# TOWARDS ENERGY-EFFICIENT AND RELIABLE COMPUTING: FROM HIGHLY-SCALED CMOS DEVICES TO RESISTIVE MEMORIES

by

SOHEIL SALEHI MOBARAKEH
B.S. Isfahan University of Technology, 2014

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Electrical Engineering & Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2016

Major Professor: Ronald F. DeMara

# ABSTRACT

The continuous increase in transistor density based on Moore's Law has led us to highly scaled Complementary Metal-Oxide Semiconductor (CMOS) technologies. These transistor-based process technologies offer improved density as well as a reduction in nominal supply voltage. An analysis regarding different aspects of 45nm and 15nm technologies, such as power consumption and cell area to compare these two technologies is proposed on an IEEE 754 Single Precision Floating-Point Unit implementation. Based on the results, using the 15nm technology offers 4-times less energy and 3-fold smaller footprint. New challenges also arise, such as relative proportion of leakage power in standby mode that can be addressed by post-CMOS technologies.

Spin-Transfer Torque 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. Cost and suitability metrics assessed include the area of nanomagmetic and CMOS components per bit, access time and complexity, Sense Margin (SM), and energy or power consumption costs versus resiliency benefits. In an attempt to further improve the Process Variation (PV) immunity of the Sense Amplifiers (SAs), a new SA has been introduced called Adaptive Sense Amplifier (ASA). ASA can benefit from low Bit Error Rate (BER) and low Energy Delay Product (EDP) by combining the properties of two of the commonly used SAs, Pre-Charge Sense Amplifier (PCSA) and Separated Pre-Charge Sense Amplifier (SPCSA). ASA can operate in either PCSA or SPCSA

mode based on the requirements of the circuit such as energy efficiency or reliability. Then, ASA is utilized to propose a novel approach to actually leverage the PV in Non-Volatile Memory (NVM) arrays using Self-Organized Sub-bank (SOS) design. SOS engages the preferred SA alternative based on the intrinsic as-built behavior of the resistive sensing timing margin to reduce the latency and power consumption while maintaining acceptable access time.

Dedicated to my dear Parents and brother who have always supported me and never stopped believing in me.

تقدیم به پدر ، مادر و برادر عزیزم برای حمایت‌های بی‌دریغشان و باور من و توانایی‌های من.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER ONE: INTRODUCTION

## Need for Nanoscale Computing Approaches

The continuous increase in transistor density based on Moore's Law as shown in Figure 1, has led us to Complementary Metal-Oxide Semiconductor (CMOS) technologies beyond 20nm process node. These highly-scaled process technologies offer improved density as well as a reduction in nominal supply voltage. New challenges also arise, such as relative proportion of leakage power in standby mode. Power density and area have always been two important challenges for CMOS devices and designers [1]. As the trends enabled by Moore's Law allow the technology to shrink to enable increased level of integration, both benefits and challenges arise [2]. One of the most promising device technologies for extending Moore's law to 20nm and beyond is the self-aligned double-gate MOSFET structure (FinFET). FinFET transistors offer solutions to conventional planar CMOS issues such as sub-threshold leakage, poor short-channel electrostatic behavior, and high device variability. Furthermore, its ability to operate at much lower supply voltage results in static and dynamic power savings [3]. Although issues such as Process Variation (PV) [4, 5], aging, and bias temperature and threshold voltage instability [6-9] can become more significant at higher levels of integration, the capability of computing devices is greatly increased while their cost is decreased. In particular, by scaling down the transistor size it is possible to reduce the overall footprint of the device and accommodate a lower supply voltage to obtain a better dynamic power profile. While technology scaling enables increased density for memory cells, the intrinsic high leakage power of CMOS technology and the demand for reduced energy consumption inspires the use of emerging technology alternatives as Non-Volatile Memory

(NVM) including Spin-Transfer Torque Random Access Memory (STT-MRAM), Phase Change Memory (PCM), and Resistive Random Access Memory (RRAM).

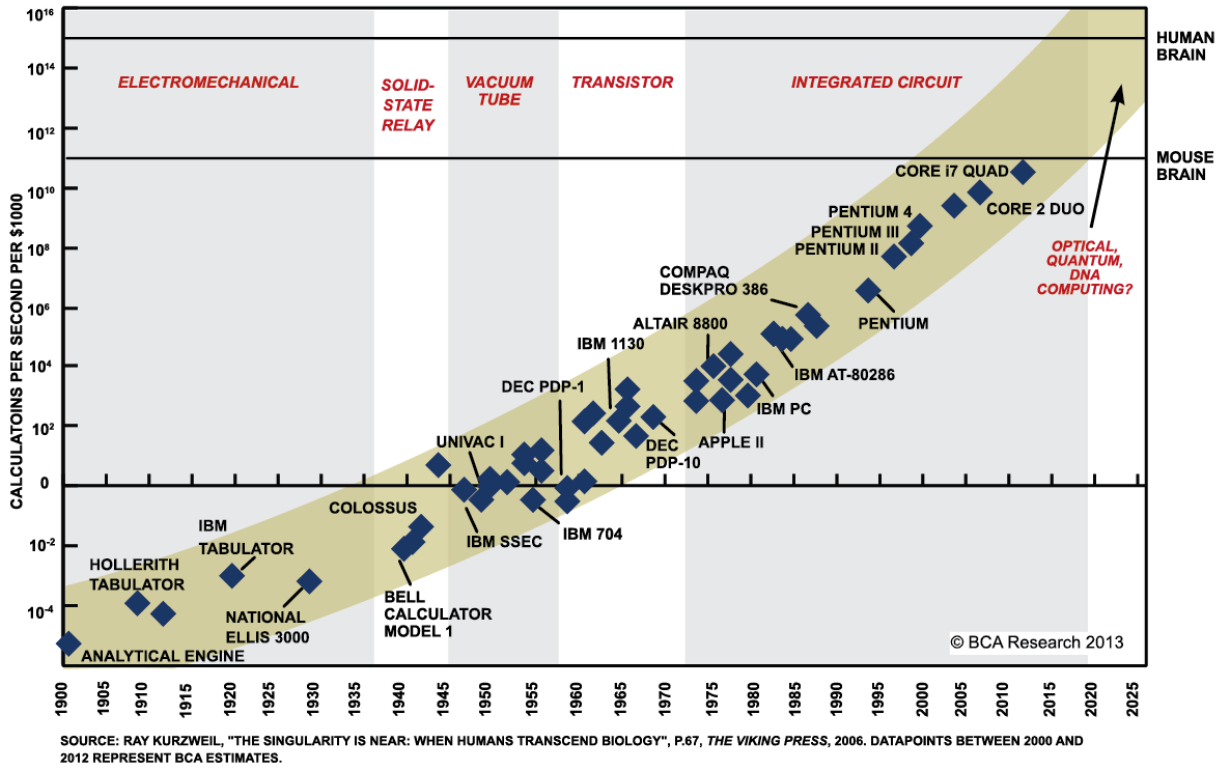



Figure 1: Moore's Law [10].

## Characteristics of post-CMOS Circuits

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 as shown in Figure 2. 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 in [11]. Cost and suitability metrics assessed include the area of

nanomagmetic and CMOS components per bit, access time and complexity, Sense Margin (SM), and energy or power consumption costs versus resiliency benefits. Solutions to the reliability issues identified are addressed within a taxonomy created to categorize the current and future approaches to reliable STT-MRAM designs. A variety of destructive and non-destructive sensing schemes are assessed for PV tolerance, read disturbance reduction, SM, and write polarization asymmetry compensation in [11].

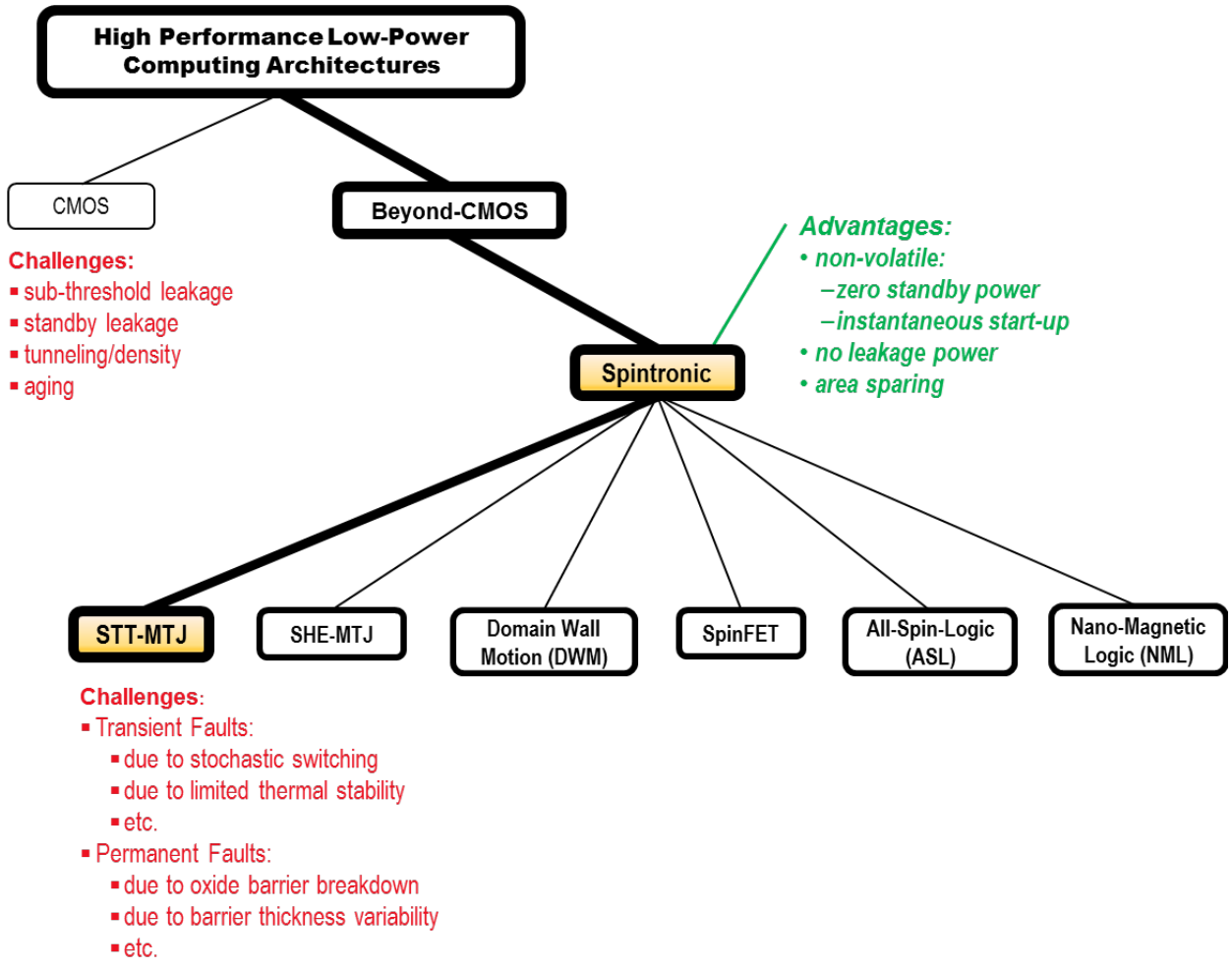



Figure 2: Taxonomy of Nanocomputing Architectures highlighting advantages of Spintronic devices.

As mentioned earlier, the intrinsic high leakage power of CMOS technology and the demand for reduced energy consumption inspires the use of emerging technology alternatives such as NVM including STT-MRAM, PCM, and RRAM. However, their narrow resistive SMs exacerbate the impact of PV in high-density NVM arrays, including on-chip cache and primary memory. Large-latency and power-hungry Sense Amplifiers (SAs) have been adapted to combat PV in the past.

## Design Tool Overview

There are several steps into the process of designing and implementing a hardware circuit. These steps are organized into two categories: Front End (FE) and Back End (BE) process as depicted in Figure 3. In FE process, designers propose the schematic and netlist of the circuit and run the simulations to collect theoretical results to prove their hypothesis. In each simulation step if the requirements and constraints of the circuit are not met, then designers will reiterate and modify their design and redo the simulation process until they achieve satisfactory results. If all the specifications of theoretical hypothesis are met, then designers will move on to the BE process phase which is more about the physical or layout design of the circuit.

In this work to implement the circuit schematics, Design Compiler and HSPICE are used as FE process softwares. Design Compiler enables the designer to extract a schematic view of the design based on the desirable technology library and constraints. In addition to the schematic view, Design Compiler provides information about power and energy consumption and timing of the circuit, which can help with the next step in the process, which is HSPICE simulation. Design Compiler's design flow is shown in Figure 4. HSPICE is used to simulate the circuit level netlist

of the design and extract waveforms and statistical analysis results. Using Monte Carlo Simulation method, a detailed statistical analysis can be performed on the circuit to check every critical corner case for the design in order to further optimize the circuit and analyze the design thoroughly. HSPICE modeling and Monte Carlo Statistical Analysis Flow is illustrated in Figure 5. After evaluating and validating the FE aspect of the design, we move on to the BE process by designing the layout for the circuit using Cadence Virtuoso as shown in Figure 6 or other layout designing tools with same characteristics. Physical design includes the layout design and also simulations and modeling. Circuit designs can also be modeled using Verilog-A language which can later reiterate with the HSPICE software if further improvements are required to be made on the circuit based on the outcome of the layout design. The hardware design flow iterates between FE and BE process phases until all the specifications and constraints are met as shown in Figure 3.

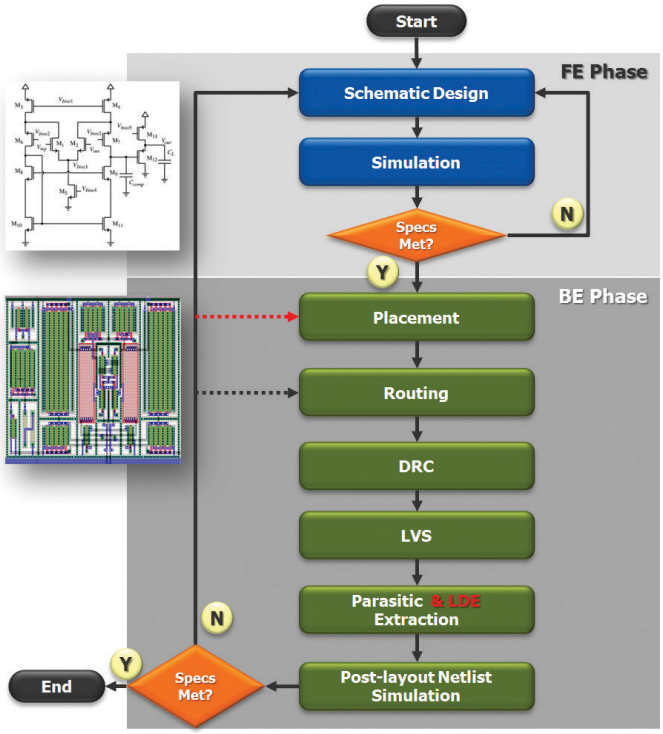



Figure 3: Hardware Design Flow for Front End (FE) and Back End (BE) designers [12].

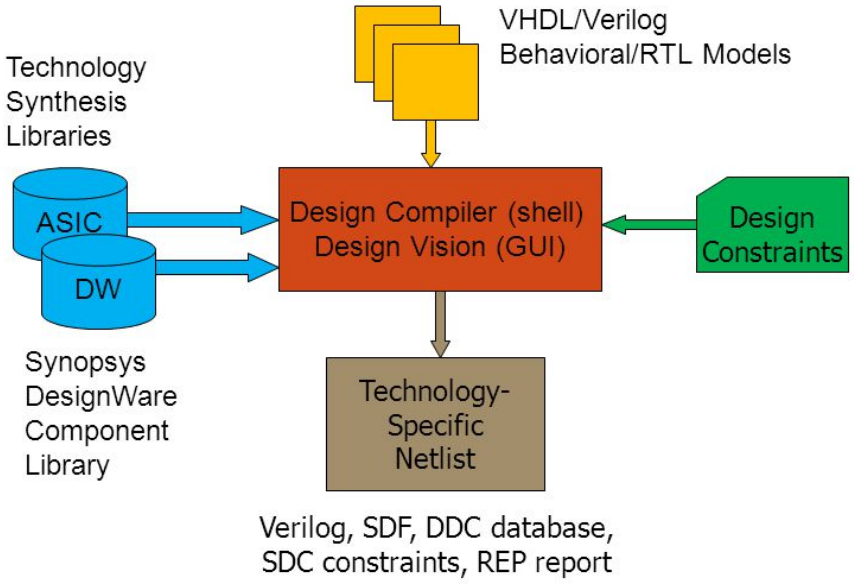



Figure 4: Automated Synthesis with Synopsys Design Compiler [13].

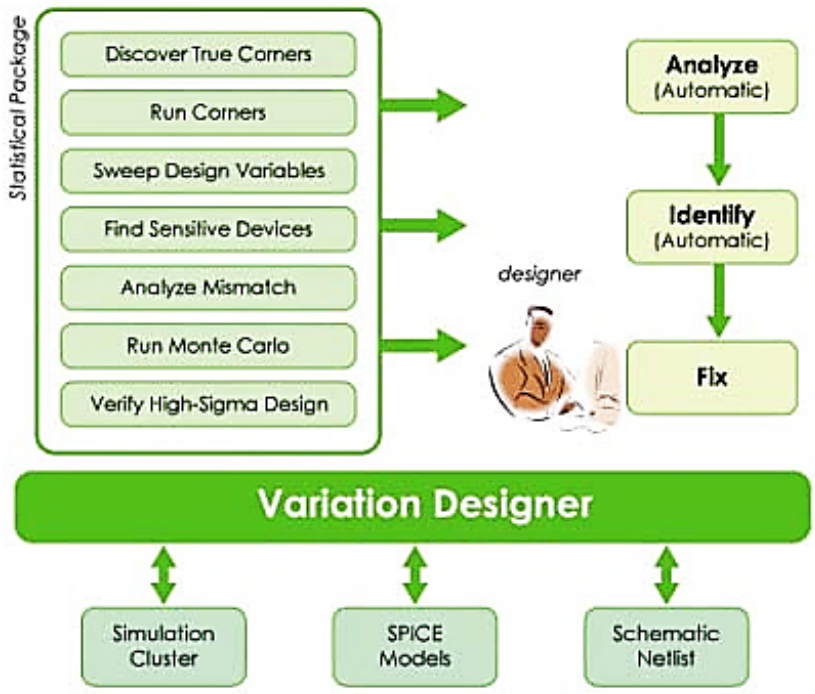



Figure 5: HSPICE modeling and Monte Carlo Statistical Analysis [14].

