# A Design Methodology for Robust, Energy-Efficient, Application-Aware Memory Systems

A Dissertation
Presented to
The Academic Faculty

by

Subho Chatterjee

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2012

# A Design Methodology for Robust, Energy-Efficient, Application-Aware Memory Systems

Approved by:


Dr. Saibal Mukhopadhyay, Advisor
School of Electrical and Computer Engineering
*Georgia Institute of Technology*

Dr. Muhannad Bakir
School of Electrical and Computer Engineering
*Georgia Institute of Technology*

Dr. Sudhakar Yalamanchili
School of Electrical and Computer Engineering
*Georgia Institute of Technology*

Dr. Satish Kumar
George W. Woodruff School of Mechanical Engineering
*Georgia Institute of Technology*

Dr. Abhijit Chatterjee
School of Electrical and Computer Engineering
*Georgia Institute of Technology*


Date Approved: Aug, 28, 2012

To my grandparents

# ACKNOWLEDGEMENTS

First and foremost I would like to acknowledge my advisor, Dr. Saibal Mukhopadhyay for being so supportive during this long journey. Whenever I have required support or have hit a barrier in my research, his doors have always been open for advice. He has been a great teacher, friend, philosopher and guide in its truest sense. I would like to thank my committee members Dr. Sudhakar Yalamanchili, Dr. Abhijit Chatterjee, Dr. Muhannad Bakir and Dr. Satish Kumar for their valuable feedback. I would also like to thank all my co-authors for their support and help. Special mention must be made of Dr. Somnath Paul and Dr. Swarup Bhunia from the Nanoscape Lab (Case Western Reserve University), Mitchelle Rasquinha from the CASL Lab (Gatech) and Dr. Sayeef Salahuddin ( LEED, UC Berkeley).

I express my gratitude to the GREEN Lab members for their co-operation and valuable technical inputs. My gratitude to the GREEN lab members: Dr. Jeremy Tolbert, Dr. Minki Cho, Amit R. Trivedi, Boris Alexandrov, Kwanyeob Chae, Zakir Khondker, Muneeb Zia, Wen Yueh, Denny Lie, Krishnamurthy Yeleswarupu, Swarrna Karthik, Sergio Carlo, Monodeep Kar and Jaeha Kung.

Life outside the campus is an important ingredient of PhD life and in this respect, I have been blessed with a set of seniors and friends who have made my stay in Atlanta memorable. My special thanks to Mrinmoy Ghosh, Atri Dutta, Arindam Basu, Padmanava Sen, Prabir Saha, Saikat Sarkar, Shreyas Sen, Debrup Das, Sabyasachi Deyati, Ananda Barua, Ayan Chakraborty, Piu Chakraborty, Aritra Banerjee, Tapobrata Bandopadhyay, Partha Chakrabarty and others.

Finally, my family has been my pillar of strength during these years. My deepest gratitude goes out to my parents, Mr. Sailen Chatterjee and Mrs. Mita Chatterjee, who have stood by me through thick and thin. I also want to mention my grandparents, late Mr. Bimal Pathak, Mrs. Shanti Pathak and late Mrs. Parul Chatterjee who infused in me the zeal to excel. And to conclude, I express my gratitude to my lovely wife Proma who with her care and love has made this journey memorable.

.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS AND ABBREVIATIONS

| | |
|---|---|
| STTRAM | Spin Transfer Torque Random Access Memory |
| MTJ | Magnetic Tunneling Junction |
| TMR | Tunneling Magneto Resistance |
| ATMR | Array Tunneling Magneto Resistance |
| WL | Word Line |
| SL | Source Line |
| BL | Bit Line |
| DWL | Dual Word Line |
| DSLB | Dual Source Line Bias |
| MBC | Memory Based Computing |
| MLB | Memory Logic Block |
| L1/L2 | Level 1/ Level 2 |
| TSV | Through Silicon Via |
| FVM | Finite Volume Method |
| NEGF | Non Equilibrium Green's Function |
| LLG Equation | Landau Lifshitz Gilbert Equation |

