSVLSI '18*, (New York, New York, USA), pp. 507–510, ACM Press, 2018.

[60] D. Suzuki and T. Hanyu, "Design of a highly reliable, high-speed MTJ-based lookup table circuit using fractured logic-in-memory structure," *Japanese Journal of Applied Physics*, vol. 58, p. SBBB10, 2 2019.

[61] D. Suzuki, Y. Lin, M. Natsui, and T. Hanyu, "A 71%-Area-Reduced Six-Input Nonvolatile Lookup-Table Circuit Using a Three-Terminal Magnetic-Tunnel-Junction-Based Single-Ended Structure," *Japanese Journal of Applied Physics*, vol. 52, p. 04CM04, 4 2013.

[62] H. Yoda, H. Sugiyama, T. Inokuchi, Y. Kato, Y. Ohsawa, K. Abe, N. Shimomura, Y. Saito, S. Shirotori, K. Koui, B. Altansargai, S. Oikawa, M. Shimizu, M. Ishikawa, K. Ikegami, Y. Kamiguchi, S. Fujita, and A. Kurobe, "High-Speed Voltage-Control Spintronics Memory (High-Speed VoCSM)," in *2017 IEEE International Memory Workshop (IMW)*, pp. 1–4, IEEE, 5 2017.

[63] D. E. Bellasi, L. Bettini, C. Benkeser, T. Burger, Q. Huang, and C. Studer, "VLSI Design of a Monolithic Compressive-Sensing Wideband Analog-to-Information Converter," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 3, pp. 552–565, 12 2013.

[64] S. Varshney, M. Goswami, and B. Singh, "4-6 Bit Variable Resolution ADC," in *2013 International Symposium on Electronic System Design*, pp. 72–76, IEEE, 12 2013.

[65] T.-F. Wu, C.-R. Ho, and M. S.-W. Chen, "A Flash-Based Non-Uniform Sampling ADC With Hybrid Quantization Enabling Digital Anti-Aliasing Filter," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 9, pp. 2335–2349, 2017.

[66] S. Naraghi, M. Courcy, and M. P. Flynn, "A 9-bit, 14 μW and 0.06 mm$ˆ{2}$ Pulse Position Modulation ADC in 90 nm Digital CMOS," *IEEE Journal of Solid-State Circuits*, vol. 45, pp. 1870–1880, 9 2010.

[67] M. Kurchuk, C. Weltin-Wu, D. Morche, and Y. Tsividis, "Event-Driven GHz-Range Continuous-Time Digital Signal Processor With Activity-Dependent Power Dissipation," *IEEE Journal of Solid-State Circuits*, vol. 47, pp. 2164–2173, 9 2012.

[68] X. Chen, E. A. Sobhy, Z. Yu, S. Hoyos, J. Silva-Martinez, S. Palermo, and B. M. Sadler, "A Sub-Nyquist Rate Compressive Sensing Data Acquisition Front-End," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, pp. 542–551, 9 2012.

[69] M. Wakin, S. Becker, E. Nakamura, M. Grant, E. Sovero, D. Ching, J. Yoo, J. Romberg, A. Emami-Neyestanak, and E. Candes, "A Nonuniform Sampler for Wideband Spectrally-Sparse Environments," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, pp. 516–529, 9 2012.

[70] A. Agarwal, S. K. Mathew, S. K. Hsu, M. A. Anders, H. Kaul, F. Sheikh, R. Ramanarayanan, S. Srinivasan, R. Krishnamurthy, and S. Borkar, "A 320mV-to-1.2V on-die fine-grained reconfigurable fabric for DSP/media accelerators in 32nm CMOS," in *2010 IEEE International Solid-State Circuits Conference - (ISSCC)*, pp. 328–329, IEEE, 2 2010.

[71] D. L. D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[72] E. J. Candès, M. B. Wakin, E. J. Candes, and M. B. Wakin, "An Introduction To Compressive Sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[73] W. Kang, Y. Ran, Y. Zhang, W. Lv, and W. Zhao, "Modeling and Exploration of the Voltage-Controlled Magnetic Anisotropy Effect for the Next-Generation Low-Power and High-Speed MRAM Applications," *IEEE Transactions on Nanotechnology*, vol. 16, no. 3, pp. 387–395, 2017.

[74] P. Khalili Amiri, J. G. Alzate, X. Q. Cai, F. Ebrahimi, Q. Hu, K. Wong, C. Grezes, H. Lee, G. Yu, X. Li, M. Akyol, Q. Shao, J. A. Katine, J. Langer, B. Ocker, and K. L. Wang, "Electric-Field-Controlled Magnetoelectric RAM: Progress, Challenges, and Scaling," *IEEE Transactions on Magnetics*, vol. 51, no. 11, pp. 1–7, 2015.

[75] H. Lee, C. Grezes, A. Lee, F. Ebrahimi, P. Khalili Amiri, and K. L. Wang, "A Spintronic Voltage-Controlled Stochastic Oscillator for Event-Driven Random Sampling," *IEEE Electron Device Letters*, vol. 38, no. 2, pp. 281–284, 2017.

[76] S. Senni, L. Torres, G. Sassatelli, A. Gamatie, and B. Mussard, "Exploring MRAM Technologies for Energy Efficient Systems-On-Chip," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 3, pp. 279–292, 2016.

[77] S. Ghosh, R. V. Joshi, D. Somasekhar, and X. Li, "Guest Editorial Emerging Memories—Technology, Architecture and Applications (First Issue)," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 2, pp. 105–108, 2016.

[78] S. Wang, H. Lee, F. Ebrahimi, P. K. Amiri, K. L. Wang, and P. Gupta, "Comparative Evaluation of Spin-Transfer-Torque and Magnetoelectric Random Access Memory," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 2, pp. 134–145, 2016.

[79] Y. Jiang, Y. Lv, M. Jamali, and J.-P. Wang, "Spin Analog-to-Digital Convertor Using Magnetic Tunnel Junction and Spin Hall Effect," *IEEE Electron Device Letters*, vol. 36, no. 5, pp. 511–513, 2015.

[80] H. Zhang, W. Kang, L. Wang, K. L. Wang, and W. Zhao, "Stateful Reconfigurable Logic via a Single-Voltage-Gated Spin Hall-Effect Driven Magnetic Tunnel Junction in a Spintronic Memory," *IEEE Transactions on Electron Devices*, vol. 64, pp. 4295–4301, 10 2017.