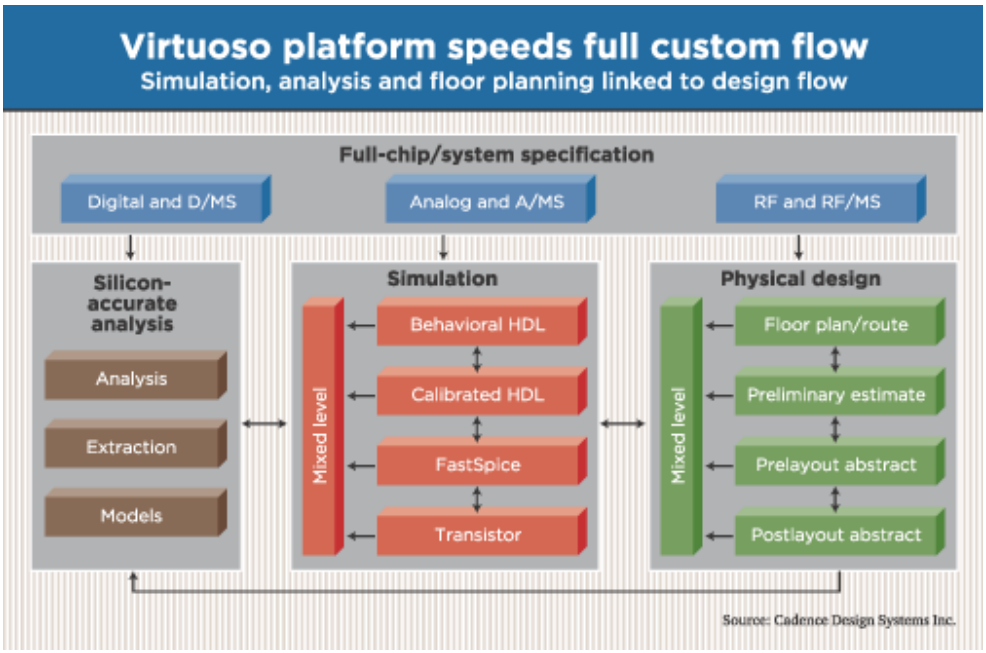



Figure 6: Cadence Virtuoso Design Space [15].

Contributions, Summary, and Organization of the Thesis: Towards Instrinsic Computation with Post CMOS Device

While spintronic-based neuromorphic architectures offer analog computation strategies [16], in this section we exploit the opportunity for reconfigurability and associative processing [17] using a Logic-In-Memory (LIM) paradigm. LIM is compatible with conventional computing algorithms and integrates logical operations with data storage, making it an ideal choice for parallel Single Instruction Multiple Data (SIMD) operations to eliminate frequent accesses to memory, which are extreme contributors to energy consumption. Spin-based LIM architectures have the capability to increase computational throughput, reduce the die area, provide instant-on functionality, and reduce static power consumption [18]. Feasibility of a low power spintronic LIM chip has recently been demonstrated in [19] for database applications.

As shown in Figure 7, in order to facilitate a variety of highly data parallel applications [20] such as Image Processing, Weather Forecasting, Big Data Analysis, and Physics Simulations, a variety of techniques have been investigated over the past two decades [21-25]. For graphics intensive applications such as real-time rendering [26], we need a novel reconfigurable fabric succeeding FPGAs to allow unprecedented gains in nanocomputation to realize signal processing [27]. Specifically, 1) energy-efficient associative computing paradigms and 2) Spintronic-based LIM reconfigurable fabric.

Unlike fixed pre-determined computing architectures that have recently been researched, a more effective approach is to realize the entire spectrum of applications by designing a Reconfigurable Spintronic Fabric (RSF). As shown in Figure 8, the RSF is a 2D array of Configurable Logic In Memory Blocks (CLIMBs) comprised of an array of Magnetic Tunnel

Junction (MTJ)-based LIM cells. The use of reconfiguration to address challenges of applications with low energy budgets while maintaining availability and resilience have been developed in recent years [27-29]. Figure 8 shows a tentative computing architecture that provides the appropriate platform for ultra-low power data-intensive processing applications. The core populates the RSF CLIMB cells with application data as well as writes the CLIMBs instruction memory with appropriate associate computing programs to perform the desired application. Only the final output data needs to be transmitted to the core. Figure 9 summarizes the features and advantages of the proposed computational system.

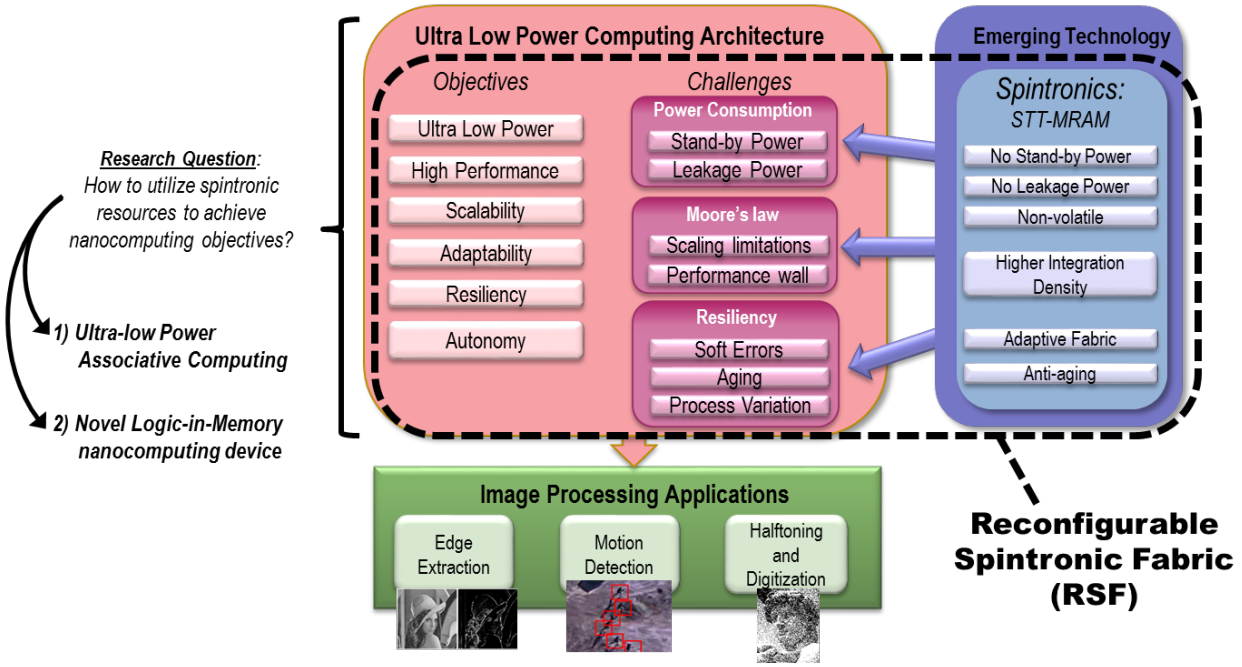



Figure 7: Proposed Research on Non-Conventional Ultra Low Power Computing Architectures.

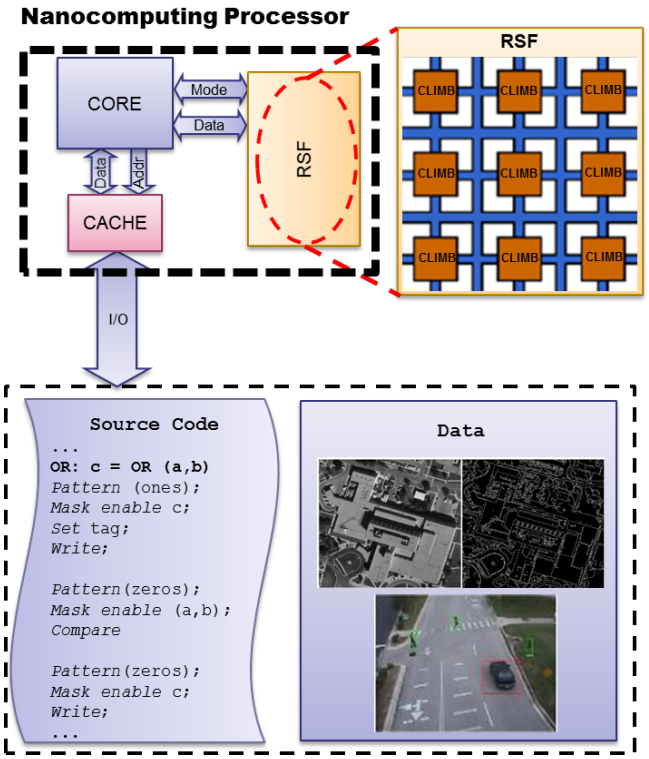



Figure 8: System Hierarchy of Nanocomputing Architecture: RSF and CLIMB.

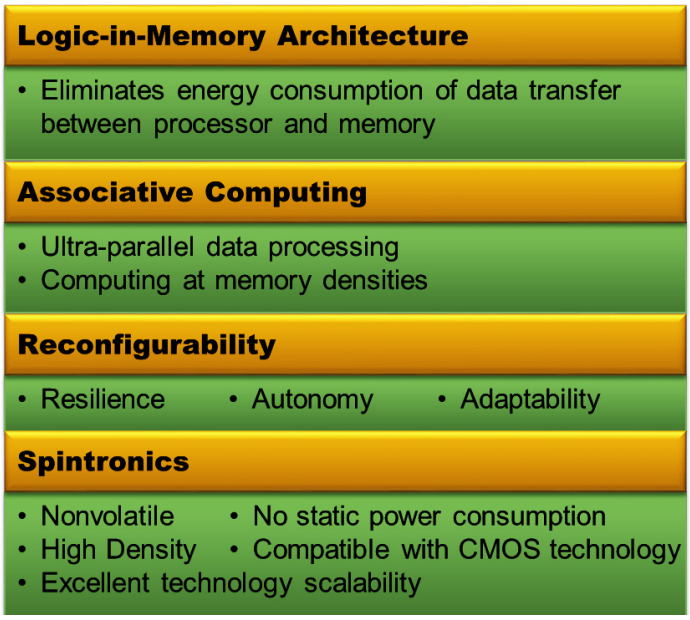



Figure 9: Features and Advantages of Reconfigurable Spintronic Fabric.

This thesis is organized into five chapters. Figure 10 outlines the materials that each chapter covers. A detailed analysis to elaborate on energy efficient CMOS design is provided in is provided in Chapter 2. Furthermore, a comprehensive analysis of reliability challenges of STT-MRAM devices is delivered in Chapter 2. In Chapter 3, an energy efficient Floating Point Unit (FPU) architecture is implemented and discussed. Chapter 4 includes a novel SA design to improve reliability of STT-MRAM. In addition, a novel approach to leverage PV to improve performance of STT-MRAM is characterized in Chapter 4. This thesis then concludes in Chapter 5.

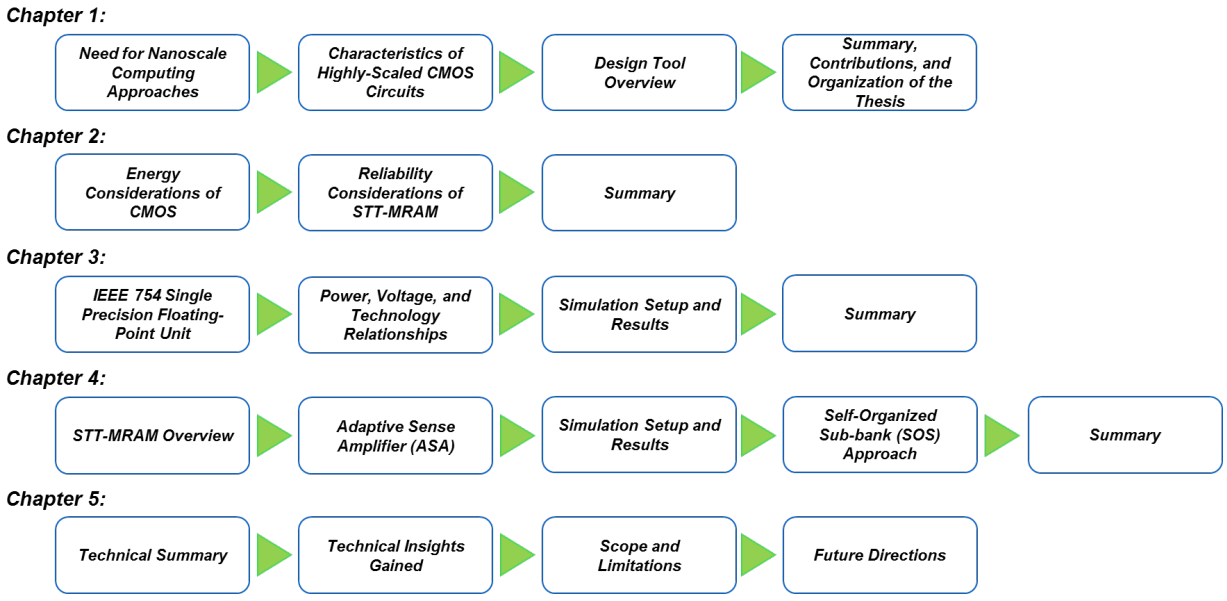



Figure 10: Organization of Thesis

# CHAPTER TWO: RELATED WORK

## Energy Considerations of CMOS

Several techniques are used to minimize the energy of CMOS logic devices for computation. Three main approaches are commonly used for energy reduction as shown in Figure 11. These three categories are 1) optimizing one or more steps of the computation procedure, 2) lowering the nominal supply voltage, and 3) allowing approximate arithmetic in applications that can tolerate reduced accuracy. In this Section, we concentrated on the lowering of nominal voltage, which can be realized through improvements in process technology. Alternate techniques of using a Near Threshold Voltage (NTV) operation are also possible, but introduce significant delay in the switching time in return for reduced energy.

These three techniques can also be synergistic. For example, [30] proposes the idea of minimizing the bit-width representation of floating-point utilizing low-resolution sensory data which results in 66% reduction in multiplier energy. In [31] a new method is proposed for improving the energy efficiency of a floating-point multiplier by partially truncating the computation of mantissa and also during different floating-point computations to allow the bit-width of mantissa in the multiplicand, multiplier, and output product to be dynamically interchangeable. Some voltage scaling techniques to reduce energy consumption are presented in [32]. In order to minimizing power consumption and energy of digital systems implemented in CMOS we can reduce the supply voltage to NTV which has an impact on logic speed and it has small performance penalties compared to operation in the sub-threshold region. Furthermore, [33] has discussed the benefits and challenges of NTV operation and its applications.

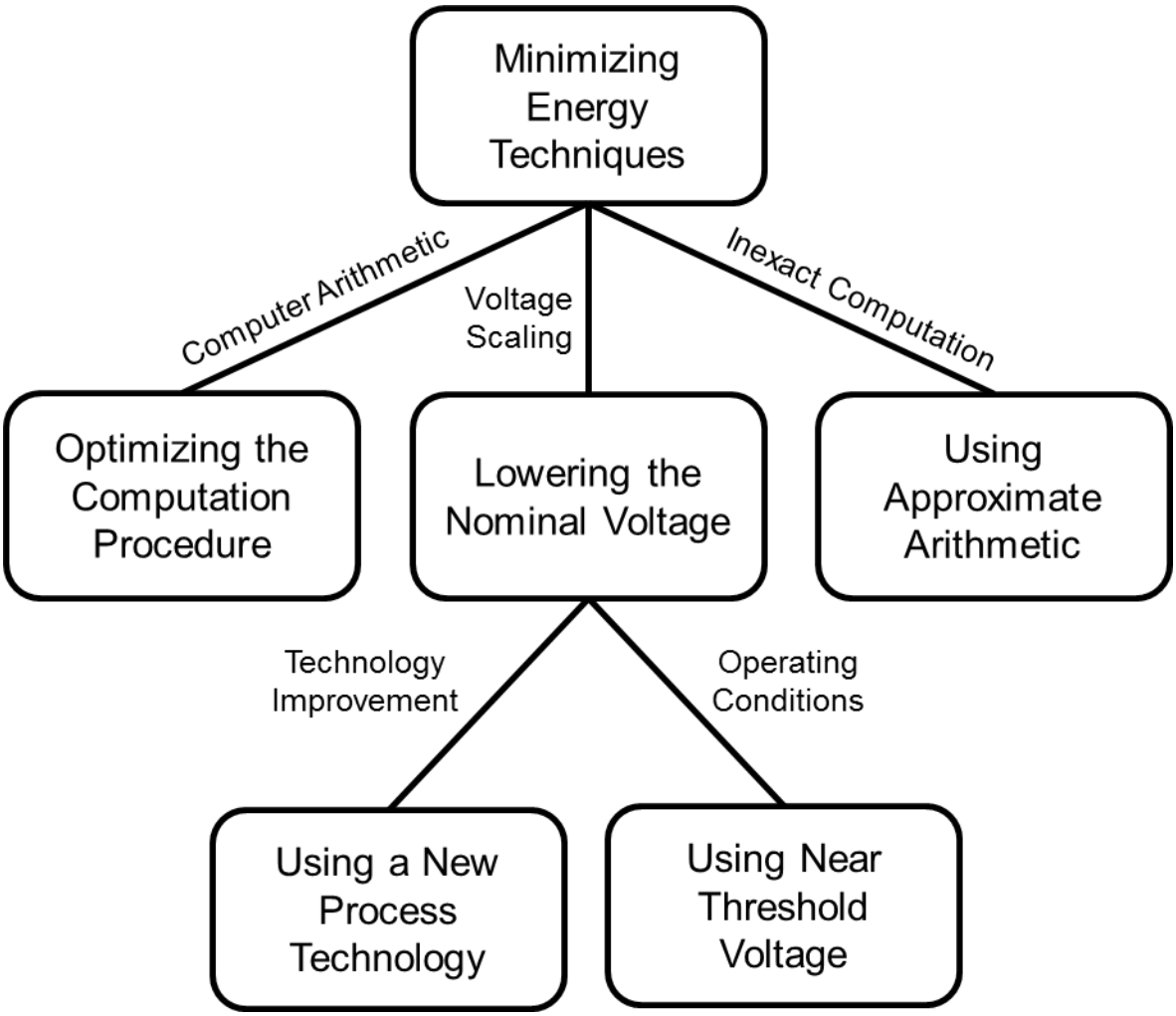



Figure 11: Energy Aware Techniques for CMOS Circuit Design.