# SUMMARY

Memory design is a crucial component of VLSI system design from area, power and performance perspectives. To meet the increasingly challenging system specifications, architecture, circuit and device level innovations are required for existing memory technologies. Emerging memory solutions are widely explored to cater to strict budgets. This thesis presents design methodologies for custom memory design with the objective of power-performance benefits across specific applications. Taking example of STTRAM (spin transfer torque random access memory) as an emerging memory candidate, the design space is explored to find optimal energy design solution. A thorough thermal reliability study is performed to estimate detection reliability challenges and circuit solutions are proposed to ensure reliable operation. Adoption of the application-specific optimal energy solution is shown to yield considerable energy benefits in a read-heavy application called MBC (memory based computing). Circuit level customizations are studied for the volatile SRAM (static random access memory) memory, which will provide improved energy-delay product (EDP) for the same MBC application. Memory design has to be aware of upcoming challenges from not only the application nature but also from the packaging front. Taking 3D die-folding as an example, SRAM performance shift under die-folding is illustrated. Overall the thesis demonstrates how knowledge of the system and packaging can help in achieving power efficient and high performance memory design.

# CHAPTER 1

# INTRODUCTION

The choice of a memory technology and its subsequent design depends on several factors- the energy constraints, the reliability requirement, and performance constraints. All these constraints arise primarily from the application for which the design is intended and from the packaging constraints. Thus, any memory design has to take these factors into consideration. For well-established memory technologies, adaptation to the application space and packaging induced constraints might result in the alteration of the standard cell parameters and in some cases even the topology. For the emerging technologies, we need a methodology which identifies the application space where it might be most useful. Beyond that point, design optimization will help us meet the energy-performance budget. Thus, application-aware tuning is a critical part of memory system design and depending upon the level of maturity technology might need an application space exploration before actual circuit optimization. In this chapter, we provide a brief introduction to the various memory technologies that are prevalent today. We discuss which memory technology benefits which market segment and application. We try to understand the desirable properties that different technologies have which make them the forerunners in corresponding markets and application spaces.

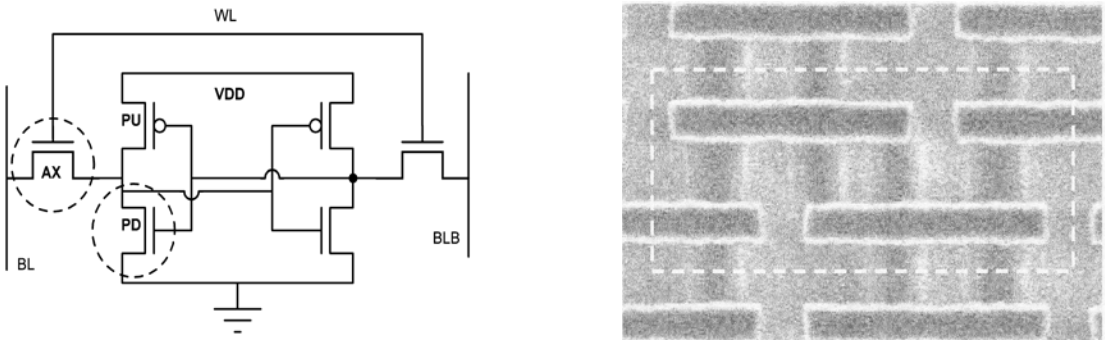
Memory systems are diverse in nature from the application perspective. The application space for embedded memory has expanded over the years, creating memory design opportunities from the microprocessor cache to reconfigurable computing

systems. The design requirements are essentially vast. Increased design complexity and scaling requirements for the embedded memory market has prompted a host of innovations from the memory technology domain to the circuit domain.

## 1.1 Memory Technologies

Let us briefly discuss some of the major memory technologies and talk about what application domains they are most suitable for. The memory technologies can be subdivided into two parts-volatile and non-volatile. Volatile memories can only retain their data when the power supply is switched on. Non-volatile memories retain their data even when the power supply is switched off. Let us discuss them in order.

### 1.1.1 Volatile Memory Technologies




(a) (b)

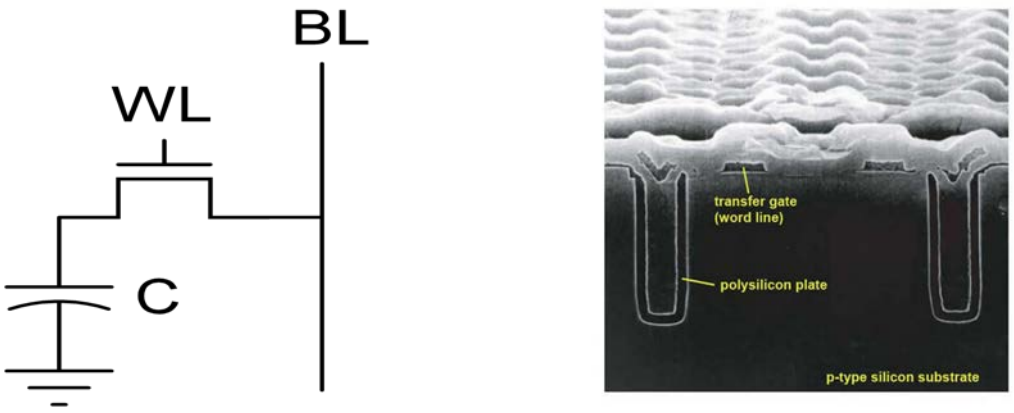
**Fig. 1.1: (a) SRAM circuit (b) 45 nm SRAM layout (courtesy Intel Corporation)**