[81] Wang Kang, Liuyang Zhang, J.-O. Klein, Youguang Zhang, D. Ravelosona, and Weisheng Zhao, "Reconfigurable Codesign of STT-MRAM Under Process Variations in Deeply Scaled Technology," *IEEE Transactions on Electron Devices*, vol. 62, pp. 1769–1777, 6 2015.

[82] K. Yogendra, M.-C. Chen, X. Fong, and K. Roy, "Domain wall motion-based low power hybrid spin-CMOS 5-bit Flash Analog Data Converter," in *Sixteenth International Symposium on Quality Electronic Design*, pp. 604–609, IEEE, 2015.

[83] W. Zhao, C. Chappert, V. Javerliac, and J.-P. Nozière, "High speed, high stability and low power sensing amplifier for MTJ/CMOS hybrid logic circuits," *Magnetics, IEEE Transactions on*, vol. 45, no. 10, pp. 3784–3787, 2009.

[84] S. Sarvotham, D. Baron, R. G. Baraniuk, S. Sarvotham, D. Baron, and R. G. Baraniuk, "Measurements vs. Bits: Compressed Sensing meets Information Theory," *Allerton Conference on Communication, Control and Computing*, 9 2006.

[85] A. Zymnis, S. Boyd, and E. Candes, "Compressed Sensing With Quantized Measurements," *IEEE Signal Processing Letters*, vol. 17, pp. 149–152, 2 2010.

[86] W. Dai and O. Milenkovic, "Information Theoretical and Algorithmic Approaches to Quantized Compressive Sensing," *IEEE Transactions on Communications*, vol. 59, pp. 1857–1866, 7 2011.

[87] J. N. Laska and R. G. Baraniuk, "Regime Change: Bit-Depth Versus Measurement-Rate in Compressive Sensing," *IEEE Transactions on Signal Processing*, vol. 60, pp. 3496–3505, 7 2012.

[88] Y. Wang, H. Yu, L. Ni, G.-B. Huang, M. Yan, C. Weng, W. Yang, and J. Zhao, "An Energy-Efficient Nonvolatile In-Memory Computing Architecture for Extreme Learning Machine by Domain-Wall Nanowire Devices," *IEEE Transactions on Nanotechnology*, vol. 14, pp. 998–1012, 11 2015.

[89] P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, "PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory," in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, pp. 27–39, IEEE, 6 2016.

[90] S. Kleinfelder, SukHwan Lim, Xinqiao Liu, and A. El Gamal, "A 10000 frames/s CMOS digital pixel sensor," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 12, pp. 2049–2059, 2001.

[91] A. Zaeemzadeh, M. Joneidi, and N. Rahnavard, "Adaptive non-uniform compressive sampling for time-varying signals," in *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, (Baltimore, MD), pp. 1–6, IEEE, 3 2017.

[92] M. Boloursaz Mashhadi, N. Salarieh, E. S. Farahani, F. Marvasti, M. B. Mashhadi, N. Salarieh, E. S. Farahani, and F. Marvasti, "Level crossing speech sampling and its sparsity promoting reconstruction using an iterative method with adaptive thresholding," *IET Signal Processing*, vol. 11, no. 6, pp. 721–726, 2017.

[93] X. Li, B. Taylor, Y. Chien, and L. T. Pileggi, "Adaptive Post-silicon Tuning for Analog Circuits: Concept, Analysis and Optimization," in *Proceedings of the 2007 IEEE/ACM International Conference on Computer-aided Design*, ICCAD '07, (Piscataway, NJ, USA), pp. 450–457, IEEE Press, 2007.

[94] B. Shahrasbi and N. Rahnavard, "Model-Based Nonuniform Compressive Sampling and Recovery of Natural Images Utilizing a Wavelet-Domain Universal Hidden Markov Model," *IEEE Transactions on Signal Processing*, vol. 65, pp. 95–104, 1 2017.

[95] M. Sharad, Deliang Fan, and K. Roy, "Low power and compact mixed-mode signal processing hardware using spin-neurons," in *International Symposium on Quality Electronic Design (ISQED)*, pp. 189–195, IEEE, 3 2013.

[96] A. Zaeemzadeh, J. Haddock, N. Rahnavard, and D. Needell, "A Bayesian Approach for Asynchronous Parallel Sparse Recovery," in *52nd Asilomar Conference on Signals, Systems, and Computers*, (Pacific Grove, CA), pp. 1980–1984, IEEE, 2018.

[97] D. Bellasi, L. Bettini, T. Burger, Q. Huang, C. Benkeser, and C. Studer, "A 1.9 GS/s 4-bit sub-Nyquist flash ADC for 3.8 GHz compressive spectrum sensing in 28 nm CMOS," in *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 101–104, IEEE, 8 2014.

[98] N. Rahnavard, A. Talari, and B. Shahrasbi, "Non-uniform compressive sensing," in *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pp. 212–219, IEEE, 2011.

[99] D. Vodenicarevic, N. Locatelli, A. Mizrahi, J. Friedman, A. Vincent, M. Romera, A. Fukushima, K. Yakushiji, H. Kubota, S. Yuasa, S. Tiwari, J. Grollier, and D. Querlioz, "Low-Energy Truly Random Number Generation with Superparamagnetic Tunnel Junctions for Unconventional Computing," *Physical Review Applied*, vol. 8, p. 054045, 11 2017.

[100] Y. Qu, J. Han, B. F. Cockburn, W. Pedrycz, Y. Zhang, and W. Zhao, "A True Random Number Generator Based on Parallel STT-MTJs," in *Proceedings of the Conference on Design, Automation & Test in Europe (DATE '17)*, pp. 606–609, 2017.

[101] Y. Wang, H. Cai, L. Alves De Barros Naviner, J.-O. Klein, and W. Zhao, "A Novel Circuit Design of True Random Number Generator Using Magnetic Tunnel Junction," in *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pp. 123–128, 2016.

[102] N. Rangarajan, A. Parthasarathy, and S. Rakheja, "A Spin-based True Random Number Generator Exploiting the Stochastic Precessional Switching of Nanomagnets," *Journal of Applied Physics*, vol. 121, p. 223905, 6 2017.