Approximate computing is another concept that recently it has been used frequently in order to reduce the energy, power and area of CMOS devices. Using approximate computing in [34] results show reduction in energy and area. Further, using approximate or inexact computing can allow tradeoffs between energy, performance and area while introducing perceptually tolerable level of error for some applications [35, 36]. Using a new process technology is the most direct way to reduce the supply voltage without sacrificing speed and still results in increased energy

efficiency of CMOS switching devices. This technique has been discussed in [37], which analyzes a floating-point unit in 90nm, 45nm, and 22nm technologies. Furthermore, Swaminathan et al. in [38] investigated the switching time and energy consumption of a 32-bit CMOS full adder circuit in 15nm node where the authors created their own cell library. Other 15nm arithmetic designs are still emerging in the literature at this time. Main concept here is to use a new technology, which has a lower supply voltage and can make our circuit more efficient in terms of energy.

## Reliability Considerations of STT-MRAM

STT-MRAM has several advantages over other emerging memory technologies, however, it faces some distinct reliability challenges involving read and write failures [39, 40] as listed in this Section. STT-MRAM scalability is greatly influenced and limited due to thermal fluctuations and issues such as MTJ PV and the CMOS access transistor have had negative effects on STT-MRAM devices. In addition, as a result of these issues, demand for an advanced sensing circuit, which can provide required SM along with low power operation has been increased.

STT-MRAM bit errors can be significantly influenced due to PV [41], 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 ($V_t$), 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 [42], here we focus on the PV of the MTJ cell. The difference between the sensed bit-line voltage and the reference voltage which is known as the SM will be small due to the wide distribution of MTJ

resistance which can also result in a false detection scenario [43]. 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 PV-induced variability of the switching current [42].

Errors due to the STT-MRAM physical nature's failures will be categorized into transient faults and permanent faults as depicted in Figure 12. Transient faults, which can also be described as an incorrect signal condition [44], is mostly caused by the parameters of free layer such as current density ($J_c$) , and thermal stability factor denoted by $\Delta$. 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 [45].

In general, sensing schemes can be classified into two categories, Destructive and Non-Destructive [46]. Based on the definition presented in this research, Destructive Schemes are more vulnerable to read reliability failures. Non-Destructive Schemes are more tolerant to PV of reference cell, however, Destructive Schemes, typically, provide smaller read/sense latency.

### A. Transient Faults:

Transient Faults for STT-MRAM are divided into two categories: a) faults happening during the write operation and b) faults happening during the read operation.

#### a. Write Reliability Issues

Write Failures can happen due to stochastic nature of write process in STT-MRAM. During a write failure, an MTJ cell does not switch properly in order to

store the required value within write period. One of the possible solutions for this failure can be increasing the write duration. Another possible solution can be increasing write current. However, both of these solutions may cause significant amount of power dissipation and area overhead as well as speed degradation, which is not favorable.

Write Polarization Asymmetry is another issue for STT-MRAMs that can cause failures during the write operation. This concern is because switching and MTJ cell from Parallel (P) state to Anti-Parallel (AP) state needs higher switching current and suffers from more error rate compared to AP to P state switching. Possible solution to this concern can be utilization of Reversed MTJ Connection, resulting in a larger $I_{MTJ}$ for the P to AP switching which alleviates the effect of the critical current ($I_c$) asymmetry (($I_c$ (P→AP)/$I_c$ (AP→P)) > 1).

**b. Read Reliability Issues**

One of the main concerns during the read operation is Failures due to Read Disturbance. Since in STT-MRAMs Read and Write operation share the same path, unwanted bit-flip might happen during a read operation. This issue is becoming more significant in scaled technology nodes, since thermal stability factor Δ and critical switching current decrease. Possible solution to this concern can be increasing the margin between read and write currents by either increasing the write current or decreasing the read current. Increasing the write current may not be feasible since write current maintains a high value in STT-MRAM devices.

Also decreasing the read current will Increase the read latency and may result in another reliability issue called decision failure. Another solution would be using Error Correction Codes (ECC) and software methods.

Another issue that affects the read operation of STT-MRAM devices is Readability Degradation at Scaled Technology Node. This problem is **d**ue to reduction in switching current, which will become a greater concern in scaled technology node since reduction in switching current will limit the upper bound of sensing current. In order to address this concern, high read current is required to provide enough SM, and ensure reliable sensing by excluding the device variation of the sense amplifier, and maintain fast read and reduce read latency. Another way to deal with this issue is using low read current to prevent stored data from being upset.

Decision Failure is another problem that needs to be addressed. This issue happens when reading an MTJ cell, being unable to distinguish whether the stored bit is zero or one. Possible solution can be increasing the read duration or increasing read current.

Retention Failure problem would be another concern for STT-MRAM due to its 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 $\Delta$. Increasing $\Delta$ results in longer read duration, larger current amplitude and increase in number of bits per word during parallel reading.

## B. Permanent Faults

Oxide Barrier Breakdown can become problematic in STT-MRAM, causing permanent faults. Due to the fact that 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: $V=R\cdot A\times J_c$. In order to achieve better switching probability, large SM is required and in order to maintain high current density, reduced Resistance Area product ($R\cdot A$) is required. In addition, reducing thickness of the oxide or increasing the bias voltage (V) can help solving this issue. However, each of these solutions can result in the oxide barrier breakdown and shorten the MTJ lifetime.

Barrier Thickness Variability is another reliability concern for STT-MRAM devices. In order to maintaining low R.A value, favorably ultra-thin insulator or oxide barrier is required. MTJ's resistance is proportional to the oxide thickness exponentially. As a result, 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, SM will be upset by VR and permanent faults will occur.

To address both of these concerns we can use Modular Redundancy technique [47]. In addition, to prevent permanent faults, oxide thickness variation is required to be less than 5%. Furthermore, using a low bias voltage for sensing is suggested since the real TMR ratio decreases during the sensing operation.

Each of the solutions to the reliability problems of the STT-MRAM leverage different properties of the MTJ switching behavior. Recent preferred designs are able to achieve less than

5ns read sensing latency while maintaining wide SMs. Non-Destructive schemes have emphasized lower energy consumption as opposed to maintaining SMs. These 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. SA performance is seen to span in three different ranges across all proposed design strategies. The highest resiliency strategies deliver a SM above 300mV while incurring low power and energy consumption in the order of picojoules and microwatts, respectively, with read sense latency of a few nanoseconds down to hundreds of picoseconds for non-destructive and destructive sensing schemes, respectively. These criteria are also summarized in Figure 13 along with the approaches that correspond with each objective.

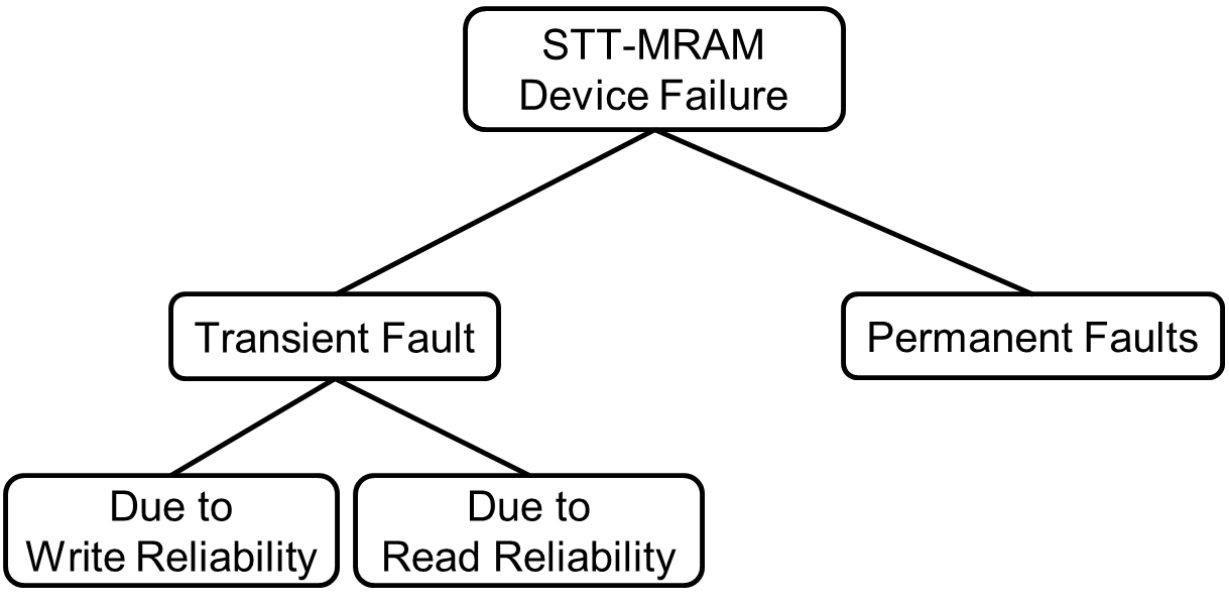
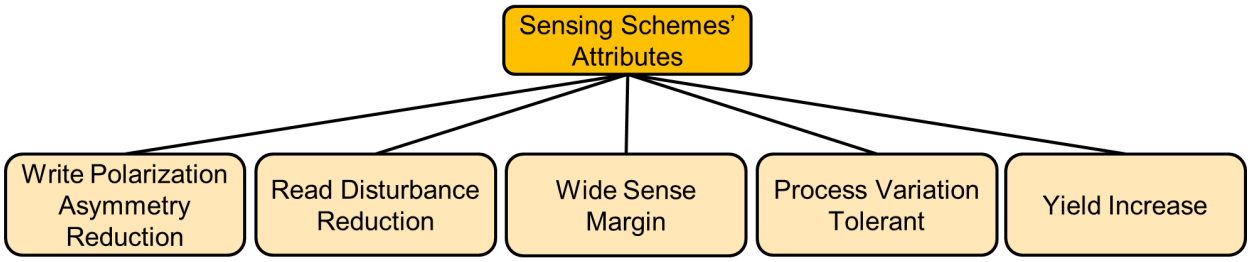



Figure 12: Sensing Reliability Challenges of STT-MRAM Devices Addressed with Techniques Developed Herein Outlined in Yellow Boxes




| Write Polarization Asymmetry Reduction | Read Disturbance Reduction | Wide Sense Margin | Process Variation Tolerant | Yield Increase |
|---|---|---|---|---|
| [42] | [39] | [57] | [42] | [57] |
| [43] | [49] | [45] | [43] | [52] |
| [48] | [50] | [42] | [46] | [67] |
| | [51] | [43] | [67] | |
| | [52] | [58] | [64] | |
| | [53] | [46] | [54] | |
| | [54] | [59] | [68] | |
| | [40] | [50] | [69] | |
| | [55] | [60] | [40] | |
| | [56] | [61] | [70] | |
| | | [62] | [66] | |
| | | [63] | | |
| | | [64] | | |
| | | [65] | | |
| | | [40] | | |
| | | [66] | | |

Figure 13: Sensing Schemes' Attributes Taxonomy

## Summary

In order to reduce the amount of power consumed in CMOS devices there are several methods explained within this thesis. Out of all the solutions, miniaturizing the device or in other words using technology nodes with smaller footprint is one of the most promising. On the other hand, due to scaling and power limitations of CMOS devices, researchers are exploring alternatives that offer better performance to overcome these limitations. As a result, STT-MRAM devices are considered a feasibly implemented solution for beyond CMOS technologies. Improving the reliability of STT-MRAM, however, is of great importance. Some notable inflection points between reliability and performance occur based on whether tolerating PV is of primary importance. In that case, destructive techniques such as [40, 42, 43, 46, 54, 64, 66-70] are recommended. On the other hand if tolerating read disturbance is a governing requirement then [39, 40, 49-56, 66] techniques are believed to be more promising. Furthermore, if wide SM is required techniques such as [40, 42, 43, 45, 46, 50, 53, 57, 58, 60-64, 66] could be a preferable, 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 [42, 43, 48] can be good candidates. Finally, if increasing the yield is the main goal then [52, 57, 67] techniques can be promising alternatives for conventional sensing schemes. These criteria are also summarized in Figure 13 along with the approaches that correspond with each objective.

# CHAPTER THREE: ENERGY AND AREA ANALYSIS OF A FLOATING-POINT UNIT IN 15NM CMOS PROCESS TECHNOLOGY

Floating-point computation can represent a large portion of the power consumed by the CPUs performing video processing and high performance scientific computation, and is a significant area component of most processors. The primary emphasis of this Chapter is to examine the use of 15nm technology process, which can allow a lower nominal supply voltage to reduce energy consumption and area. We are using 15nm technology [71] for IEEE-745 Floating-point Standard [72] in order to assess out about the relative advantages of the 15nm technology over 45nm technology [73]. In this Chapter, first we will introduce the design of the IEEE 754 floating-point unit that we used as a case study. Then, power, voltage and technology relationships are discussed. In addition, the simulation environment and the technology libraries used in our research are described followed by the experimental results are presented. We show realization for area reduction of about 3-fold and 3-times less energy consumption in 15nm technology [74].

## IEEE 754 Single Precision Floating-Point Unit

IEEE 754 is a standard for floating-point arithmetic, which is a well-known standard frequently used in processors. Details about the IEEE 754 standard can be found in [72], and we will utilize this standard as our case study. Numbers in this standard are represented using an exponent and a significand where the sign is represented using one bit. We can categorize floating-point numbers based on their exponent and based on their significand. Categories based on the exponent are basic and extended where if the floating point's significand is 32 bits long then it is single precision format and if it is 64 bits long then it is referred to as double precision format.

IEEE 754 standard supports different types of operation such as addition, subtraction, multiplication, comparisons, division, square root, remainder, and conversions between integer and floating-point formats. During the arithmetic operations, our result might be a Not-A-Number (NAN) if there is some overflow or underflow or a division by zero event, which all need to be handled as an exception. After every floating-point operation also needs the result to be rounded based on the format so that the result fits within the standard specifications mentioned in [72].

In this Chapter, we used a single precision Floating-Point Unit (FPU) [75] which is fully IEEE 754 compliant and it can perform a floating-point operation every cycle. It will latch internally the operation type, rounding mode, and operands. This FPU delivers the result after four clock cycles. This unit will only assert Signaling NAN (**`snan`**) if operand a (**`opa`**) or operand b (**`opb`**) signals NAN which in this case the output will be a quiet NAN (**`qnan`**). It uses two pre-normalization units, one for addition and subtraction and another for multiplication and division to adjust the exponents and mantissas and we have a post normalization block, which does the normalization of the output's fraction and then rounds the output. Finally, the result will be provided in single precision floating-point format. The FPU block diagram is shown in Figure 14.

## Power, Voltage, and Technology Relationships

Power calculation is an important metric for a CMOS device performance. Utilizing the power analysis, we can determine important factors such as power-supply sizing, current requirements, criteria for device selection, and the maximum reliable operating frequency. As shown in Eq. (1), total power of a CMOS device is determined by two main components, which are dynamic power and static power, respectively:

$$P_{Total} = P_{Dynamic} + P_{Static} \quad \text{Eq. (1)}$$

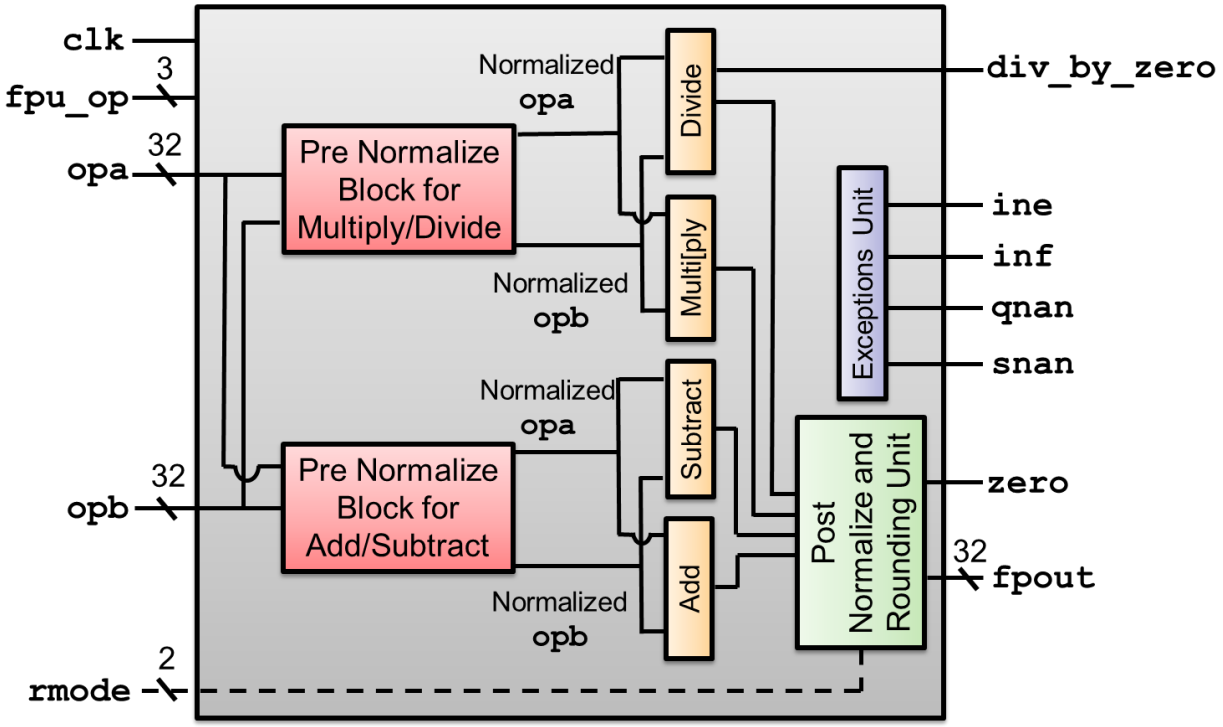



Figure 14: FPU Functional Elements.