In this section there are two major components- SRAM (static random access memory) and DRAM (dynamic random access memory). Random access memory signifies the fact that any of the bit cells can be accessed completely randomly. As we shall be dealing with them later we discuss their basics in a bit of detail here.

**SRAM**: A standard SRAM cell consists of six transistors (Fig 1.1 (a) and (b)). It has two cross coupled inverters and two NMOS access devices to sense/write to these inverters. When the power supply is present and access devices are off, the cross-coupled inverter holds the data. While sensing, the bit line (BL) and bit line complement (BLB) are raised to the same potential. The word line (WL) is switched on, turning on the access devices. Depending on whether a "1" or "0" is stored at a node, one of BL or BLB registers a voltage drop. This drop is sensed across a sense amplifier. The small voltage drop is magnified rail to rail. For writing, one of the BL/BLB is maintained at "1" and other at "0". The access path is opened to write the values at the storage nodes.

Proper sizing of the transistors can make SRAM a fast and robust solution. Due to its fast access speed, SRAM finds use extensively in the different cache levels e.g. L1, L2 and L3. The speed-power requirements out of these cache levels are different and the sizing of the transistors varies from one to another. SRAM is also used extensively in FPGA, MBC and other lookup based computations. However, the area penalty for using SRAM is quite large. A typical 6T cell macro layout occupies 100-120 $F^2$ where F is the feature size of the technology. Hence, in applications where area compaction is more crucial than speed (e.g. memory or storage) SRAM is not the preferred memory technology.

SRAM read, write and hold imply contrasting requirements on transistor sizing. Also all these transistors are subject to global and local process variations. Designing the macro at lower technologies to meet the system specifications in presence of ever increasing process variations is a huge challenge. Hence, 8T and 10T cells with the intent to decouple read and write have come up as an alternate robust, low power option.




**(a)** **(b)**

**Fig. 1.2: (a) The 1T DRAM cell (b) A DRAM cell with trench capacitance [3]**

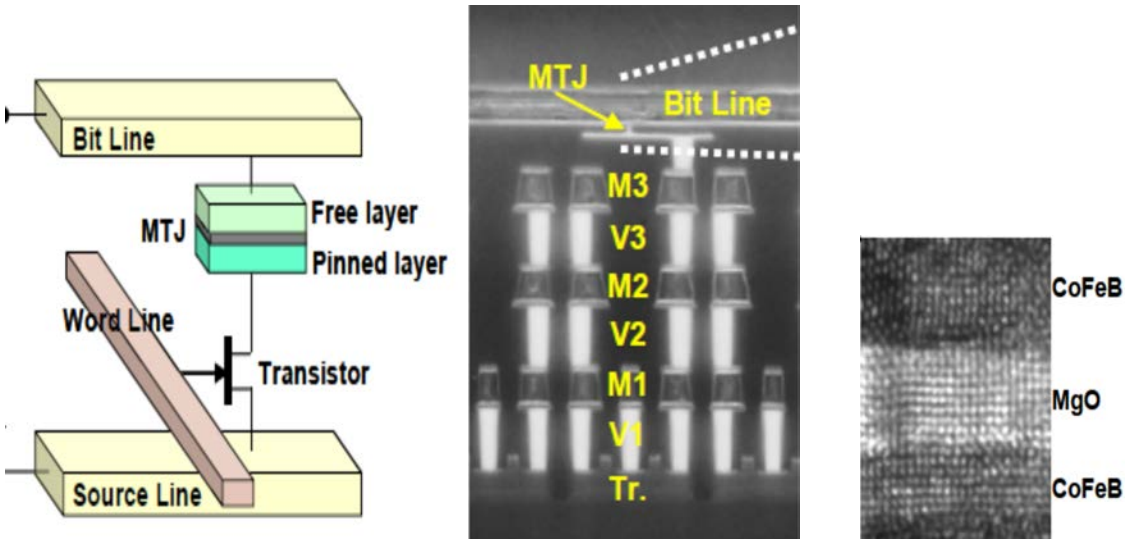
**DRAM**: A DRAM consists of a single cell and a capacitor (usually trench capacitor) where the charge is stored. It can be used to store a value of "0" or "1". Because the DRAM cell consists of a single transistor, DRAM is extremely area efficient. But it is some 10-100X slower than SRAM. Area of a DRAM cell is usually 6-10 $F^2$.However there is one major problem. DRAM gets its name from the fact that the storage node has to be written to or refreshed dynamically. The periodic refresh operation demands additional power. Hence it is clear that power wise DRAM is a more expensive solution. Also reads in DRAM are self-destructive meaning every read there is a write to the same location meaning power and performance overheads. This charge sharing occurs as the bit line capacitance and cell capacitance share their charge thereby destroying the original voltage in the cell. Also as the sensing is single ended unlike in SRAM, the response time is larger. Hence, DRAM finds usage as the microprocessor memory where a high integration density is the prime requirement [5, 6], for which a certain amount of performance and power penalty can be accepted.