[103] H. Lee, F. Ebrahimi, P. K. Amiri, and K. L. Wang, "Design of high-throughput and low-power true random number generator utilizing perpendicularly magnetized voltage-controlled magnetic tunnel junction," *AIP Advances*, vol. 7, p. 055934, 5 2017.

[104] K. Chen, J. Han, and F. Lombardi, "On the restore operation in MTJ-based nonvolatile SRAM cells," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 11, pp. 2695–2699, 2015.

[105] R. H. Koch, J. A. Katine, and J. Z. Sun, "Time-resolved reversal of spin-transfer switching in a nanomagnet," *Physical review letters*, vol. 92, no. 8, p. 88302, 2004.

[106] T. Devolder, C. Chappert, J. A. Katine, M. J. Carey, and K. Ito, "Distribution of the magnetization reversal duration in subnanosecond spin-transfer switching," *Physical Review B*, vol. 75, no. 6, p. 64402, 2007.

[107] 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, p. 103001, 10 2014.

[108] E. Eken, I. Bayram, Y. Zhang, B. Yan, W. Wu, H. H. Li, and Y. Chen, "Giant Spin-Hall assisted STT-RAM and logic design," *Integration, the VLSI Journal*, pp. –, 2017.

[109] 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, 2017.

[110] 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, 2012.

[111] Y. K. Upadhyaya, M. K. Gupta, M. Hasan, and S. Maheshwari, "High-Density Magnetic Flash ADC Using Domain-Wall Motion and Pre-Charge Sense Amplifiers," *IEEE Transactions on Magnetics*, vol. 52, pp. 1–10, 6 2016.

[112] Q. Dong, K. Yang, L. Fick, D. Fick, D. Blaauw, and D. Sylvester, "Low-Power and Compact Analog-to-Digital Converter Using Spintronic Racetrack Memory Devices," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, pp. 907–918, 1 2017.

[113] J. Torrejon, J. Kim, J. Sinha, S. Mitani, M. Hayashi, M. Yamanouchi, and H. Ohno, "Interface control of the magnetic chirality in CoFeB/MgO heterostructures with heavy-metal underlayers," *Nature Communications*, vol. 5, p. 4655, 8 2014.

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

[115] R. Zand, K. Y. Camsari, I. Ahmed, S. D. Pyle, C. H. Kim, S. Datta, and R. F. DeMara, "R-DBN: A Resistive Deep Belief Network Architecture Leveraging the Intrinsic Behavior of Probabilistic Devices," 2017.

[116] 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.

[117] W. Kang, E. Deng, J.-O. Klein, Y. Y. Zhang, Y. Y. Zhang, C. Chappert, D. Ravelosona, and W. Zhao, "Separated precharge sensing amplifier for deep submicrometer MTJ/CMOS hybrid logic circuits," *IEEE Transactions on Magnetics*, vol. 50, no. 6, pp. 1–5, 2014.

[118] S. Motaman, S. Ghosh, and J. P. Kulkarni, "A novel slope detection technique for robust STTRAM sensing," in *2015 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 7–12, IEEE, 7 2015.

[119] T. Na, J. Kim, J. P. Kim, S. H. Kang, and S.-O. Jung, "An Offset-Canceling Triple-Stage Sensing Circuit for Deep Submicrometer STT-RAM," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, pp. 1620–1624, 7 2014.

[120] W. Kang, Z. Li, J.-O. Klein, Y. Chen, Y. Zhang, D. Ravelosona, C. Chappert, and W. Zhao, "Variation-Tolerant and Disturbance-Free Sensing Circuit for Deep Nanometer STT-MRAM," *IEEE Transactions on Nanotechnology*, vol. 13, pp. 1088–1092, 11 2014.

[121] K. Kim and C. Yoo, "Variation-Tolerant Sensing Circuit for Spin-Transfer Torque MRAM," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, pp. 1134–1138, 12 2015.

[122] F. Ren, H. Park, R. Dorrance, Y. Toriyama, C.-K. K. Yang, and D. Markovic, "A body-voltage-sensing-based short pulse reading circuit for spin-torque transfer RAMs (STT-RAMs)," in *Thirteenth International Symposium on Quality Electronic Design (ISQED)*, pp. 275–282, IEEE, 3 2012.

[123] Jung Pill Kim, Taehyun Kim,Wuyang Hao, Hari M. Rao, Kangho Lee, Xiaochun Zhu, Xia Li and N. M. N. Y. Wah Hsu, Seung H. Kang, "A 45nm 1Mb Embedded STT-MRAM with design techniques to minimize read-disturbance," *VLSI Circuits (VLSIC), 2011 Symposium on*, 2011.

[124] W. S. Zhao, Y. Zhang, T. Devolder, J.-O. Klein, D. Ravelosona, C. Chappert, and P. Mazoyer, "Failure and reliability analysis of STT-MRAM," *Microelectronics Reliability*, vol. 52, no. 9, pp. 1848–1852, 2012.

[125] J.-T. Choi, G.-H. Kil, K.-B. Kim, and Y.-H. Song, "Novel Self-Reference Sense Amplifier for Spin-Transfer-Torque Magneto-Resistive Random Access Memory," *JSTS:Journal of Semiconductor Technology and Science*, vol. 16, pp. 31–38, 2 2016.

[126] Hochul Lee, J. G. Alzate, R. Dorrance, Xue Qing Cai, D. Markovic, P. Khalili Amiri, and K. L. Wang, "Design of a Fast and Low-Power Sense Amplifier and Writing Circuit for High-Speed MRAM," *IEEE Transactions on Magnetics*, vol. 51, pp. 1–7, 5 2015.

[127] Y. Emre, C. Yang, K. Sutaria, Y. Cao, and C. Chakrabarti, "Enhancing the Reliability of STT-RAM through Circuit and System Level Techniques," in *2012 IEEE Workshop on Signal Processing Systems*, pp. 125–130, IEEE, 10 2012.

[128] Y. Lakys, W. S. Zhao, T. Devolder, Y. Zhang, J.-O. Klein, D. Ravelosona, and C. Chappert, "Self-Enabled "Error-Free" Switching Circuit for Spin Transfer Torque MRAM and Logic," *IEEE Transactions on Magnetics*, vol. 48, pp. 2403–2406, 9 2012.

[129] W. Zhao, T. Devolder, Y. Lakys, J. Klein, C. Chappert, and P. Mazoyer, "Design considerations and strategies for high-reliable STT-MRAM," *Microelectronics Reliability*, vol. 51, pp. 1454–1458, 9 2011.