CMOS static power consumption is a result of the leakage current while the transistor is off. In general, static power consumption is the product of the device leakage current and the supply voltage as shown in Eq. (2). However, dynamic power consumption can have a significant impact on the total power when the device's operating frequency is high. In addition to the high operating frequency, charging and discharging a capacitive load can also increase the dynamic power consumption. Dynamic power consists of two components 1) signal transitions power (transient power) and 2) short circuit power as shown in Eq. (3) where $P_T$ and $P_{SC}$ stand for transient power and short circuit power respectively.

$$P_{Static} = I_{Static} \times V_{dd} \quad \text{Eq. (2)}$$

$$P_{Dynamic} = P_T + P_{SC} \quad \text{Eq. (3)}$$

The dynamic power is the power consumed for legitimate logic transitions and spurious glitches due to switching which is a result of input transitions. The first component is the current required to charge the internal nodes called switching current, which is shown in Eq. (4). Second component is the current that flows from $V_{dd}$ to GND when the p-channel transistor and n-channel transistor simultaneously turn on briefly during the logic transition called through or short circuit current. The transient power and the short circuit power are given by the following equations:

$$P_T = E_T \times f_{clk} \times \alpha = C_L \times {V_{dd}}^2 \times f_{clk} \times \alpha \quad \text{Eq. (4)}$$

$$P_{SC} = E_{SC} \times f_{clk} \times \alpha \quad \text{Eq. (5)}$$

$$E_{SCf} = (t_f \times (V_{dd} - |V_{Tp}| - V_{Tn}) \times I_{scmaxf})/2 \quad \text{Eq. (6)}$$

$$E_{SCr} = (t_r \times (V_{dd} - |V_{Tp}| - V_{Tn}) \times I_{scmaxr})/2 \quad \text{Eq. (7)}$$

where $E_T$ is the transient energy, $E_{SC}$ is short circuit energy which is related to rise and fall times of the input signal, $E_{SCr}$ is rise time short circuit energy, $E_{SCf}$ is fall time short circuit energy, $V_{tp}$ and $V_{tn}$ are the threshold voltages of the p-channel and n-channel transistors respectively, $I_{SCmaxf}$ and $I_{SCmaxr}$ are maximum short circuit currents flowing during the fall time and rise time respectively, $f_{clk}$ is the operating frequency, α is switching activity factor, $C_L$ is the capacitive load and $V_{dd}$ is the supply voltage [76]. As it can be inferred from Eq. (5), Eq. (6), and Eq. (7), the duration of the short circuit current impulse is directly affected by operating frequency, rise and fall times, and the internal nodes of the device. The short circuit current that flows through the gate is negligible compared to the switching current, when the operating frequency is high.

## Simulation Environment

In order to compare the two technologies and to simulate the FPU design we used Design Compiler [77] which is an RTL Synthesis tool by Synopsys. We simulated the FPU circuit using the 45nm and 15nm libraries from NANGATE and extracted the results. In order to use the Design Compiler, first we have to express the hardware description of our circuit and then synthesize it to extract the gate-level netlist using the library components defined in technology library file for RTL synthesis. We used the Design Compiler in order to create the gate-level netlist for our FPU design. Figure 15 depicts the flow of a gate-level netlist extraction.

## Simulation Results

Simulating the FPU using Design Compiler, we could extract the information about the resources used in the design after RTL synthesis. Information about the gates that have been used for the FPU design are listed in Table 1, Table 2, and Table 3. The gates used for the design are all standard cells defined in the corresponding technology libraries. Due to technology scaling, the anticipation is that the cell area in 15nm technology would be significantly less than 45nm technology and after simulation, the results validated our hypothesis with specific area values. The Total Cell Area of the FPU in 15nm technology is about 30% less than that of the FPU in 45nm technology. Figure 16 depicts the graph for Cell Area analysis of the two technologies used for simulation.

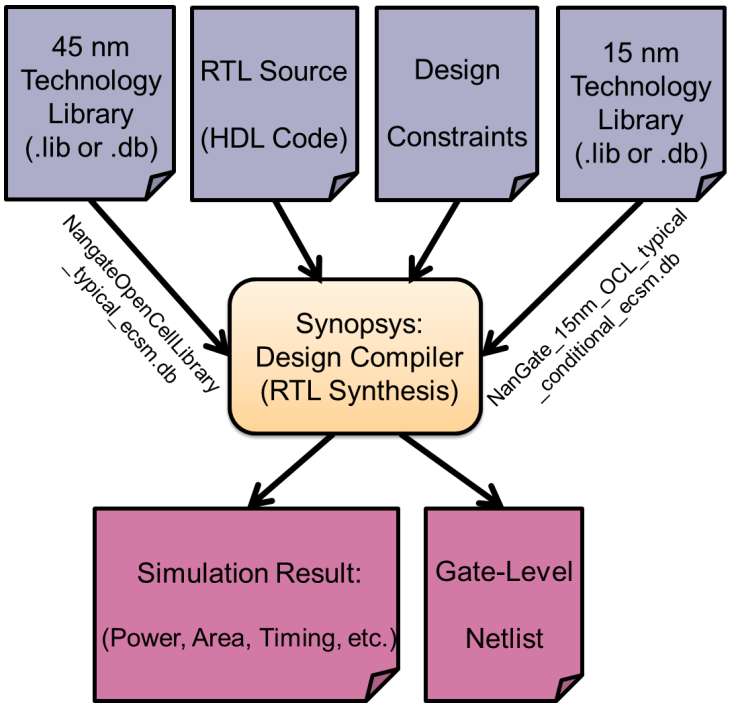



Figure 15: Modeling Environment and Synthesis Flow.

Identical HDL was synthesized using the same tool under identical synthesis parameters. As noted in Table 1, Table 2, and Table 3, library differences can result in some diversity between gate selection and gate count. Nonetheless, the predominant trend for energy consumption between the two designs is realistic for synthesis using two process technologies. Table 4lists power consumption estimates for the FPU using the default testbench inputs from Design Compiler. The 45nm column indicates power consumption for a zero-negative slack clock period of 5ns. These values are seen to be 3.15-fold to 4.56-fold larger than the same design synthesized using the 15nm with default parameters. The rightmost column indicates that the FPU design can also operate significantly faster in 15nm technology than in 45nm technology. It is observed that the minimum clock rate, which avoids negative timing slack, is 400ps. Thus, for the default testbench, the FPU in 15nm technology can operate about 12.5 times faster than the same FPU in 45nm technology,

albeit at a higher power consumption due to the faster clock. Finally, Figure 17 shows the components of energy consumption and total energy consumption for the FPU in 45nm and 15nm technologies. Results indicate that using 15nm technology allows the FPU to consume about four times less energy than 45nm technology.

Table 1: Constituent Gate Types and Usage Count (Simplex Gates).

| Gate Function | Simplex Gates | Quantity | | Gate Area ($\mu m^2$) | |
|---|---|---|---|---|---|
| | | *45nm* | *15nm* | *45nm* | *15nm* |
| AND | AND2_X1 | 38 | 19 | 1.0640 | 0.2949 |
| | AND3_X1 | 6 | 9 | 1.3300 | 0.3932 |
| | AND3_X2 | 0 | 2 | 1.5960 | 0.3932 |
| | AND4_X1 | 2 | 4 | 1.5960 | 0.4424 |
| | AND4_X2 | 0 | 1 | 1.8620 | 0.4915 |
| NAND | NAND2_X1 | 63 | 120 | 0.7980 | 0.1966 |
| | NAND2_X2 | 0 | 5 | 1.3300 | 0.2949 |
| | NAND3_X1 | 26 | 28 | 1.0640 | 0.2949 |
| | NAND3_X2 | 2 | 17 | 1.8620 | 0.4424 |
| | NAND4_X1 | 29 | 16 | 1.3300 | 0.3441 |
| | NAND4_X2 | 0 | 1 | 2.3940 | 0.5407 |
| OR | OR2_X1 | 4 | 7 | 1.0640 | 0.2949 |
| | OR3_X1 | 7 | 7 | 1.3300 | 0.3932 |
| | OR3_X2 | 9 | 1 | 1.5960 | 0.3932 |
| | OR4_X1 | 2 | 3 | 1.5960 | 0.4424 |
| | OR4_X2 | 0 | 2 | 2.3940 | 0.4915 |
| NOR | NOR2_X1 | 27 | 28 | 0.7980 | 0.1966 |
| | NOR2_X2 | 0 | 1 | 1.3300 | 0.2949 |
| | NOR3_X1 | 37 | 23 | 1.0640 | 0.2949 |
| | NOR4_X1 | 27 | 24 | 1.3300 | 0.3441 |
| | NOR4_X2 | 0 | 4 | 2.3940 | 0.5407 |
| XNOR | XNOR2_X1 | 2 | 0 | 1.5960 | 0.4424 |
| BUFFER | BUF_X1 | 26 | 0 | 0.7980 | 0.2458 |
| | BUF_X2 | 4 | 22 | 1.0640 | 0.2458 |
| | BUF_X4 | 0 | 1 | 1.8620 | 0.3932 |
| | BUF_X8 | 0 | 2 | 3.4580 | 0.6881 |
| | CLKBUF_X1 | 16 | 0 | 0.7980 | 0.2458 |
| | CLKBUF_X2 | 0 | 1 | 1.0640 | 0.2458 |
| | CLKBUF_X4 | 0 | 1 | N/A | 0.3932 |
| | CLKBUF_X8 | 0 | 1 | N/A | 0.6881 |
| | CLKBUF_X12 | 0 | 5 | N/A | 0.9830 |
| INV | INV_X1 | 237 | 173 | 0.5320 | 0.1475 |
| | INV_X2 | 4 | 8 | 0.7980 | 0.1966 |
| | INV_X4 | 0 | 7 | 1.3300 | 0.2949 |
| | INV_X8 | 6 | 2 | 2.3940 | 0.4915 |

Table 2: Constituent Gate Types and Usage Count (Complex Gates).

| Gate Function | Complex Gates | Quantity | | Gate Area (μm²) | |
|---|---|---|---|---|---|
| | | *45nm* | *15nm* | *45nm* | *15nm* |
| AND-OR-INV | AOI21_X1 | 5 | 31 | 1.0640 | 0.2949 |
| | AOI22_X1 | 59 | 91 | 1.3300 | 0.3441 |
| | AOI22_X2 | 0 | 11 | 2.3940 | 0.5898 |
| | AOI211_X1 | 3 | 0 | 1.3300 | N/A |
| | AOI221_X1 | 6 | 0 | 1.5960 | N/A |
| | AOI221_X4 | 2 | 0 | 3.4580 | N/A |
| | AOI222_X1 | 19 | 0 | 2.1280 | N/A |
| OR-AND-INV | OAI21_X1 | 8 | 83 | 1.0640 | 0.2949 |
| | OAI21_X2 | 0 | 10 | 1.8620 | 0.4424 |
| | OAI22_X1 | 4 | 8 | 1.3300 | 0.3440 |
| | OAI22_X2 | 0 | 10 | 2.3940 | 0.5898 |
| | OAI211_X1 | 3 | 0 | 1.3300 | N/A |
| | OAI221_X1 | 50 | 0 | 1.5960 | N/A |
| | OAI221_X4 | 15 | 0 | 3.4580 | N/A |

Table 3: Constituent Gate Types and Usage Count (Registers).

| Technology | Registers | Quantity | Area (μm²) |
|---|---|---|---|
| 45nm | DFF_X1 | 234 | 4.5220 |
| | DFF_X2 | 2 | 5.0540 |
| | SDFF_X1 | 24 | 6.1180 |
| 15nm | DFFSNQ_X1 | 256 | 1.2779 |
| | DFFRNQ_X1 | 4 | 1.2779 |

Table 4: Energy and Delay Analysis of the Floating Point Unit.

| | **45nm τ = 5ns** | **15nm τ = 5ns** | **15nm τ = 0.4ns** |
|---|---|---|---|
| Cell Internal Power (mW) | 1.2367 | 0.3922 | 4.8297 |
| Net Switching Power (mW) | 0.5863 | 0.1284 | 1.6604 |
| Total Dynamic Power (mW) | 1.8230 | 0.5206 | 6.4901 |
| Cell Leakage Power (mW) | 0.2250 | 0.1134 | 0.1215 |
| Total Power (mW) | 2.0480 | 0.6340 | 6.6116 |
| Clock Period (ns) | 5.0 | 5.0 | 0.4 |
| Global Operating Voltage (v) | 1.1 | 0.8 | 0.8 |

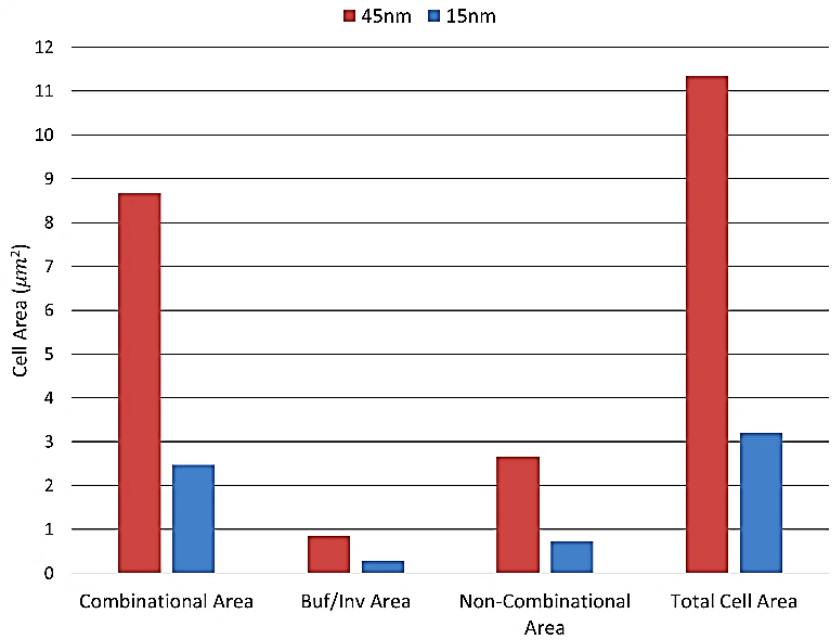



Figure 16: Comparison of Floating Point Unit Area in 45nm and 15nm.

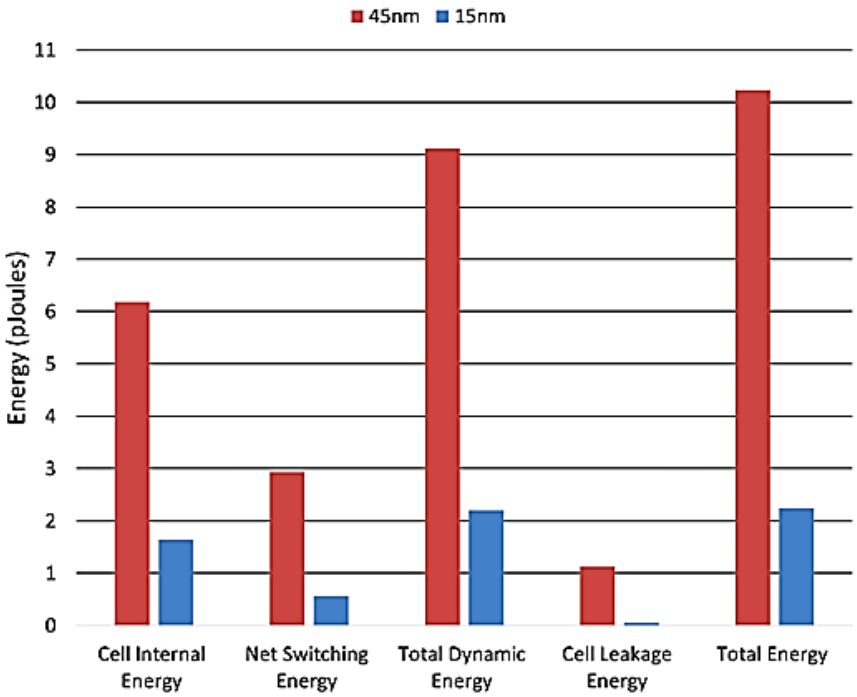



Figure 17: Comparison of Floating Point Unit Energy Consumption in 45nm and 15nm.

## Summary

Power density and area are two important challenges for CMOS devices. As discussed in this Chapter, using a new process technology is the most direct way to reduce the supply voltage which results in increased energy efficiency of CMOS switching devices without sacrificing speed. Results have proven that 15nm technology suggests 3-fold to 4-fold improvement energy efficiency than 45nm technology and it offers about 30% less cell area using this Predictive Technology Model.

Despite the fact that FinFET devices are one of the most promising alternative for planar CMOS, these devices may suffer from some reliability issues that need to be addressed. Self-heating is one of the problems that FinFET devices may face due to their complex geometry and confined dimensions. Self-heating can be a cause for electro-migration and other such issues because it decreases the reliability of the device. As the number of fins grow, self-heating impact will be increased; however, increase in the number of gates doesn't have any significant effect on self-heating [9]. Other important reliability issues, which can influence FinFET's performance and can affect the behavior of the device, are Negative Bias Temperature Instability (NBTI) aging and Positive Bias Temperature Instability (PBTI) aging [6-8]. These issues can result in an alteration in the $V_t$ of the device which is a function of three main factors: $V_{GS}$, temperature, and time. In long term use of the device, $V_t$ can undergo a significant degradation which influences the critical path's delay by as much as 7% to 10% [3]. PV is another reliability concern that needs to be taken into account. PV is a result of small geometries of FinFET devices and as the technology shrinks, their impact has become more significant. Generally, these variations are caused by factors such as random dopant fluctuations, line edge roughness, layout induced stress, and other PV which can

result in changes in $V_t$, power and timing [4]. Migrating to new device technologies such as 15nm can help reduce the energy due to reduction in supply voltage, however, as mentioned earlier, PV, aging, etc. can cause some reliability issues, which need to be solved and addressed at the 15nm technology node [78].