### 1.1.2 Non-Volatile Memory Technologies

In the non-volatile memory space, cost per bit is an important factor. A segment of non-volatile memories are intended for storage. Here performance is greatly sacrificed for storage density. Magnetic memory and flash are the representatives of this type of memory. Then there are some newer technologies like STTRAM, FeRAM, PCRAM, ReRAM etc which could be a viable alternative for the embedded memory domain. Each has its relative advantages and disadvantages. We will briefly describe flash and move on to the emerging technologies. With the rapid growth of the embedded and hand held market, the application space for these memories is ever increasing. For our study, we select STTRAM from this domain. We will very briefly put forward the relative advantage and disadvantage of each and only discuss STTRAM in detail.

**Flash**: For NAND flash memory, cell size can go down to as low as $5F^2$. This coupled with non volatility and "not so high" write (0.1ms) and read times (50ns) make it a good choice for storage. However, during erase procedure it requires a high voltage (16-20V) and consequently the write power consumption is very high. Due to these later properties, flash is restricted to the external memory market and is not a candidate for the embedded market.

**Emerging Memories**: We have put forward several memory technologies in the emerging memory technology umbrella. Let us begin with Spin Transfer Torque RAM (STTRAM). Spin polarized electrons are used here to alter the state of the spin content of a magnetic tunneling junction (MTJ). For different spin states, the resistive contributions of MTJ are different. This resistive signature is used for bit storage. In phase change




**(a)** **(b)** **(c)**

**Fig 1.3: (a) STTRAM circuit (b) Cross sectional SEM image of STTRAM (c) Cross sectional TEM-image of MTJ [all pictures courtesy: [2]]**

RAM (PCRAM), the change in resistive signature is brought about by changing the crystalline structure of chalcogenide glass. The change is brought about by a high current pulse of a certain pulse width. Ferroelectric RAM (FeRAM) uses ferroelectric material to achieve non volatility. The ferroelectric material polarity changes on applying electric field across it. In ReRAM a current pulse can change the resistance signature much like its PCRAM counterpart with lower current requirements. Let us discuss about STTRAM in a greater detail at this point.

**STTRAM**: We have talked about the STTRAM mechanism very briefly. Let us now try to understand in greater detail the physics behind the operation. Electrons have a property called spin. Depending upon the direction, the spin could be +-1/2. However, in unpolarized current flow, the spins balance out. When unpolarized current passes through the fixed layer in an MTJ, it becomes polarized. On reaching the free layer, the polarized current exerts a magnetic torque which causes the magnetic elements to flip.

The projections for STTRAM are a cell size of 6-20$F^2$. A read and write time requirements of 2-20 ns in the precessional regime and a high endurance. The drawback of STTRAM is that increased switching current density (1$MA/cm^2$) requirement is becoming a bottleneck for cell area scaling. Added to it is the problem of thermal stability where to retain data for a certain number of years, the temperature has to be below a certain prescribed value. Also, there are problems of excess process variation in a developing technology. Hence, STTRAM has its own set of advantages and disadvantages.

## 1.2 Thesis Statements and Contributions

From the discussions we conclude that the embedded memory space is an area of growth and both volatile and non-volatile memories could find use there. Our first task is to identify an embedded memory application.