[130] T. Kawahara, K. Ito, R. Takemura, and H. Ohno, "Spin-transfer torque RAM technology: Review and prospect," *Microelectronics Reliability*, vol. 52, pp. 613–627, 4 2012.

[131] J. Li, P. Ndai, A. Goel, S. Salahuddin, and K. Roy, "Design Paradigm for Robust Spin-Torque Transfer Magnetic RAM (STT MRAM) From Circuit/Architecture Perspective," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, pp. 1710–1723, 12 2010.

[132] D. Chabi, W. Zhao, J.-O. Klein, and C. Chappert, "Design and Analysis of Radiation Hardened Sensing Circuits for Spin Transfer Torque Magnetic Memory and Logic," *IEEE Transactions on Nuclear Science*, vol. 61, pp. 3258–3264, 12 2014.

[133] J. Yang, P. Wang, Y. Zhang, Y. Cheng, W. Zhao, Y. Chen, and H. H. Li, "Radiation-Induced Soft Error Analysis of STT-MRAM: A Device to Circuit Approach," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 35, pp. 380–393, 3 2016.

[134] 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, p. 405003, 10 2014.

[135] G. Tsiligiannis, L. Dilillo, A. Bosio, P. Girard, A. Todri, A. Virazel, S. S. McClure, A. D. Touboul, F. Wrobel, and F. Saigne, "Testing a Commercial MRAM Under Neutron and Alpha Radiation in Dynamic Mode," *IEEE Transactions on Nuclear Science*, vol. 60, pp. 2617–2622, 8 2013.

[136] W. Kang, W. Zhao, Y. Zhang, J.-O. Klein, C. Chappert, D. Ravelosona, J.-O. Klein, C. Chappert, W. Kang, W. Zhao, and Y. Zhang, "High reliability sensing circuit for deep submicron spin transfer torque magnetic random access memory," *Electronics Letters*, vol. 49, pp. 1283–1285, 9 2013.

[137] W. Kang, L. Zhang, W. Zhao, J.-O. Klein, Y. Zhang, D. Ravelosona, and C. Chappert, "Yield and Reliability Improvement Techniques for Emerging Nonvolatile STT-MRAM," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 5, pp. 28–39, 3 2015.

[138] W. Xu, H. Sun, X. Wang, Y. Chen, and T. Zhang, "Design of Last-Level On-Chip Cache Using Spin-Torque Transfer RAM (STT RAM)," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, pp. 483–493, 3 2011.

[139] X. Zhou, N. Cui, Z. Li, F. Liang, and T. Huang, "Hierarchical Gaussianization for image classification," 2009.

[140] Xiaobin Wang, Yiran Chen, Hai Li, D. Dimitrov, and H. Liu, "Spin Torque Random Access Memory Down to 22 nm Technology," *IEEE Transactions on Magnetics*, vol. 44, pp. 2479–2482, 11 2008.

[141] K. Ono, T. Kawahara, R. Takemura, K. Miura, H. Yamamoto, M. Yamanouchi, J. Hayakawa, K. Ito, H. Takahashi, S. Ikeda, H. Hasegawa, H. Matsuoka, and H. Ohno, "A disturbance-free read scheme and a compact stochastic-spin-dynamics-based MTJ circuit model for Gb-scale SPRAM," in *2009 IEEE International Electron Devices Meeting (IEDM)*, pp. 1–4, IEEE, 12 2009.

[142] A. Raychowdhury, "Pulsed READ in spin transfer torque (STT) memory bitcell for lower READ disturb," in *2013 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pp. 34–35, IEEE, 7 2013.

[143] R. Takemura, T. Kawahara, K. Ono, K. Miura, H. Matsuoka, and H. Ohno, "Highly-scalable disruptive reading scheme for Gb-scale SPRAM and beyond," in *2010 IEEE International Memory Workshop*, pp. 1–2, IEEE, 2010.

[144] H. Tanizaki, T. Tsuji, J. Otani, Y. Yamaguchi, Y. Murai, H. Furuta, S. Ueno, T. Oishi, M. Hayashikoshi, and H. Hidaka, "A high-density and high-speed 1T-4MTJ MRAM with Voltage Offset Self-Reference Sensing Scheme," in *2006 IEEE Asian Solid-State Circuits Conference*, pp. 303–306, IEEE, 2006.

[145] Gitae Jeong, Wooyoung Cho, Sujin Ahn, Hongsik Jeong, Gwanhyeob Koh, Youngnam Hwang, and Kinam Kim, "A 0.24μm 2.0V 1T1MTJ 16kb NV magnetoresistance RAM with self reference sensing," in *2003 IEEE International Solid-State Circuits Conference, 2003. Digest of Technical Papers. ISSCC.*, vol. 1, pp. 280–281, IEEE.

[146] E. Au, W.-H. Ki, W. Mow, S. Hung, and C. Wong, "A Novel Current-Mode Sensing Scheme for Magnetic Tunnel Junction MRAM," *IEEE Transactions on Magnetics*, vol. 40, pp. 483–488, 3 2004.

[147] K. Tsuchida, T. Inaba, K. Fujita, Y. Ueda, T. Shimizu, Y. Asao, T. Kajiyama, M. Iwayama, K. Sugiura, S. Ikegawa, T. Kishi, T. Kai, M. Amano, N. Shimomura, H. Yoda, and Y. Watanabe, "A 64Mb MRAM with clamped-reference and adequate-reference schemes," in *2010 IEEE International Solid-State Circuits Conference - (ISSCC)*, pp. 258–259, IEEE, 2 2010.

[148] D. Halupka, S. Huda, W. Song, A. Sheikholeslami, K. Tsunoda, C. Yoshida, and M. Aoki, "Negative-resistance read and write schemes for STT-MRAM in 0.13µm CMOS," in *2010 IEEE International Solid-State Circuits Conference - (ISSCC)*, pp. 256–257, IEEE, 2 2010.

[149] Yiran Chen, Hai Li, Xiaobin Wang, Wenzhong Zhu, Wei Xu, and Tong Zhang, "A nondestructive self-reference scheme for Spin-Transfer Torque Random Access Memory (STT-RAM)," in *2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010)*, pp. 148–153, IEEE, 3 2010.