# CHAPTER FOUR: ADAPTIVE SENSE AMPLIFIER DESIGN FOR POST CMOS RESISTIVE NON-VOLATILE MEMORIES

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 [57], innovations to mitigate the power wall and reduce leakage power consumption occupy the forefront of on-chip memory design considerations. [39, 79]. Power consumed by memory elements can become a significant portion of total power in active modes whereby the processing cores rely on these memory arrays that are significant contributors to standby mode power consumption [80-82].

To attain these goals and deliver the necessary operational characteristics, emerging memory devices such as RRAM, PCM, and Magnetic RAM (MRAM) offer several potential advantages. Among promising devices, the 2014 ITRS Magnetism Roadmap identifies nanomagnetic devices such as STT-MRAM, as capable post-CMOS candidates, of which Nano Magnetic Logic (NML), Domain Wall Motion (DWM), STT-MTJ/Spin Hall Effect (SHE)-MTJ are considered feasibly-implemented [83-85]. 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.

In this Chapter, the primary focus is on reliability issues that may affect the performance of the STT-MRAM. First, an overview of STT-MRAM functionality and technology aspects is provided and then a novel SA design is proposed. Finally, an innovative circuit-architecture approach is proposed using the proposed SA.

## STT-MRAM Overview

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. STT switching is one of the most promising alternatives for data storage. MTJ device which consists of two ferromagnetic layers called the reference layer and the free layer and one tunnel barrier oxide layer, is used in STT-MRAM cells to store data as binary values 0 or 1 [57]. Figure 18A shows an STT-MRAM cell, which has an access transistor that connects the storage device, and the bit-line, which is the same as DRAM cell, however, an STT-MRAM cell, differs from DRAM since the other end of the storage device is connected to the sense line or source line instead of ground. This STT-MRAM cell structure is being known as “one-transistor-one-MTJ (1T-1R)” [86, 87].

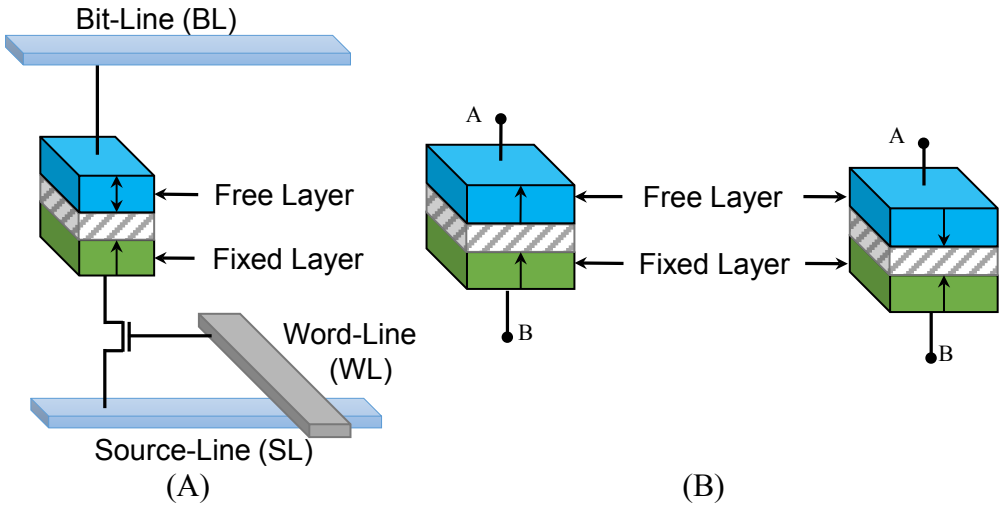



Figure 18: (A) 1T-1R STT-MRAM cell structure, (B) Right: Anti-Parallel (high resistance) state, Left: Parallel (low resistance) state.

As mentioned the stored data in an MTJ can be represented as the magnetic orientation of the free layer compared to fixed layer, which these two layers introduce a resistance that shows either 0 or 1. As shown in Figure 18B, identical magnetic orientation results in parallel configuration, which introduces a low resistance that can be represented as logical "0" and opposite magnetic orientation results in anti-parallel configuration, which introduces high resistance that can be represented as logical "1". Since in STT-MRAM, relative resistance values are used to determine the bit value, then reliable sensing schemes in order to read the stored data compared to DRAM in which the data is the charge stored in a capacitor, which is sensed using voltage sensing circuitry. In order to read the stored data, we need to apply a small voltage between source and bit-lines, and using a SA to sense the amount of current flow. Large write amplifiers are used since writing a data to an MTJ, requires significantly larger current than reading the stored value. In order to write into an MTJ cell, we need to apply a large current through the MTJ to change the magnetic orientation of its free layer to form a parallel or an anti-parallel configuration compared

to the fixed layer [87]. Maintaining lower critical current while having faster speed as well as large energy barrier for high-density device, researchers have proposed data storage using Perpendicular Magnetic Anisotropy (PMA) instead of conventional In-plane Magnetic Anisotropy (IMA) in order to achieve high anisotropy field ($H_k$),. This will result in better integration of STT-MRAM into logic circuits [45]. Detailed information about STT-MRAM circuit design and operation can be found in [57] and [80].

STT-MRAM writing has three different operating regions for switching which include Thermal Activation Region where the write current is less than eighty percent of the critical current, Dynamic Reversal Region wherein the write current is greater than eighty percent of the critical current and less than the critical current itself, and Precessional Region where the write current is greater than the critical current [86]. We normally operate this in either the Thermal Activation Region or the Precessional Region. If fast switching is required Precessional Region is a better option compared to Thermal Activation Region since the latter will cause slower switching.

STT-MRAM bit errors can be significantly influenced due to PV [41] 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 $V_t$, 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 [42], here we focus on the PV of the MTJ cell. SM, also known as The difference between the sensed bit-line voltage and the reference voltage, will be small due to the wide distribution of MTJ resistance which can also

result in a false detection scenario [43]. 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 PV-induced variability of the switching current [42].

## Adaptive Sense Amplifier (ASA)

Due to increase in PV as the technology shrinks, this element has become a major concern in high density memory arrays and cache designs [88]. In order to reduce the effects of device mismatch and variation due to scaling of the devices, different SAs were studied and among them Pre-Charge Sense Amplifier (PCSA) [58] and Separated Pre-Charge Sense Amplifier (SPCSA) [60] were chosen as promising solutions. PCSA has better sense latency and power consumption compared to SPCSA, however, it suffers from more error rate [60]. SPCSA, however, provides better reliability while having negligible increase in sense latency and power consumption and negligible area overhead compared to PCSA [60].

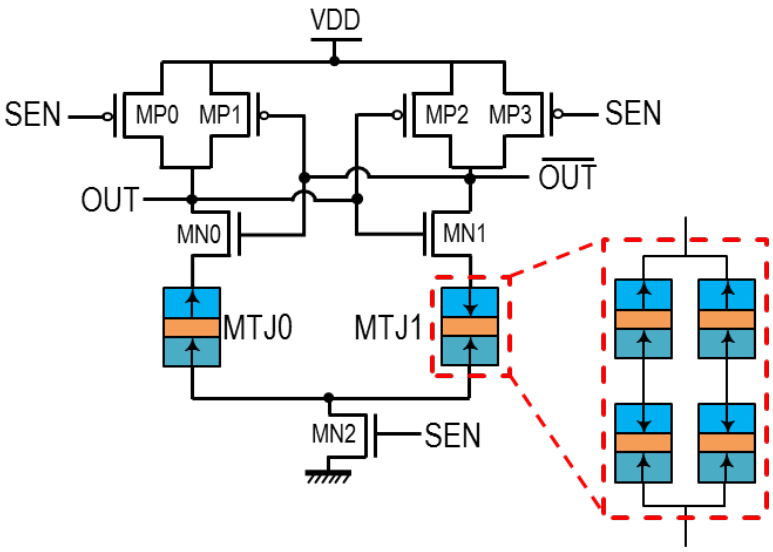



Figure 19: Pre-Charge Sense Amplifier (PCSA).

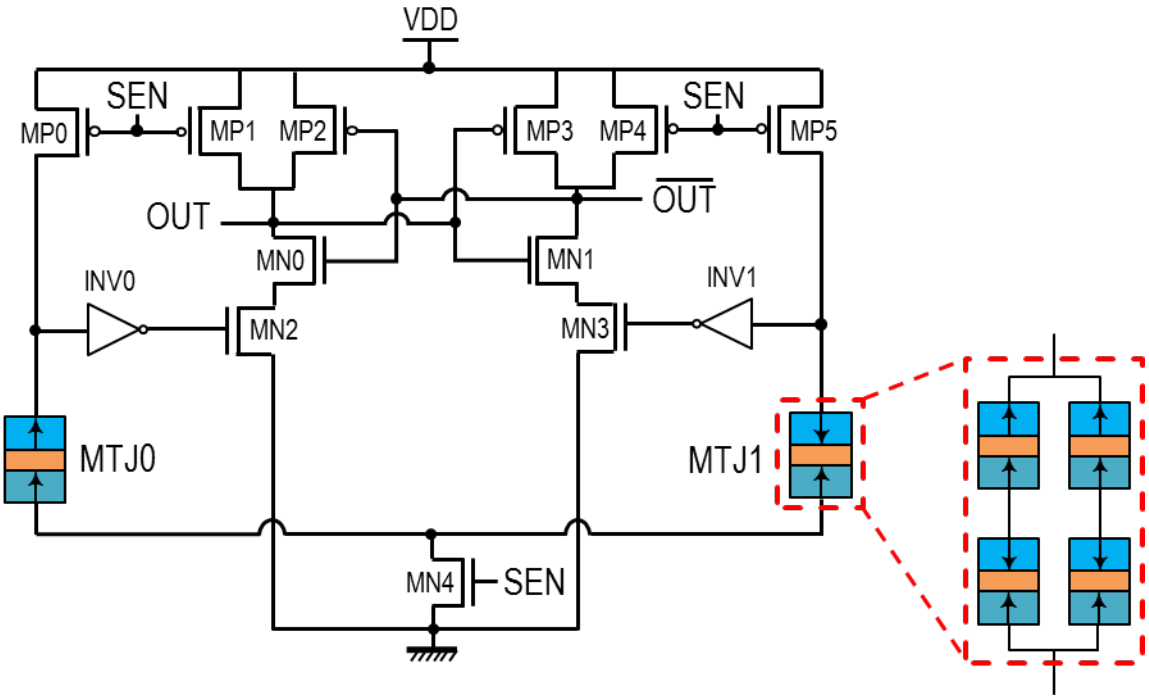



Figure 20: Separated Pre-Charge Sense Amplifier (SPCSA).

Since PCSA consists of fewer CMOS transistors, it offers enhanced performance in terms of sensing delay and Energy Delay Product (EDP) compared to SPCSA. In the branch containing the main MTJ (**MTJ0**) in PCSA, we have four transistors namely MP0, MP1, MN0, and MN2 and in the branch that includes reference MTJ (**MTJ1**) we have also four transistors MP2, MP3, MN1, and MN2 as depicted in Figure 19. However, the main MTJ (**MTJ0**) branch in SPCSA consists of two transistors MP0 and MN4 and two transistors MP5 and MN4 in the reference MTJ (**MTJ1**) branch, which makes it less vulnerable to PV by increasing the SM, as shown in Figure 20. This redesign of the SA introduces an elevated SM, which in turn results in decreased Bit Error Rate (BER). Nonetheless, the reduced BER comes with the cost of utilizing greater number of transistors in SPCSA design, which incurs higher EDP.

A new approach called Adaptive Sense Amplifier (ASA) is proposed herein, which combines PCSA and SPCSA and utilizes their properties in order to increase the performance and reliability of the memory. In ASA design, a select input is used in the circuit called **MODE** to choose between the two SAs based on whether energy efficiency is important or reliability. **MODE** signal controls the operation mode of the circuit to either operate in PCSA mode or SPCSA mode. If the input **MODE** is high, then the circuit will operate in PCSA mode. On the other hand, if **MODE** is low will change the operation of the circuit to SPCSA mode. In order to further reduce the PV effect on the reference cell we can use the configuration shown in Figure 19 and Figure 20 for the reference MTJ (**MTJ1**) which consists of ($MTJ_P$+$MTJ_{AP}$) || ($MTJ_P$+$MTJ_{AP}$). This configuration will provide a resistance of ($MTJ_P$+$MTJ_{AP}$)/2 that will result in good SM and increased PV immunity [53, 89]. The proposed design has been depicted in Figure 21 and waveform of the output of all the designs are provided in Figure 22 in which the two operation of the proposed design is shown.

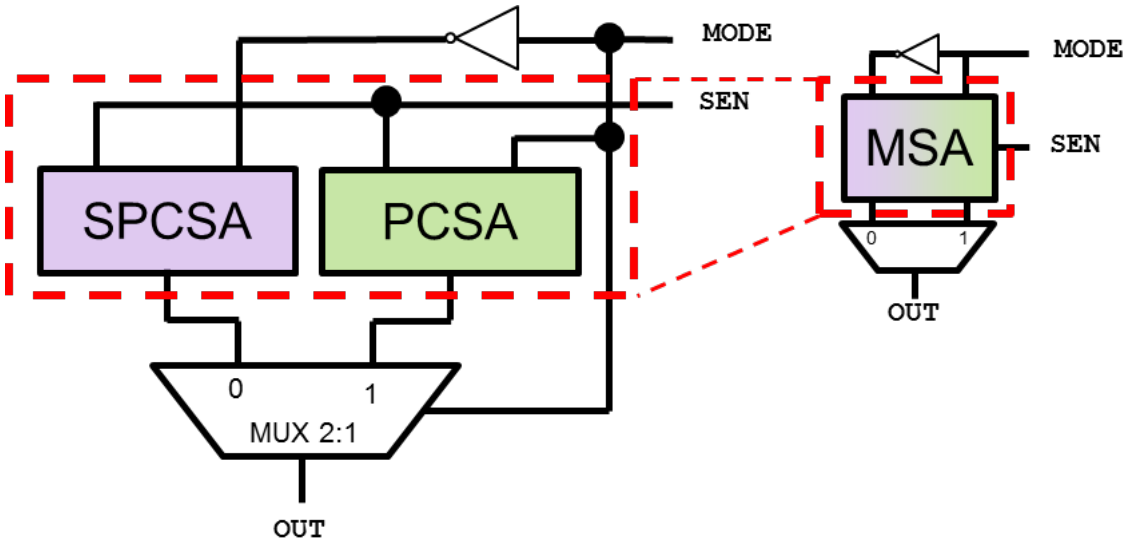



Figure 21: Adaptive Sense Amplifier (ASA).

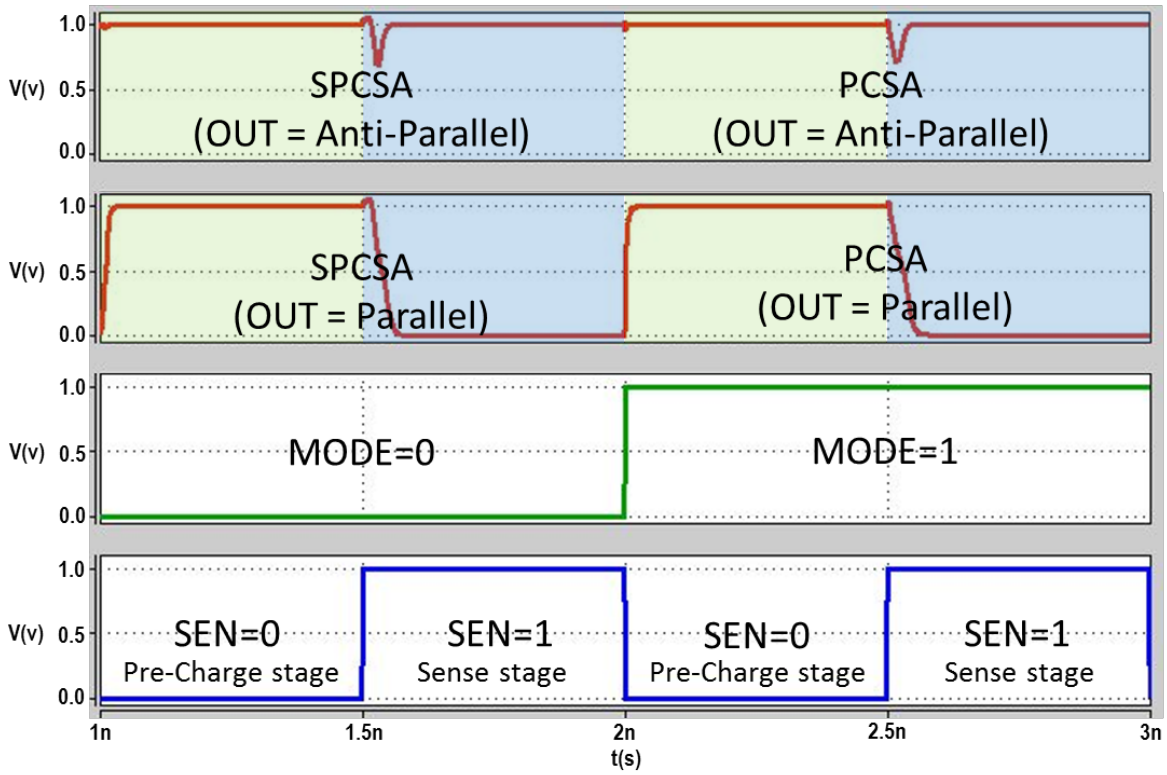



Figure 22: Adaptive Sense Amplifier (ASA) Waveform for Parallel and Anti-Parallel Configurations.

## Simulation Environment

Significant amount of research has analyzed the Power-on Self-Test (POST) and its power and delay overhead [90-94] but since it's a one-time operation it is not going to affect the performance of the memory as a whole and it just introduces a negligible overhead. Taking advantage of this feature, we will be able to analyze memory cells before starting the main operation, which helps us, find out which cells suffer more from PV. An algorithm has been suggested herein has been depicted in Figure 23 which describes the process of the proposed circuit. Simulation results have been extracted using 22nm Predictive Technology Model (PTM) [95] and parameters and PV elements have been provided in Table 5. Every design has been

analyzed in an ideal case where no PV taken into account as well as Monte Carlo simulation in presence of PV elements. The results for the analysis of ideal case have been listed in Table 6.