[150] Y. Chen, H. Li, X. Wang, W. Zhu, W. Xu, and T. Zhang, "A 130 nm 1.2 V/3.3 V 16 Kb Spin-Transfer Torque Random Access Memory With Nondestructive Self-Reference Sensing Scheme," *IEEE Journal of Solid-State Circuits*, vol. 47, pp. 560–573, 2 2012.

[151] J. Kim, K. Ryu, S. H. Kang, and S.-O. Jung, "A Novel Sensing Circuit for Deep Submicron Spin Transfer Torque MRAM (STT-MRAM)," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, pp. 181–186, 1 2012.

[152] J. Das, S. M. Alam, and S. Bhanja, "Non-destructive variability tolerant differential read for non-volatile logic," in *2012 IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 178–181, IEEE, 8 2012.

[153] W. Kang, W. Zhao, Z. Wang, Y. Zhang, J.-O. Klein, Y. Zhang, C. Chappert, and D. Ravelosona, "A low-cost built-in error correction circuit design for STT-MRAM reliability improvement," *Microelectronics Reliability*, vol. 53, pp. 1224–1229, 9 2013.

[154] J. Kim, T. Na, J. P. Kim, S. H. Kang, and S.-O. Jung, "A Split-Path Sensing Circuit for Spin Torque Transfer MRAM," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 61, pp. 193–197, 3 2014.

[155] T. M. Maffitt, J. K. DeBrosse, J. A. Gabric, E. T. Gow, M. C. Lamorey, J. S. Parenteau, D. R. Willmott, M. A. Wood, and W. J. Gallagher, "Design considerations for MRAM," *IBM Journal of Research and Development*, vol. 50, pp. 25–39, 1 2006.

[156] J. Kim, K. Ryu, J. P. Kim, S. H. Kang, and S.-O. Jung, "STT-MRAM Sensing Circuit With Self-Body Biasing in Deep Submicron Technologies," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, pp. 1630–1634, 7 2014.

[157] W. Kang, Z. Li, Z. Wang, E. Deng, J.-O. O. Klein, Y. Zhang, C. Chappert, D. Ravelosona, and W. Zhao, "Variation-tolerant high-reliability sensing scheme for deep submicrometer STT-MRAM," *IEEE Transactions on Magnetics*, vol. 50, pp. 1–4, 11 2014.

[158] C. Kim, K. Kwon, C. Park, S. Jang, and J. Choi, "7.4 A covalent-bonded cross-coupled current-mode sense amplifier for STT-MRAM with 1T1MTJ common source-line structure array," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, pp. 1–3, IEEE, 2 2015.

[159] L. Yang, Y. Cheng, Y. Wang, H. Yu, W. Zhao, and A. Todri-Sanial, "A body-biasing of readout circuit for STT-RAM with improved thermal reliability," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1530–1533, IEEE, 5 2015.

[160] W. Kang, T. Pang, Y. Zhang, D. Ravelosona, and W. Zhao, "Dynamic Reference Sensing Scheme for Deeply Scaled STT-MRAM," in *2015 IEEE International Memory Workshop (IMW)*, pp. 1–4, IEEE, 5 2015.

[161] Y. Ran, W. Kang, Y. Zhang, J.-O. Klein, and W. Zhao, "Read disturbance issue for nanoscale STT-MRAM," in *2015 IEEE Non-Volatile Memory System and Applications Symposium (NVMSA)*, pp. 1–6, IEEE, 8 2015.

[162] C.-L. Su and A. M. Despain, "Cache design trade-offs for power and performance optimization: a case study," in *Proceedings of the 1995 international symposium on Low power design*, pp. 63–68, 1995.

[163] C.-L. Su and A. M. Despain, "Cache designs for energy efficiency," in *System Sciences, 1995. Proceedings of the Twenty-Eighth Hawaii International Conference on*, vol. 1, pp. 306–315, 1995.

[164] N. S. Kim, K. Flautner, D. Blaauw, and T. Mudge, "Drowsy instruction caches. leakage power reduction using dynamic voltage scaling and cache sub-bank prediction," in *Microarchitecture, 2002.(MICRO-35). Proceedings. 35th Annual IEEE/ACM International Symposium on*, pp. 219–230, 2002.

[165] O. Olorode and M. Nourani, "Improving Performance in Sub-Block Caches with Optimized Replacement Policies," *J. Emerg. Technol. Comput. Syst.*, vol. 11, no. 4, p. 41:1–41:22, 2015.

[166] J. Bunda, W. Athas, and D. Fussell, "Evaluating power implications of CMOS microprocessor design decisions," in *Proceedings of the 1994 International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 147–152, 1994.

[167] S. Mittal, J. S. Vetter, and D. Li, "A Survey Of Architectural Approaches for Managing Embedded DRAM and Non-Volatile On-Chip Caches," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, pp. 1524–1537, 6 2015.

[168] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen, "A Novel Architecture of the 3D Stacked MRAM L2 Cache for CMPs," in *Proceedings of 15th International Symposium on High Performance Computer Architecture*, pp. 239–249, 2009.

[169] Z. Wang, D. A. Jiménez, C. Xu, G. Sun, and Y. Xie, "Adaptive Placement and Migration Policy for an STT-RAM-based Hybrid Cache," in *Proceedings of 20th International Symposium on High Performance Computer Architecture (HPCA)*, pp. 13–24, 2014.

[170] J. Ahn, S. Yoo, and K. Choi, "Prediction Hybrid Cache: An Energy-Efficient STT-RAM Cache Architecture," *IEEE Transactions on Computers*, vol. 65, pp. 940–951, 3 2016.

[171] A. Jadidi, M. Arjomand, and H. Sarbazi-Azad, "High-endurance and performance-efficient design of hybrid cache architectures through adaptive line replacement," in *IEEE/ACM International Symposium on Low Power Electronics and Design*, pp. 79–84, IEEE, 8 2011.

[172] S. Kirolos, J. Laska, M. Wakin, M. Duarte, D. Baron, T. Ragheb, Y. Massoud, and R. Baraniuk, "Analog-to-Information Conversion via Random Demodulation," in *2006 IEEE Dallas/CAS Workshop on Design, Applications, Integration and Software*, pp. 71–74, IEEE, 2006.

[173] J. N. Laska, S. Kirolos, M. F. Duarte, T. S. Ragheb, R. G. Baraniuk, and Y. Massoud, "Theory and implementation of an analog-to-information converter using random demodulation," in *IEEE International Symposium on Circuits and Systems, 2007. ISCAS 2007*, pp. 1959–1962, 2007.

[174] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond Nyquist: Efficient Sampling of Sparse Bandlimited Signals," *Information Theory, IEEE Transactions on*, vol. 56, no. 1, pp. 520–544, 2010.

[175] M. Mishali and Y. C. Eldar, "Blind Multiband Signal Reconstruction: Compressed Sensing for Analog Signals," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 993–1009, 2009.

[176] M. Mishali and Y. C. Eldar, "From Theory to Practice: Sub-Nyquist Sampling of Sparse Wideband Analog Signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 375–391, 2010.

[177] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, "Xampling: analog to digital at sub-Nyquist rates," *IET Circuits, Devices & Systems*, vol. 5, no. 1, p. 8, 2011.

[178] L. Jacques, K. Degraux, and C. De Vleeschouwer, "Quantized Iterative Hard Thresholding: Bridging 1-bit and High-Resolution Quantized Compressed Sensing," 2013.

[179] H.-J. M. Shi, M. Case, X. Gu, S. Tu, and D. Needell, "Methods for quantized compressed sensing," in *2016 Information Theory and Applications Workshop (ITA)*, pp. 1–9, IEEE, 2016.

[180] P. T. Boufounos, "Greedy sparse signal reconstruction from sign measurements," in *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, pp. 1305–1309, IEEE, 2009.

[181] Y. Plan and R. Vershynin, "Robust 1-bit Compressed Sensing and Sparse Logistic Regression: A Convex Programming Approach," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 482–494, 2013.

[182] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-Bit Compressive Sensing via Binary Stable Embeddings of Sparse Vectors," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2082–2102, 2013.

[183] J. N. Laska, Zaiwen Wen, Wotao Yin, and R. G. Baraniuk, "Trust, But Verify: Fast and Accurate Signal Recovery From 1-Bit Compressive Measurements," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5289–5301, 2011.

[184] K. Knudson, R. Saab, and R. Ward, "One-Bit Compressive Sensing With Norm Estimation," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2748–2758, 2016.

[185] U. S. Kamilov, A. Bourquard, A. Amini, and M. Unser, "One-Bit Measurements With Adaptive Thresholds," *IEEE Signal Processing Letters*, vol. 19, no. 10, pp. 607–610, 2012.

[186] C. Qian and J. Li, "ADMM for harmonic retrieval from one-bit sampling with time-varying thresholds," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3699–3703, IEEE, 2017.

[187] S. Marano, V. Matta, and P. Willett, "Quantizer precision for distributed estimation in a large sensor network," *IEEE Transactions on Signal Processing*, 2006.

[188] M. A. Davenport, J. N. Laska, J. R. Treichler, and R. G. Baraniuk, "The Pros and Cons of Compressive Sensing for Wideband Signal Acquisition: Noise Folding versus Dynamic Range," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4628–4642, 2012.

[189] H. Fu and Y. Chi, "Quantized Spectral Compressed Sensing: Cramer-Rao Bounds and Recovery Algorithms," *Arxiv preprint arXiv:1710.03654*, 2017.

[190] A. Gilbert and P. Indyk, "Sparse recovery using sparse matrices," in *Proceedings of the IEEE*, vol. 98, pp. 937–947, Institute of Electrical and Electronics Engineers, 2010.

[191] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank, "Efficient and robust compressed sensing using optimized expander graphs," *IEEE Transactions on information theory*, vol. 55, no. 9, pp. 4299–4308, 2009.

[192] H. T. Kung and S. J. Tarsa, "Partitioned compressive sensing with neighbor-weighted decoding," in *MILITARY COMMUNICATIONS CONFERENCE, 2011 - MILCOM 2011*, pp. 149–156, 2011.

[193] L. Gan, "Block Compressed Sensing of Natural Images," in *15th International Conference on Digital Signal Processing*, pp. 403–406, 5 2007.

[194] Y. Yu, B. Wang, and L. Zhang, "Saliency-based compressive sampling for image signals," *IEEE signal processing letters*, vol. 17, no. 11, pp. 973–976, 2010.

[195] Y. Shen, W. Hu, R. Rana, and C. T. Chou, "Nonuniform Compressive Sensing for Heterogeneous Wireless Sensor Networks," in *IEEE Sensors Journal,*, vol. 13, pp. 2120–2128, 1 2013.

[196] Y. Liu, X. Zhu, L. Zhang, and S. H. Cho, "Expanding Window Compressed Sensing for Non-Uniform Compressible Signals," *Sensors*, vol. 12, pp. 13034–13057, 1 2012.

[197] M. B. Mashhadi, S. Gazor, N. Rahnavard, and F. Marvasti, "Feedback Acquisition and Reconstruction of Spectrum-Sparse Signals by Predictive Level Comparisons," *IEEE Signal Processing Letters*, vol. 25, pp. 496–500, 4 2017.

[198] 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)*, pp. 52–57, IEEE, 2017.

[199] X. Chen, N. Khoshavi, J. Zhou, D. Huang, R. DeMara, J. Wang, W. Wen, and Y. Chen, "AOS: Adaptive Overwrite Scheme for Energy Efficient MLC STT-RAM Cache," in *Proceedings of 53nd Annual Design Automation Conference (DAC)*, 2016.

[200] X. Chen, N. Khoshavi, R. F. DeMara, J. Wang, D. Huang, W. Wen, and Y. Chen, "Energy-Aware Adaptive Restore Schemes for MLC STT-RAM Cache," *IEEE Transactions on Computers*, vol. PP, no. 99, p. 1, 2016.

[201] 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)*, pp. 1–4, IEEE, 8 2015.

[202] Arizona State University (ASU), "22nm Predictive Technology Model (PTM), accessed on 20 March 2018, available at: http://ptm.asu.edu/."

[203] Y. Ye, F. Liu, M. Chen, S. Nassif, and Y. Cao, "Statistical Modeling and Simulation of Threshold Variation Under Random Dopant Fluctuations and Line-Edge Roughness," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, pp. 987–996, 6 2011.

[204] 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 CoFeB/MgO(001)/CoFeB magnetic tunnel junctions for read-head applications," *Applied Physics Letters*, vol. 87, p. 072503, 8 2005.