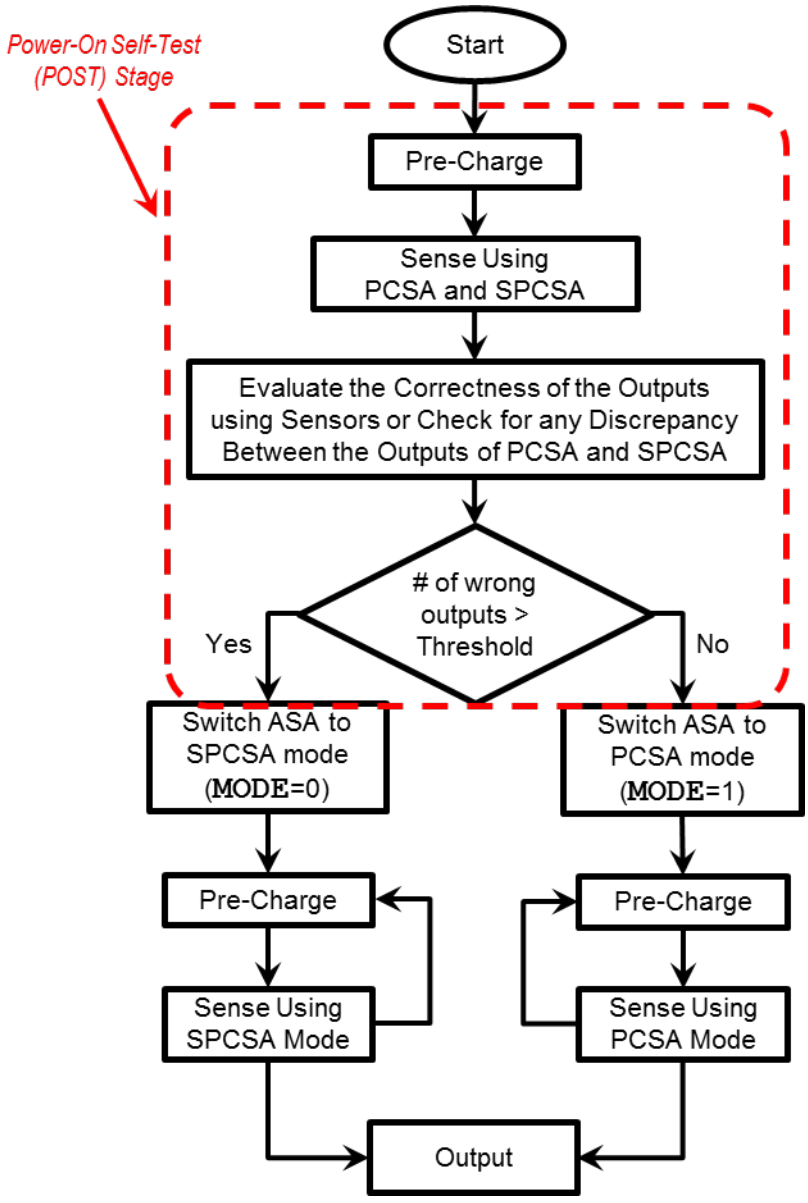



Figure 23: ASA Operational Algorithm.

Furthermore, 10,000 Monte Carlo simulations were performed considering different standard deviations for CMOS transistors' $V_t$ and also MTJ's MgO thickness and shape area in

order to have a variety of cases to analyze during the Simulation. These simulations vary the $V_t$, width, and length of the transistors in the netlist based on a Gaussian distribution having a mean equal to the nominal model card for PTM and $\sigma V_t$ as provided in [96]. Ideally, the $\sigma V_t$ can be adapted to accommodate local and global variations, or their combined effects as considered in this work [32]. Overall variation that has been taken into account here for the MTJs has an effect of 1% and 10% on the MTJs' resistance, which is included in the circuit for the Monte Carlo simulation.

Table 5: Simulation Parameters

| ***Parameter*** | | | | ***Value*** | ***Std. Dev.*** |
|---|---|---|---|---|---|
| *PMOS* | *$V_{th}$ (Threshold Voltage)* | | | 460mV | 50mV (10%) |
| | *Width (=2×Length)* | | | 44nm | 0.44nm (1%) |
| *NMOS* | *$V_{th}$ (Threshold Voltage)* | | | 500mV | 50mV (10%) |
| | *Width (=Length)* | | | 22nm | 0.22nm (1%) |
| *MTJ* | *MgO Thickness* | | | 0.85nm | Effects of variation are applied to TMR |
| | *Shape Area* | *main MTJ (MTJ0)* | | $(\frac{\pi}{4})$x40x40nm$^2$ | |
| | | *reference MTJ (MTJ 1)* | *$MTJ_{AP}$* | $(\frac{\pi}{4})$x30x30nm$^2$ | |
| | | | *$(MTJ_P+MTJ_{AP})$ \|\| $(MTJ_P+MTJ_{AP})$* | $(\frac{\pi}{4})$x40x40nm$^2$ | |
| | *R·A (Resistance×Area)* | | | 5Ω·μm$^2$ | N/A |
| | *α (Damping Factor)* | | | 0.01 | N/A |
| | *TMR (Tunnel Magneto Resistance)* | | | 100% | 1% & 10% |
| *Nominal Voltage* | | | | 1.0V | N/A |
| *SEN Signal Period (T)* | | | | 1ns | N/A |

## Simulation Results

Based on the results listed in Table 6, it can be concluded that ASA operating in PCSA mode, performs better than ASA operating in SPCSA mode, however, suffers more from PV. On the contrary, ASA operating in SPCSA mode offers better performance in terms of reliability and PV immunity compared to ASA operating in PCSA mode due to less BER. PCSA and SPCSA design space for TMR=100% is illustrated in Figure 24. Monte Carlo reliability simulation results are depicted in Figure 25 and Figure 26.

Based on the results listed in Table 6, Table 7, and Table 8 and as depicted in Figure 25 and Figure 26, it can be concluded that, ASA operating in PCSA mode attains 6-fold improvement over ASA operating in SPCSA mode on average in terms of EDP by maintaining on average 2.43μW and 8.7ps, less power consumption and reduced sensing latency, respectively. On the contrary, ASA operating in SPCSA mode increases the reliability by having 6% reduced BER (for TMR=100%) on average caused by PV compared to ASA operating in PCSA mode.

Furthermore, it can be observed that by optimizing the reference MTJ and using $(MTJ_P+MTJ_{AP}) \parallel (MTJ_P+MTJ_{AP})$ configuration, the BER can be reduced by 15% on average (for TMR=100%). In addition, based on the results of Monte Carlo Simulation, it is clear that larger MTJ resistance reduces the impact of variation on sensing output. From physical layout designs of PCSA, SPCSA, and ASA are depicted in Figure 27 it is observed that the area overhead of ASA compared to PCSA is about 3-fold and compared to SPCSA is approximately 1.5-fold. However, since the SA is shared among memory cells within memory, the area over head is negligible. As a trade-off factor, by sacrificing the area we are gaining reliability.

Table 6: Simulation Results for Baseline Design with no PV ($MTJ_{REF}$=5.7KΩ)

| *Design* | *Tech. Node* | *Area (Device Count)* | | | *Anti-Parallel (6.4 KΩ)* | | | *Parallel (3.2 KΩ)* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *PMOS Trans.* | *NMOS Trans.* | *MTJ* | *Delay (ps)* | *Power (μW)* | *EDP (J*ps)* | *Delay (ps)* | *Power (μW)* | *EDP (J*ps)* |
| *PCSA (implemented herein)* | 22nm | 4 | 3 | 2 | 20.8 | 0.79 | 16.43 | 13.5 | 0.77 | 10.24 |
| *SPCSA (implemented herein)* | 22nm | 8 | 5 | 2 | 28.8 | 3.21 | 92.45 | 22.9 | 3.21 | 73.51 |

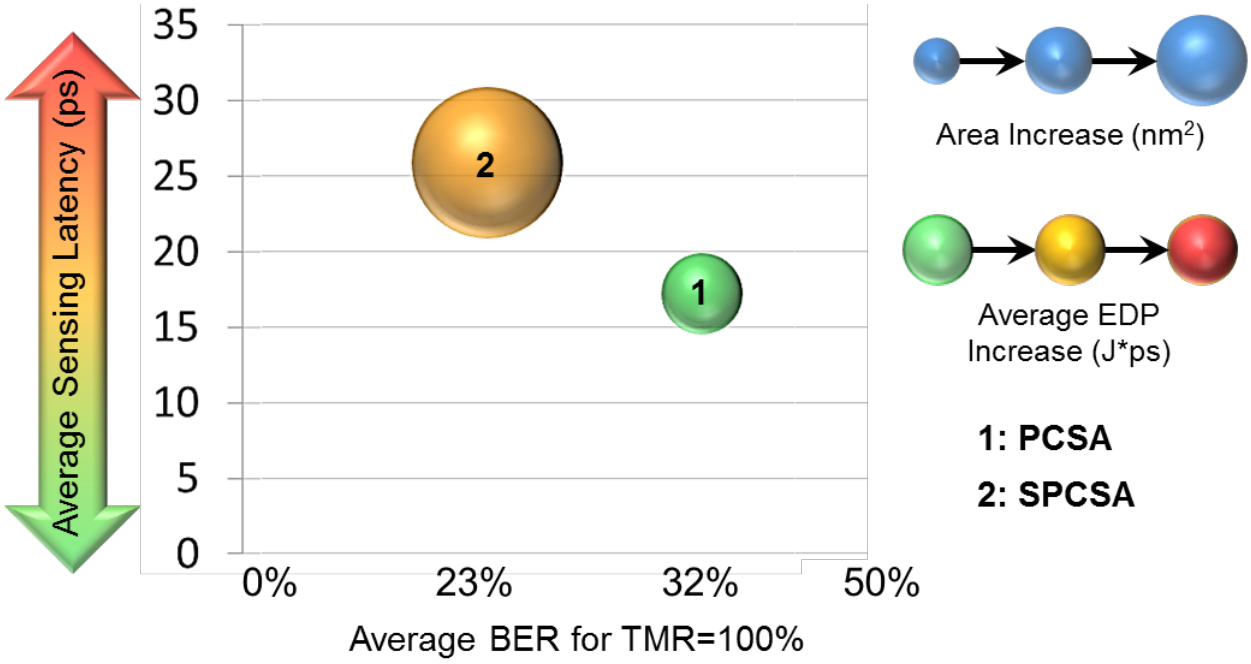



Figure 24: PCSA and SPCSA design space.

Table 7: Monte Carlo Simulation 10,000 Run Results ($MTJ_{REF}$=5.7KΩ, $MTJ_P$=3.2KΩ, and $MTJ_{AP}$=6.4KΩ for TMR=100%).

| *Design* | *Bit Error Rate (BER)(%)(1% Variation in TMR and 10% in Vth)* | | | | | |
|---|---|---|---|---|---|---|
| | *TMR=100%* | *TMR=150%* | *TMR=200%* | *TMR=250%* | *TMR=300%* | *TMR=350%* |
| *PCSA (implemented herein)* | 38.29 | 17.15 | 5.89 | 1.65 | 0.30 | 0.03 |
| *SPCSA (implemented herein)* | 34.04 | 9.41 | 1.35 | 0.09 | 0.01 | 0.00 |
| *Design* | *Bit Error Rate (BER)(%)(10% Variation in TMR and 10% in Vth)* | | | | | |
| | *TMR=100%* | *TMR=150%* | *TMR=200%* | *TMR=250%* | *TMR=300%* | *TMR=350%* |
| *PCSA (implemented herein)* | 38.44 | 17.18 | 6.06 | 1.62 | 0.34 | 0.04 |
| *SPCSA (implemented herein)* | 34.32 | 10.00 | 1.44 | 0.13 | 0.02 | 0.00 |

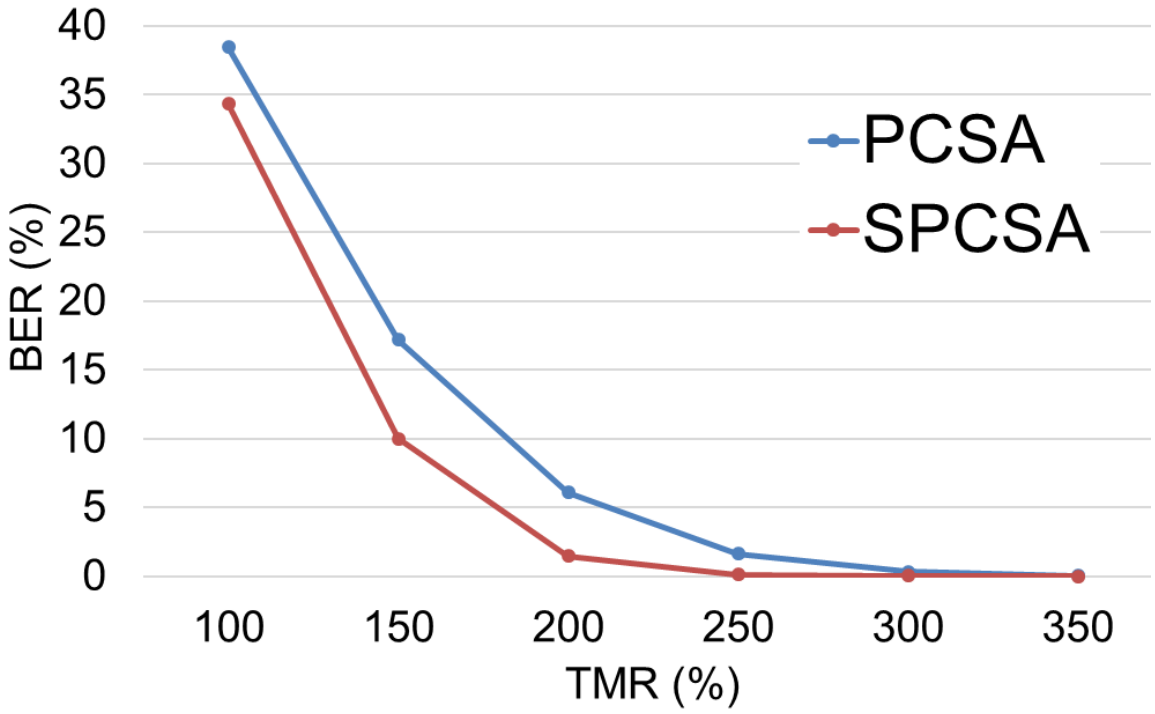



Figure 25: BER (%) Monte Carlo Simulation 10,000 run results for 10% Variation in TMR and 10% in $V_t$ ($MTJ_{Ref}$=5.7KΩ, $MTJ_P$=3.2KΩ, and $MTJ_{AP}$=6.4KΩ for TMR=100%).

Table 8: Monte Carlo Simulation 10,000 Run Results ($MTJ_{REF}$=4.8KΩ, $MTJ_P$=3.2KΩ, and $MTJ_{AP}$=6.4KΩ for TMR=100%).

| *Design* | *Bit Error Rate (BER)(%)(1% Variation in TMR and 10% in Vth)* | | | | | |
|---|---|---|---|---|---|---|
| | *TMR=100%* | *TMR=150%* | *TMR=200%* | *TMR=250%* | *TMR=300%* | *TMR=350%* |
| *PCSA (implemented herein)* | 24.87 | 9.01 | 2.71 | 0.54 | 0.05 | 0.00 |
| *SPCSA (implemented herein)* | 17.68 | 3.41 | 0.32 | 0.02 | 0.00 | 0.00 |
| *Design* | *Bit Error Rate (BER)(%)(10% Variation in TMR and 10% in Vth)* | | | | | |
| | *TMR=100%* | *TMR=150%* | *TMR=200%* | *TMR=250%* | *TMR=300%* | *TMR=350%* |
| *PCSA (implemented herein)* | 24.90 | 9.31 | 2.69 | 0.59 | 0.07 | 0.00 |
| *SPCSA (implemented herein)* | 17.78 | 3.58 | 0.35 | 0.02 | 0.00 | 0.00 |

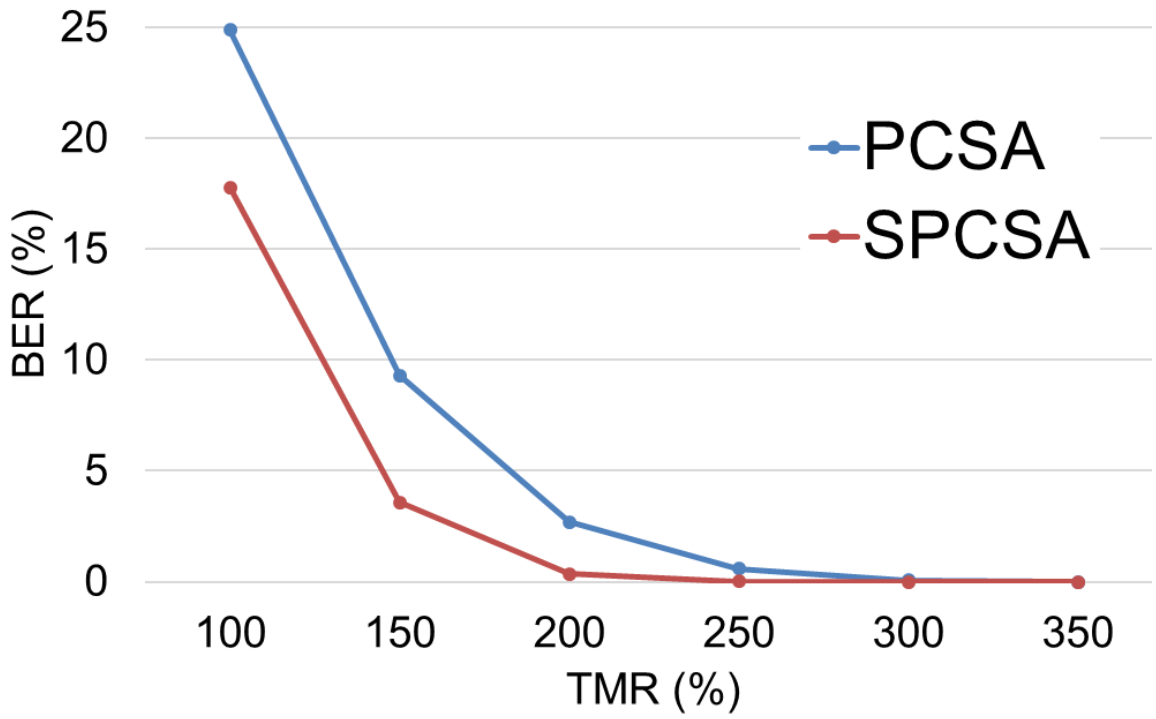



Figure 26: BER (%) Monte Carlo Simulation 10,000 run results for 10% Variation in TMR and 10% in $V_t$ ($MTJ_{Ref}$=4.8KΩ, $MTJ_P$=3.2KΩ, and $MTJ_{AP}$=6.4KΩ for TMR=100%).