[205] E. Deng, W. Kang, Y. Zhang, J.-O. Klein, C. Chappert, and W. Zhao, "Design Optimization and Analysis of Multicontext STT-MTJ/CMOS Logic Circuits," *IEEE Transactions on Nanotechnology*, vol. 14, pp. 169–177, 1 2015.

[206] D. Cheng, H. Hsiung, B. Liu, J. Chen, J. Zeng, R. Govindan, and S. K. Gupta, "A new march test for process-variation induced delay faults in srams," in *2013 22nd Asian Test Symposium*, pp. 115–122, 2013.

[207] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The PARSEC benchmark suite," in *Proceedings of the 17th international conference on Parallel architectures and compilation techniques - PACT '08*, (New York, New York, USA), p. 72, ACM Press, 2008.

[208] A. Chintaluri, H. Naeimi, S. Natarajan, and A. Raychowdhury, "Analysis of Defects and Variations in Embedded Spin Transfer Torque (STT) MRAM Arrays," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 3, pp. 319–329, 2016.

[209] A. Valero, J. Sahuquillo, P. Lopez, and J. Duato, "Design of Hybrid Second-Level Caches," *IEEE Transactions on Computers*, vol. 64, no. 7, pp. 1884–1897, 2015.

[210] N. Khoshavi, X. Chen, J. Wang, and R. F. DeMara, "Read-Tuned STT-RAM and eDRAM Cache Hierarchies for Throughput and Energy Enhancement," in *arXiv preprint arXiv:1607.08086*, 2016.

[211] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "Nvsim: A Circuit-level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 31, no. 7, pp. 994–1007, 2012.

[212] A. Patel, F. Afram, and K. Ghose, "Marss-x86: A qemu-based micro-architectural and systems simulator for x86 multicore processors," in *1st International Qemu Users' Forum*, pp. 29–30, 2011.

[213] Xiaoxia Wu, Jian Li, Lixin Zhang, E. Speight, and Yuan Xie, "Power and performance of read-write aware Hybrid Caches with non-volatile memories," in *2009 Design, Automation & Test in Europe Conference & Exhibition*, pp. 737–742, IEEE, 4 2009.

[214] C. W. Smullen, V. Mohan, A. Nigam, S. Gurumurthi, and M. R. Stan, "Relaxing Non-volatility for Fast and Energy-efficient STT-RAM Caches," in *Proceedings of 17th International Symposium on High Performance Computer Architecture (HPCA)*, pp. 50–61, 2011.

[215] S. H. Choi, B. C. Paul, and K. Roy, "Novel Sizing Algorithm for Yield Improvement Under Process Variation in Nanometer Technology," in *Proceedings of the 41st Annual Design Automation Conference*, DAC '04, (New York, NY, USA), pp. 454–459, ACM, 2004.

[216] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, "The gem5 Simulator,"

[217] A. Percey, "Advantages of the Virtex-5 FPGA 6-Input LUT Architecture," 2007.

[218] J. Kim, A. Chen, B. Behin-Aein, S. Kumar, J.-P. 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)*, pp. 1–4, IEEE, 9 2015.

[219] K. Y. Camsari, S. Ganguly, and S. Datta, "Modular approach to spintronics," *Scientific reports*, vol. 5, 2015.

[220] T. Sundstrom, B. Murmann, and C. Svensson, "Power Dissipation Bounds for High-Speed Nyquist Analog-to-Digital Converters," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 56, no. 3, pp. 509–518, 2009.

[221] B. Murmann, "A/D converter trends: Power dissipation, scaling and digitally assisted architectures," in *2008 IEEE Custom Integrated Circuits Conference*, pp. 105–112, IEEE, 2008.

[222] A. S. U. (ASU), "14nm HP-FinFET Predictive Technology Model (PTM), accessed on 26 November 2018, available at: http://ptm.asu.edu/." http://ptm.asu.edu/.

[223] M. Osama, L. Gaber, and A. Hussein, "Design of high performance Pseudorandom Clock Generator for compressive sampling applications," in *2016 33rd National Radio Science Conference (NRSC)*, pp. 257–265, IEEE, 2 2016.

[224] R. Z. Bhatti, K. M. Chugg, and J. Draper, "Standard cell based pseudo-random clock generator for statistical random sampling of digital signals," in *2007 50th Midwest Symposium on Circuits and Systems*, pp. 1110–1113, IEEE, 8 2007.

[225] A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of CMOS device performance from 180 nm to 7 nm," *Integration*, vol. 58, pp. 74–81, 6 2017.

[226] S. J. Noh, Y. Miyamoto, M. Okuda, N. Hayashi, and Y. Keun Kim, "Effects of notch shape on the magnetic domain wall motion in nanowires with in-plane or perpendicular magnetic anisotropy," *Journal of Applied Physics*, vol. 111, p. 07D123, 4 2012.

[227] Chang-Joon Park, H. M. Geddada, A. I. Karsilayan, J. Silva-Martinez, and M. Onabajo, "A current-mode flash ADC for low-power continuous-time sigma delta modulators," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, pp. 141–144, IEEE, 5 2013.

[228] Z. He and D. Fan, "A Low Power Current-Mode Flash ADC with Spin Hall Effect based Multi-Threshold Comparator," in *Proceedings of the 2016 International Symposium on Low Power Electronics and Design*, (New York, New York, USA), pp. 314–319, ACM, 2016.

[229] S. Matsunaga, J. Hayakawa, S. Ikeda, K. Miura, H. Hasegawa, T. Endoh, H. Ohno, and T. Hanyu, "Fabrication of a nonvolatile full adder based on logic-in-memory architecture using magnetic tunnel junctions," *Applied Physics Express*, vol. 1, p. 91301, 8 2008.

[230] E. Deng, Y. Zhang, W. Kang, B. Dieny, J.-O. Klein, G. Prenat, and W. Zhao, "Synchronous 8-bit Non-Volatile Full-Adder based on Spin Transfer Torque Magnetic Tunnel Junction," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, pp. 1757–1765, 7 2015.

[231] D. Ravelsona, C. Chappert, J.-O. Klein, H.-P. Trinh, Y. Zhang, and W. Zhao, "Domain wall motion based magnetic adder," *Electronics Letters*, vol. 48, pp. 1049–1051, 8 2012.