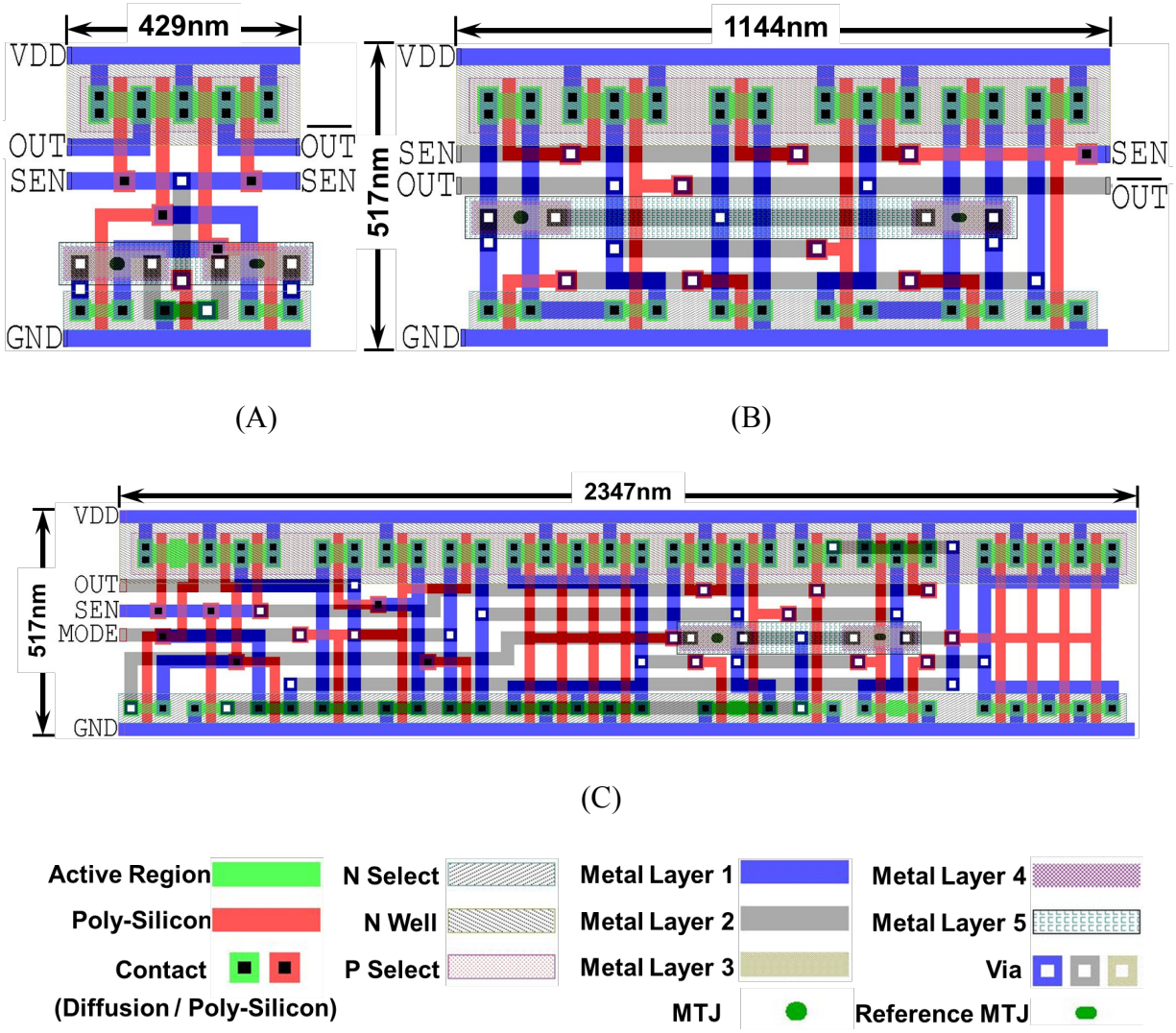



Figure 27: A) PCSA Layout, B) SPCSA Layout, C) ASA Layout, and D) Layout legend.

## Self-Organized Sub-bank (SOS) Approach

Self-Organized Sub-bank (SOS) partitions NVM arrays into several banks to directly access requested data while introducing individualized sensing. Sub-banks are evaluated and tagged during an initial Power-On Self-Test (POST) phase to identify the preferred SA for that

particular sub-bank. To be specific, if the error rate of the impacted NVM cells in a sub-bank exceeds the pre-defined threshold, a High Resilience (HR) SA is assigned. Otherwise, a Low Energy Delay Product (LEDP) SA offering reduced delay and power consumption is assigned. For instance, Figure 28 depicts an SOS-enabled on-chip STT-MRAM cache whereby SOS maximizes NVM sensing reliability while minimizing power consumption. Additionally, SOS enables new means to increase yield of high capacity NVM arrays using intrinsic adaptation via sub-banking. Consider an $x$ MB NVM array organized into $m$ cache lines entailing $n$ Bytes each. SOS device cost totals $\frac{2}{m}$ compared to $\frac{1}{m}$ for conventional NVM arrays. Whereas conventional designs require a higher supply voltage to ensure adequate SM, SOS allows reduced voltage operation reliably while incurring a very small overhead as $m \approx 512$ or more.

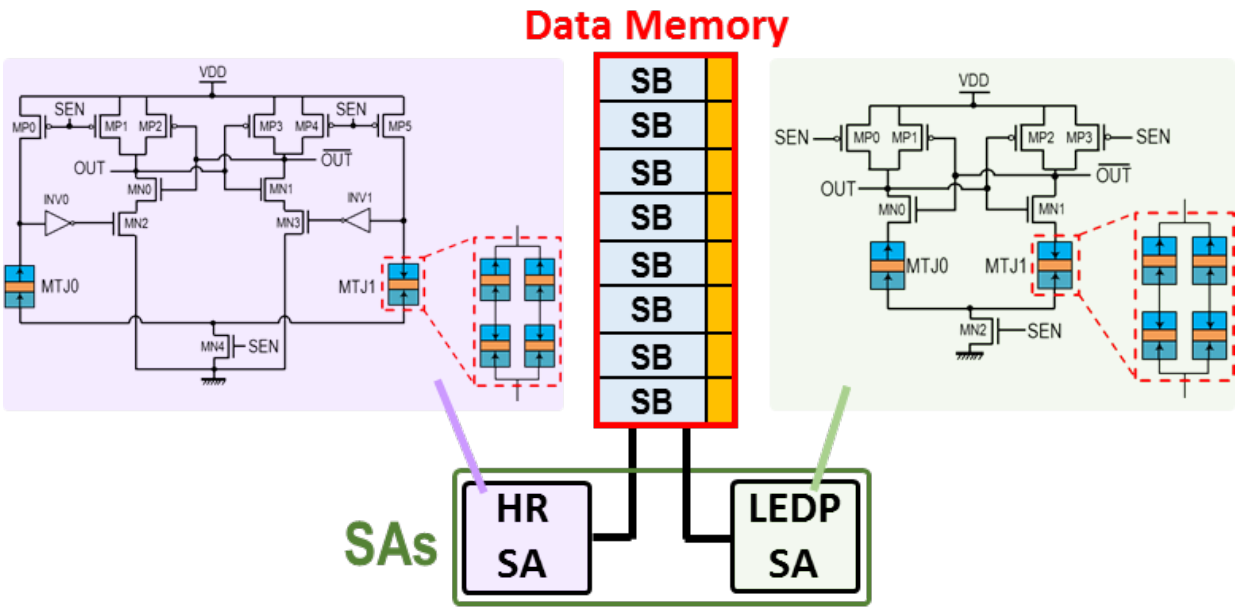



Figure 28: SOS Strategy applied to NVM Cache Array, SB: Sub-Bank, SA: Sense Amplifier, HR: High Resilience, LEDP: Low Energy Delay Product.

## Summary

Proposed circuit can be used in large (e.g. 96MB) Last Level Cache (LLC) to sub-organize it to many sub-banks, as usual, in which each sub-bank has some light-weight alternate sensing mechanism (either SPCSA or PCSA). Furthermore, based on error rate of Error Correcting Code (ECC) circuit already presented in cache lines, memory banks can be demoted to lower or promoted to higher sub-hierarchies. As a result, PV, which will always occur at highly scaled technology nodes and especially NVM technologies like MTJs, will be leveraged as a self-configuring advantage of creating a sub-hierarchy in this level of the memory hierarchy. To take advantage of the proposed design, the more frequently accessed data goes into the faster half of the sub hierarchy using a new preference placement strategy.

Results have shown that the ASA has small power and delay overhead and negligible area overhead since the SA is shared among all memory cells within an array. In conclusion, SOS is a circuit-architecture cross-layer solution, which combats the common PV problem in the emerging NVM technologies by engaging PV-resilient SAs array offering acceptable resistive SM. In addition, SOS manages to meet the power budget constraint in IoT devices through low-power SAs in sub-banks that experience lower rates of PV. Furthermore, we classify the output of the sensed data based on its impact on the execution flow of the workload. This classification of experimental outcomes is vital to identify the efficiency of SOS to accommodate critical read operations. Our experimental results indicate that the energy consumption of SOS is as high as LLC with conventional SAs. The confluence of these factors in turn significantly increase the reliability of realistic program execution.

# CHAPTER FIVE: CONCLUSION

This thesis analyzed STT-MRAM, a device to overcome scaling and power limitations of CMOS devices. As shown in Figure 29, several conclusions can be drawn from the results developed herein. First, it was validated that CMOS scaling can result in significant power and energy reduction. Next, it was demonstrated in this thesis that it is possible to STT-MRAM elements can be a promising alternative for scaled CMOS memory elements due to its zero leakage and non-volatility features. Finally, a novel SA was proposed alongside with a memory configuration to mitigate bit errors while incurring high performance. A conclusion drawn from these results is that ASA combined with SOS offers a circuit-architecture solution that is clearly beneficial to minimizing BER due to PV in STT-MRAM.

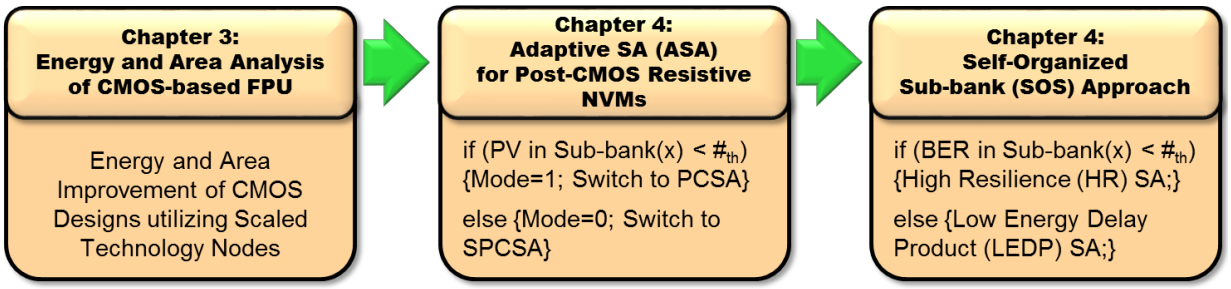



Figure 29: Conclusions Drawn from Study Herein.

## Technical Summary and Insight Gained

In order to reduce the amount of power consumed in CMOS devices there are several methods explained within this thesis. Out of all the solutions, miniaturizing the device or in other words using technology nodes with smaller footprint is one of the most promising. In order to demonstrate this, a single precision FPU was proposed and designed herein using 45nm and

15nm libraries. Results have shown that using the 45nm library, power consumption for a zero-negative slack clock period of 5ns is 3.15-fold to 4.56-fold larger than the same design synthesized using the 15nm with default parameters. Results indicate that using 15nm technology allows the FPU to consume about four times less energy than 45nm technology.

Due to scaling and power limitations of CMOS devices, researchers are exploring alternatives that offer better performance to overcome these limitations. As a result, STT-MRAM devices are considered a feasibly implemented solution for beyond CMOS technologies. Improving the reliability of STT-MRAM, however, is of great importance. The use of ASA was shown to improve the reliability and performance of PCSA and SPCSA circuits within an acceptable area overhead margin. In SPCSA mode, ASA reduced the average BER by 6%. In PCSA mode, ASA improves the average EDP by 6-fold. In summary, we addressed reliability and performance challenges related to STT-MRAM Sensing devices as depicted in Figure 12. First, we utilized ASA to address the challenging task of designing a reliable SA, especially on resource-constrained devices. Next, the ASA developed is able to be utilized in SOS. We developed SOS, which is able to improve the reliability and performance of our memory hierarchy by Self-Organizing the memory sub-banks to provide the most effective, efficient, and reliable result as shown in the results.

The SOS has shown great merit to realize functionality that would be incredibly difficult to hand-design given a resource-constrained platform. In an ideal system with no constraints, lack of PV, temperature variations, and device mismatch, as well as a possibly unconstrained amount of resources cannot be used as a source of comparison. Our simulation with accurate models allows

all these characteristics to be considered during modeling since all of them intrinsically manipulate reliability and performance of the circuits.

When developing the techniques herein, the aspects that were the most straightforward to develop were:

- implementing Floating Point Unit using 45nm and 15nm NANGATE Open Cell libraries,
- and implementing the ASA and SOS using 22nm PTM library,

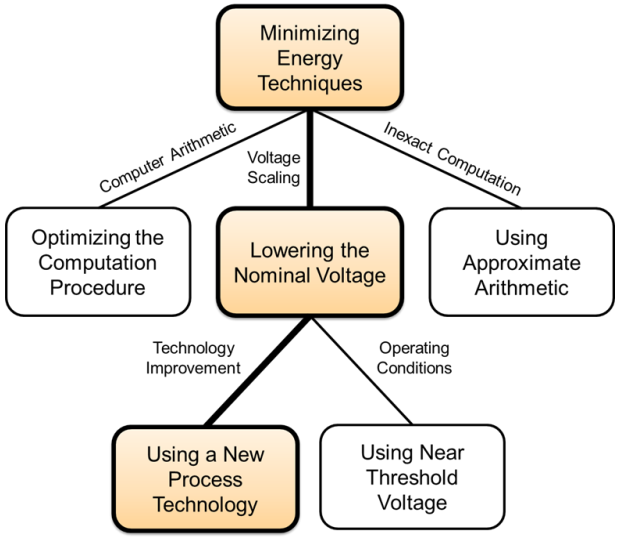
Some of the most challenging aspects faced when developing the techniques herein include:

- designing a reliable circuit that also can offer improved performance parameters,
- finding the right parameters for accurate PV, reliability, and energy analysis using Monte Carlo Simulation,
- and time consuming simulations on a relatively large circuit in order to have more accurate results.

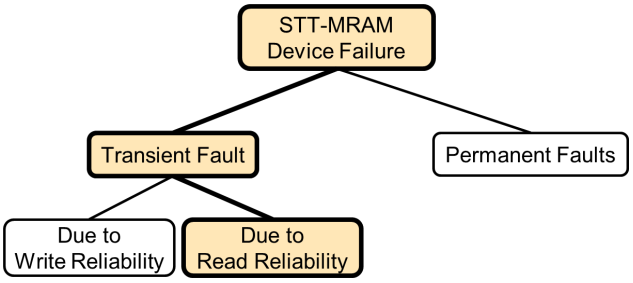
## Scope, Limitations, and Future Directions

The scope of this research centers around reliable and high performance for the potential benefit of greater efficiency in computation at current technology scaling limits as shown in Figure 30. The performance of SOS could benefit by more exploration by applying SOS techniques to larger Memories with more storage capacity. Using such devices could perhaps show that a large of a range of accurate computation is possible.

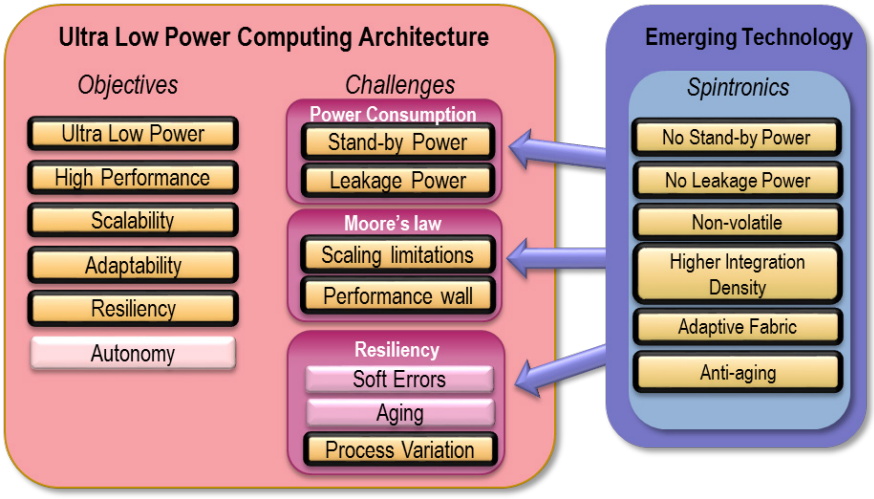
Future research can be directed towards improved memory hierarchies to mitigate reliability challenges such as Single Event Upsets (SEUs) or Multiple Event Upsets (MEUs), implementing LIM that reduces number of interconnects in the design, resulting in energy efficiency, and Reconfigurable Spintronics Architectures that introduce adaptability to the design to realize evolvability [28, 36, 97, 98]. In addition, exploration, improvement, and expansion of SOS technique for LIM and RSF can be further investigated.




(A)




(B)




(C)

Figure 30: (A) Eenrgy Efficient Scaled CMOS Designs, (B) Sensing Reliability Challenges of STT-MRAM Devices, and (C) Emerging Technology Benefits and Challenges with the Thesis Scope Outlined in Yellow Boxes.

# REFERENCES


[1] T. Karnik, S. Borkar, and V. De, "Sub-90nm technologies: challenges and opportunities for CAD," in *Proceedings of the International Conference on Computer-Aided Design (ICCAD)*, 2002, pp. 203-206.

[2] H. Bahr and R. DeMara, "OTBSAF Scalability on Pentium III/4 and Athlon 64/XP3000 Architectures," *MSIAC Modeling and Simulation Journal,* vol. 6, pp. 1-4, 2005.

[3] J. Kawa. (2013, FinFET design, manufacturability, and reliability. *Synopsys DesignWare Technical Bulletin 1(1)*.

[4] H. B. Jang, J. Lee, J. Kong, T. Suh, and S. W. Chung, "Leveraging Process Variation for Performance and Energy: In the Perspective of Overclocking," *IEEE Transactions on Computers,* vol. 63, pp. 1316-1322, 2014.

[5] J. Kawa and A. Biddle. (2012, FinFET: The Promises and the Challenges. *Synopsys Insight Newsletter (3)*.

[6] S. Hamdioui, "NBTI modeling in the framework of temperature variation," in *Proceedings of Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2010, pp. 283-286.

[7] V. De, "Energy efficient computing in nanoscale CMOS: Challenges and opportunities," in *Proceedings of Asian Solid-State Circuits Conference (A-SSCC)*, 2014, pp. 121-124.

[8] A. Kerber and T. Nigam, "Challenges in the characterization and modeling of BTI induced variability in metal gate/High-k CMOS technologies," in *Proceedings of International Reliability Physics Symposium (IRPS)*, 2013, pp. 2D. 4.1-2D. 4.6.

[9] M. I. Khan, A. R. Buzdar, and F. Lin, "Self-heating and reliability issues in FinFET and 3D ICs," in *Proceedings of 12th International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, 2014, pp. 1-3.

[10] G. Templeton. (2015). *What is Moore's Law?* Available: http://www.extremetech.com/extreme/210872-extremetech-explains-what-is-moores-law

[11] S. Salehi, D. Fan, and R. DeMara, "Survey of STT-MRAM Cell Design Strategies: Taxonomy and Sense Amplifier Tradeoffs for Resiliency," *ACM Journal on Emerging Technologies in Computing Systems (JETC),* p. 15, in press.

[12] A. Ramadan. (2013). *Catching layout-dependent effects on-the-fly*. Available: http://www.techdesignforums.com/practice/technique/lde-layout-dependent-effects-fly/

[13] V. P. Nelson, "Computer-Aided Design of ASICs Concept to Silicon," ed, 2015.

[14] K. Cheung. (2009). *Solido Variation Designer*. Available: http://edablog.com/2009/01/22/transistor-level-design/

[15] M. Santarini. (2003). *Cadence rolls custom-IC tools into one platform*. Available: http://www.eetimes.com/document.asp?doc_id=1217217

[16] M. Sharad, D. Fan, K. Aitken, and K. Roy, "Energy-efficient non-boolean computing with spin neurons and resistive memory," *IEEE Transactions on Nanotechnology,* vol. 13, pp. 23-34, 2014.

[17] H. Bahr, R. DeMara, and M. Georgiopoulos, "Integer-Encoded Massively Parallel Processing of Fast-Learning ARTMAP Networks," in *Proceedings of the SPIE AeroSense Symposium (AeroSense)*, Orlando, Florida, USA, 1997, pp. 678-689.

[18] Y. Zhang, W. Zhao, J.-O. Klein, W. Kang, D. Querlioz, Y. Zhang, *et al.*, "Spintronics for low-power computing," in *Proceedings of Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2014, pp. 1-6.

[19] H. Jarollahi, N. Onizawa, V. Gripon, N. Sakimura, T. Sugibayashi, T. Endoh, *et al.*, "A Nonvolatile Associative Memory-Based Context-Driven Search Engine Using 90 nm CMOS/MTJ-Hybrid Logic-in-Memory Architecture," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems,* vol. 4, pp. 460-474, 2014.

[20] R. F. DeMara and D. I. Moldovan, "Performance Indices for Parallel Marker-Propagation," in *International Conference on Parallel Processing (ICPP)*, 1991, pp. 658-659.

[21] R. F. DeMara, Y. Tseng, and A. Ejnioui, "Tiered algorithm for distributed process quiescence and termination detection," *IEEE Transactions on Parallel and Distributed Systems,* vol. 18, pp. 1529-1538, 2007.

[22] R. F. DeMara, C. Lin, S. Kuo, and B. S. Motlagh, "Barrier Synchronization Techniques for Distributed Process Creation," in *IPPS*, 1994, pp. 597-603.

[23] Y. Tseng, R. F. DeMara, and P. Wilder, "Distributed-sum termination detection supporting multithreaded execution," *Parallel Computing,* vol. 29, pp. 953-968, 2003.

[24] Y. Tseng and R. F. DeMara, "Communication pattern based methodology for performance analysis of termination detection schemes," in *Proceedings of 9th International Conference on Parallel and Distributed Systems*, 2002, pp. 535-541.

[25] R. F. DeMara, Y. Tseng, K. Drake, and A. Ejnioui, "Capability classes of multiprocessor synchronization techniques," *International Journal of Computers and Applications,* vol. 28, pp. 342-349, 2006.

[26] A. J. Gonzalez, J. Leigh, R. F. DeMara, A. Johnson, S. Jones, S. Lee*, et al.*, "Passing an enhanced Turing test–interacting with lifelike computer representations of specific individuals," *Journal of Intelligent Systems,* vol. 22, pp. 365-415, 2013.

[27] N. Imran, R. F. DeMara, J. Lee, and J. Huang, "Self-adapting Resource Escalation for Resilient Signal Processing Architectures," *Journal of Signal Processing Systems,* vol. 77, pp. 257-280, 2014.

[28] R. Al-Haddad, R. Oreifej, R. Ashraf, and R. F. DeMara, "Sustainable modular adaptive redundancy technique emphasizing partial reconfiguration for reduced power consumption," *International Journal of Reconfigurable Computing,* vol. 2011, 2011.

[29] M. Alawad, Y. Bai, R. DeMara, and M. Lin, "Energy-efficient multiplier-less discrete convolver through probabilistic domain transformation," in *Proceedings of the International Symposium on Field-Programmable Gate Arrays*, 2014, pp. 185-188.

[30] J. Y. F. Tong, D. Nagle, and R. A. Rutenbar, "Reducing power by optimizing the necessary precision/range of floating-point arithmetic," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 8, pp. 273-286, 2000.

[31] S.-R. Kuang, K.-Y. Wu, and K.-K. Yu, "Energy-efficient multiple-precision floating-point multiplier for embedded applications," *Journal of Signal Processing Systems,* vol. 72, pp. 43-55, 2013.

[32] R. A. Ashraf, A. Alzahrani, and R. F. DeMara, "Extending modular redundancy to NTV: Costs and limits of resiliency at reduced supply voltage," in *Proceedings of Workshop on Near Threshold Computing (WNTC)*, 2014.

[33] H. Kaul, M. Anders, S. Hsu, A. Agarwal, R. Krishnamurthy, and S. Borkar, "Near-threshold voltage (NTV) design: opportunities and challenges," in *Proceedings of 49th Design Automation Conference (DAC)*, 2012, pp. 1153-1158.

[34] J. Han and M. Orshansky, "Approximate computing: An emerging paradigm for energy-efficient design," in *Proceedings of 18th European Test Symposium (ETS)*, 2013, pp. 1-6.

[35] N. Imran, R. A. Ashraf, and R. F. DeMara, "Power and quality-aware image processing soft-resilience using online multi-objective GAs," *International Journal of Computational Vision and Robotics,* vol. 5, pp. 72-98, 2015.

[36] R. S. Oreifej, C. A. Sharma, and R. F. DeMara, "Expediting GA-based evolution using group testing techniques for reconfigurable hardware," in *Proceedings of International Conference on Reconfigurable Computing and FPGA's (ReConFig)*, 2006, pp. 1-8.

[37] S. Galal and M. Horowitz, "Energy-efficient floating-point unit design," *IEEE Transactions on Computers,* vol. 60, pp. 913-922, 2011.

[38] K. Swaminathan, M. S. Kim, N. Chandramoorthy, B. Sedighi, R. Perricone, J. Sampson*, et al.*, "Modeling steep slope devices: From circuits to architectures," in *Proceedings of Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2014, pp. 1-6.

[39] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. B. Tahoori, "Read disturb fault detection in STT-MRAM," in *Proceedings of International Test Conference (ITC)*, 2014, pp. 1-7.

[40] W. Kang, Z. Li, J.-O. Klein, Y. Chen, Y. Zhang, D. Ravelosona*, et al.*, "Variation-tolerant and disturbance-free sensing circuit for deep nanometer STT-MRAM," *IEEE Transactions on Nanotechnology,* vol. 13, pp. 1088-1092, 2014.

[41] Y. Emre, C. Yang, K. Sutaria, Y. Cao, and C. Chakrabarti, "Enhancing the reliability of STT-RAM through circuit and system level techniques," in *Proceedings of Workshop on Signal Processing Systems (SiPS)*, 2012, pp. 125-130.

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

[43] 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 (ICCAD)*, 2013, pp. 9-16.

[44] 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, 2010.

[45] W. Zhao, Y. Zhang, T. Devolder, J.-O. Klein, D. Ravelosona, C. Chappert*, et al.*, "Failure and reliability analysis of STT-MRAM," *Microelectronics Reliability,* vol. 52, pp. 1848-1852, 2012.

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

[47] K. Zhang, G. Bedette, and R. F. DeMara, "Triple modular redundancy with standby (TMRSB) supporting dynamic resource reconfiguration," in *AutotestCon*, 2006, pp. 690-696.

[48] H. Lee, J. G. Alzate, R. Dorrance, X. Q. Cai, D. Markovic, P. Khalili Amiri*, et al.*, "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, 2015.

[49] Y. Lakys, W. S. Zhao, T. Devolder, Y. Zhang, J.-O. Klein, D. Ravelosona*, et al.*, "Self-enabled "error-free" switching circuit for spin transfer torque MRAM and logic," *IEEE Transactions on Magnetics,* vol. 48, pp. 2403-2406, 2012.

[50] F. Ren, H. Park, R. Dorrance, Y. Toriyama, C.-K. K. Yang, and D. Marković, "A body-voltage-sensing-based short pulse reading circuit for spin-torque transfer RAMs (STT-RAMs)," in *Proceedings of 13th International Symposium on Quality Electronic Design (ISQED)*, 2012, pp. 275-282.

[51] D. Halupka, S. Huda, W. Song, A. Sheikholeslami, K. Tsunoda, C. Yoshida*, et al.*, "Negative-resistance read and write schemes for STT-MRAM in 0.13µm CMOS," in *Proceedings of International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2010, pp. 256-257.

[52] J. P. Kim, T. Kim, W. Hao, H. M. Rao, K. Lee, X. Zhu*, et al.*, "A 45nm 1Mb embedded STT-MRAM with design techniques to minimize read-disturbance," in *Proceedings of Symposium on VLSI Circuits-Digest of Technical Papers*, 2011.

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

[54] 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, 2014.

[55] 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 *Proceedings of International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 1530-1533.

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

[57] 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, 2012.

[58] 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," *IEEE Transactions on Magnetics,* vol. 45, pp. 3784-3787, 2009.

[59] 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, 2014.

[60] W. Kang, E. Deng, J.-O. Klein, Y. Zhang, Y. Zhang, C. Chappert, *et al.*, "Separated precharge sensing amplifier for deep submicrometer MTJ/CMOS hybrid logic circuits," *IEEE Transactions on Magnetics,* vol. 50, pp. 1-5, 2014.

[61] 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," *Journal of Semiconductor Technology and Science,* vol. 16, pp. 31-38, 2016.

[62] K. Tsuchida, T. Inaba, K. Fujita, Y. Ueda, T. Shimizu, Y. Asao, *et al.*, "A 64Mb MRAM with Clamped-Reference and Adequate-Reference Schemes," in *Proceedings of International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2010, p. 3.

[63] 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, 2012.

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

[65] W. Kang, W. Zhao, Z. Wang, Y. Zhang, J.-O. Klein, Y. Zhang, *et al.*, "A low-cost built-in error correction circuit design for STT-MRAM reliability improvement," *Microelectronics Reliability,* vol. 53, pp. 1224-1229, 2013.

[66] 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, 2015.

[67] 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, 2012.

[68] 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, 2014.

[69] W. Kang, Z. Li, Z. Wang, E. Deng, J.-O. Klein, Y. Zhang*, et al.*, "Variation-tolerant high-reliability sensing scheme for deep submicrometer STT-MRAM," *IEEE Transactions on Magnetics,* vol. 50, pp. 1-4, 2014.

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

[71] M. Martins, J. M. Matos, R. P. Ribas, A. Reis, G. Schlinker, L. Rech*, et al.*, "Open Cell Library in 15nm FreePDK Technology," in *Proceedings of International Symposium on Physical Design (ISPD)*, 2015, pp. 171-178.

[72] I. o. Electrical and E. Engineers, "IEEE Standard for Binary Floating-point Arithmetic," ed: IEEE, 1985.

[73] J. Knudsen. (2008, Nangate 45nm Open Cell Library. *CDNLive, EMEA*.

[74] S. Salehi and R. F. DeMara, "Energy and Area Analysis of a Floating-Point Unit in 15nm CMOS Process Technology," in *Proceedings of SoutheastCon*, 2015, pp. 1-5.

[75] R. Usselman, "Documentation for Floating Point Unit," ed.

[76] M. Ortega and J. Figueras, "Short Circuit Power Modeling in Submicron CMOS," in *Proceedings of International Workshop on Power And Timing Modeling, Optimization and Simulation (PATMOS)*, 1996, pp. 147-166.

[77] (2010). *Design Compiler*. Available: http://www.synopsys.com/Tools/Implementation/RTLSynthesis/DesignCompiler/Pages/default

[78] F. Alghareb, R. Ashraf, A. Alzahrani, and R. DeMara, "Energy and Delay Tradeoffs of Soft Error Masking for 16nm FinFET Logic Paths: Survey and Impact of Process Variation in Near Threshold Region," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. PP, p. 5, 2016.

[79] R. A. Ashraf, A. Al-Zahrani, N. Khoshavi, R. Zand, S. Salehi, A. Roohi*, et al.*, "Reactive rejuvenation of CMOS logic paths using self-activating voltage domains," in *Proceedings of International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 2944-2947.

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

[81] S. Crawford and R. DeMara, "Cache Coherence in Multiport Memory Architecture," in Proceedings of the Second," in *Proceedings of the Second International Conference on Massively Parallel Computing Systems (MPCS-95)*, 1995.

[82] X. Chen, N. Khoshavi, J. Zhou, D. Huang, R. F. DeMara, J. Wang*, et al.*, "AOS: adaptive overwrite scheme for energy-efficient MLC STT-RAM cache," in *Proceedings of 53rd Design Automation Conference (DAC)*, 2016, p. 170.

[83] R. Zand, A. Roohi, S. Salehi, and R. DeMara, "Scalable Adaptive Spintronic Reconfigurable Logic using Area-Matched MTJ Design," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 63, p. 5, July 2016.

[84] A. Roohi, R. Zand, and R. DeMara, "A Tunable Majority Gate based Full Adder using Current-Induced Domain Wall Nanomagnets," *IEEE Transactions on Magnetics,* vol. 52, p. 7, August 2016.

[85] S. D. Pyle, H. Li, and R. F. DeMara, "Compact low-power instant store and restore D flip-flop using a self-complementing spintronic device," *Electronics Letters,* 2016.

[86] W. Kang, Y. Zhang, Z. Wang, J.-O. Klein, C. Chappert, D. Ravelosona*, et al.*, "Spintronics: Emerging ultra-low-power circuits and systems beyond MOS technology," *ACM Journal on Emerging Technologies in Computing Systems (JETC),* vol. 12, p. 16, 2015.

[87] X. Fong, Y. Kim, K. Yogendra, D. Fan, A. Sengupta, A. Raghunathan*, et al.*, "Spin-Transfer Torque Devices for Logic and Memory: Prospects and Perspectives," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 35, pp. 1-22, 2016.

[88] A. Agrawal, A. Ansari, and J. Torrellas, "Mosaic: Exploiting the Spatial Locality of Process Variation to Reduce Refresh Energy in On-Chip eDram Modules," in *Proceedings of 20th International Symposium on High Performance Computer Architecture (HPCA)*, 2014, pp. 84-95.

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

[90] Y.-J. Park and I.-g. Jung, "Computer system having non-volatile memory and method of operating the computer system," ed: Google Patents, 2013.

[91] K.-s. Park, "Method For Repairing Defective Memory Cells in Semiconductor Memory Device," ed: US Patent 20,160,062,819, 2016.

[92] O.-s. Kwon, P. Chulwoo, and L. Yunsang, "User device having nonvolatile random access memory and method of booting the same," ed: Google Patents, 2013.

[93] O.-s. Kwon and J. Oh, "Nonvolatile random access memory and data management method," ed: Google Patents, 2015.

[94] Y.-J. Park and J. Ilguy, "Main memory system storing operating system program and computer system including the same," ed: Google Patents, 2015.

[95] (2008). *22nm Predictive Technology Model (PTM)*. Available: http://ptm.asu.edu/modelcard/HP/22nm_HP.pm

[96] Y. Ye, F. Liu, S. Nassif, and Y. Cao, "Statistical modeling and simulation of threshold variation under dopant fluctuations and line-edge roughness," in *Proceedings of 45th Design Automation Conference (DAC)*, 2008, pp. 900-905.

[97] A. Alzahrani and R. F. DeMara, "Non-adaptive sparse recovery and fault evasion using disjunct design configurations," in *Proceedings of the International Symposium on Field-Programmable Gate Arrays*, 2014, pp. 251-251.

[98] K. Zhang, R. F. DeMara, and C. A. Sharma, "Consensus-based evaluation for fault isolation and on-line evolutionary regeneration," in *International Conference on Evolvable Systems*, 2005, pp. 12-24.