[232] V. Bhatia, M. Goel, S. Gupta, P. Iswerya, N. Pandey, and A. Bhattacharyya, "Low power delay proficient current mode ADC design," in *2012 2nd International Conference on Power, Control and Embedded Systems*, pp. 1–4, IEEE, 12 2012.

[233] A. Zaeemzadeh, M. Joneidi, N. Rahnavard, and G. J. Qi, "Co-SpOT: Cooperative Spectrum Opportunity Detection Using Bayesian Clustering in Spectrum-Heterogeneous Cognitive Radio Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, pp. 206–219, 6 2018.

[234] N. Karim, A. Zaeemzadeh, and N. Rahnavard, "Rl-Ncs: Reinforcement Learning Based Data-Driven Approach for Nonuniform Compressed Sensing," in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, vol. 2019-October, IEEE Computer Society, 10 2019.

[235] H. Tan and R. F. DeMara, "A multilayer framework supporting autonomous run-time partial reconfiguration," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 16, pp. 504–516, 5 2008.

[236] 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, 2013.

[237] M. G. Parris, C. A. Sharma, and R. F. DeMara, "Progress in autonomous fault recovery of Field Programmable Gate Arrays," *ACM Computing Surveys*, vol. 43, pp. 1–30, 10 2011.

[238] R. F. DeMara and K. Zhang, "Autonomous FPGA fault handling through competitive runtime reconfiguration," in *Proceedings - NASA/DoD Conference on Evolvable Hardware, EH*, vol. 2005, pp. 109–116, 2005.

[239] R. DeMara, "Runtime-Competitive Fault Handling for Reconfigurable Logic Devices," *UCF Patents*, 6 2008.

[240] R. F. DeMara, K. Zhang, and C. A. Sharma, "Autonomic fault-handling and refurbishment using throughput-driven assessment," in *Applied Soft Computing Journal*, vol. 11, pp. 1588–1599, Elsevier, 3 2011.

[241] J. Lohn, G. Larchev, and R. De Mara, "Evolutionary fault recovery in a Virtex FPGA using a representation that incorporates routing," in *Proceedings - International Parallel and Distributed Processing Symposium, IPDPS 2003*, Institute of Electrical and Electronics Engineers Inc., 2003.

[242] K. Zhang, R. F. DeMara, and C. A. Sharma, "Consensus-based evaluation for fault isolation and on-line evolutionary regeneration," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3637 LNCS, pp. 12–24, Springer, Berlin, Heidelberg, 2005.

[243] S. C. Smith, R. F. DeMara, J. S. Yuan, D. Ferguson, and D. Lamb, "Optimization of NULL convention self-timed circuits," *Integration, the VLSI Journal*, vol. 37, pp. 135–165, 8 2004.

[244] 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, pp. 410–422, 3 2010.

[245] K. Zhang, G. Bedette, and R. F. DeMara, "Triple Modular Redundancy with Standby (TMRSB) supporting dynamic resource reconfiguration," in *AUTOTESTCON (Proceedings)*, pp. 690–696, Institute of Electrical and Electronics Engineers Inc., 2006.

[246] R. Al-Haddad, R. Oreifej, R. A. Ashraf, and R. F. DeMara, "Sustainable modular adaptive redundancy technique emphasizing partial reconfiguration for reduced power consumption," *International Journal of Reconfigurable Computing*, 2011.

[247] J. Huang, M. Parris, J. Lee, and R. F. DeMara, "Scalable FPGA-based architecture for DCT computation using dynamic partial reconfiguration," *Transactions on Embedded Computing Systems*, vol. 9, p. 9, 10 2009.

[248] J. Lohn, G. Larchev, and R. DeMara, "A genetic representation for evolutionary fault recovery in virtex FPGAs," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2606, pp. 47–56, Springer Verlag, 2003.

[249] R. S. Oreifej, R. N. Al-Haddad, H. Tan, and R. F. DeMara, "Layered approach to intrinsic evolvable hardware using direct bitstream manipulation of Virtex II Pro devices," in *Proceedings - 2007 International Conference on Field Programmable Logic and Applications, FPL*, pp. 299–304, 2007.

[250] R. B. Wunderlich, F. Adil, and P. Hasler, "Floating gate-based field programmable mixed-signal array," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 8, pp. 1496–1505, 2013.

[251] R. F. DeMara and D. I. Moldovan, "The Snap-1 Parallel AI Prototype," *IEEE Transactions on Parallel and Distributed Systems*, vol. 4, no. 8, pp. 841–854, 1993.

[252] S. Sheikhfaal, S. D. Pyle, S. Salehi, and R. F. Demara, "An Ultra-Low Power Spintronic Stochastic Spiking Neuron with Self-Adaptive Discrete Sampling," in *Midwest Symposium on Circuits and Systems*, vol. 2019-August, pp. 49–52, Institute of Electrical and Electronics Engineers Inc., 8 2019.

[253] A. Tatulian, S. Salehi, and R. F. DeMara, "Mixed-Signal Spin/Charge Reconfigurable Array for Energy-Aware Compressive Signal Processing," in *2019 International Conference on Reconfigurable Computing and FPGAs, ReConFig 2019*, Institute of Electrical and Electronics Engineers Inc., 12 2019.

[254] R. S. Oreifej and R. F. DeMara, "Intrinsic evolvable hardware platform for digital circuit design and repair using genetic algorithms," *Applied Soft Computing Journal*, vol. 12, pp. 2470–2480, 8 2012.

[255] T. A. Oghaz, E. C. Mutlu, J. Jasser, N. Yousefi, and I. Garibay, "Probabilistic Model of Narratives Over Topical Trends in Social Media: A Discrete Time Model," 4 2020.

[256] M. Sedghi, G. Atia, and M. Georgiopoulos, "Robust manifold learning via conformity pursuit," *IEEE Signal Processing Letters*, vol. 26, pp. 425–429, 3 2019.

[257] M. Sedghi, G. Atia, and M. Georgiopoulos, "Low-Dimensional Decomposition of Manifolds in Presence of Outliers," in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, vol. 2019-October, IEEE Computer Society, 10 2019.

[258] M. Sedghi, G. Atia, and M. Georgiopoulos, "Kernel Coherence Pursuit: A Manifold Learning-based Outlier Detection Technique," in *Conference Record - Asilomar Conference on Signals, Systems and Computers*, vol. 2018-October, pp. 2017–2021, IEEE Computer Society, 2 2019.