We choose a lookup based computation framework called memory based computing (MBC) to illustrate how memory solutions can be tailored for application characteristics. Memory based computing is a spatial computing style which is getting significant attention at present [7, 8]. As technology scales, FPGA systems are not being able to leverage the benefits of scaling as interconnect delay becomes the major performance bottleneck. This delay component does not scale with technology appreciably. MBC tries to be superior to FPGA by adopting larger LUT, less interconnect and multi-cycle operation inside each component.

In the volatile domain, we use SRAM as a test vehicle to find out how application characteristics and packaging solution impacts the design choice. Embedded memory

systems have traditionally suffered from increasing leakage and robustness concerns, especially with scaling technology generations. A lot of research emphasis in recent times has been laid on the usage of scalable non-volatile technology solutions in the embedded space to reduce leakage, and thereby improve energy-efficiency and robustness. However, achieving the suggested end goals require a consistent effort in modeling the energy expenditure and reliability challenges for the non-volatile domain. The knowledge gained thereby has to be incorporated in the design methodology to make it efficient. Thus to summarize, modeling methodologies need to be developed for emerging technologies to investigate the power-performance-robustness achievable for a given application. For our study, we use Spin Transfer Torque Random Access Memory (STTRAM) as a prototype for emerging scalable non-volatile memory solution to build energy-reliability models and suggest design innovations. In particular, the contributions of this thesis can be summarized as follows:

**Application Specific STTRAM Design Methodology**: In this phase of the work, we identify the crucial STTRAM cell design parameters. The information from the cell level are abstracted to create an array level power-performance model. This model can be used to decide how suitable STTRAM is from a given application perspective. This section proposes a design methodology involving a design space exploration of the STTRAM space which helps us take architecturally correct decisions.

**STTRAM Reliability Evaluation**: Before STTRAM is adopted for an application or package, it is necessary to know its thermal reliability. The data storage longevity, the read and write failures are all related to this. Also a large write current flows through the small MTJ device. Hence self-heating becomes a crucial determining factor for thermal

stability. A detailed self-heating analysis is performed across STTRAM and impacts are studied on detection reliability and different other crucial cell parameters. The study is extended to foresee how the nature of application might determine the reliability behavior.

**Circuit Solutions for STTRAM**: Thermal reliability points to the growing concerns for detection reliability which is magnified in presence of process variation. The thesis discusses a circuit solution to reduce its impact on the STTRAM array. We use a source biasing scheme where the STTRAM bit lines and source lines are held at different voltages for reads and writes to ensure minimal leakage for unselected cells of selected column. Results show considerable suppression of the detection failure rate. A design optimization between the dynamic energy overhead and failure rate helps us choose the operating point.

**Design Modifications for STTRAM in MBC**: After the evaluation of the emerging technology and developing an application-aware methodology, we test it for a specific embedded application (MBC). We abstract the degree of read intensity of the application. The design methodology suggests the device sizing and voltage levels based on the abstracted information. The energy model also suggests changes at the software level that could potentially lead to energy profitability.

**SRAM design modification in MBC**: Beyond this point, we take up the case of SRAM design modification in MBC. The main aim is to achieve voltage scalability and thereby power savings. An alternate cell structure is adopted which protects the storage node from the bit lines thereby eliminating the read reliability concerns. Depending on the size

of the read transistor, the read current increases and consequently, the access time can come down quite a bit. However, both read and write are single ended. Hence additional measures are necessary to ensure correct operation. The macro topology, sizing constraints, failure mechanisms for the cell etc. are discussed in detail. The idea is verified at an HSPICE simulation level and the impact is evaluated across a MBC platform using architectural simulations. A single MCB (memory cell block) prototype has been designed in IBM 130nm process with the above mentioned memory cell structure. We present extracted results from the fabricated chip.

**Stacking induced SRAM constraints in 3D**: Both application and packaging are crucial determining factors in determining the SRAM design constraints. In this case, we take 3D stacking of SRAM as a packaging platform to determine how the SRAM design requirements might be modulated. A core-cache stack with its increased vertical coupling presents a degraded worst case temperature across the SRAM cache. This gives rise to increased leakage, read failures, hold failures and enhanced aging. Study and analysis of this degraded cell design margins is necessary to prevent failure for such stacking.

The thesis is organized as follows. In chapter 2, we discuss the prior works pertaining to each of the topics mentioned. Chapter 3 talks in detail about the STTRAM design methodology. The self-heating induced STTRAM temperature distribution and its impact evaluation is done in Chapter 4. Chapter 5 talks about a circuit solution to reduce the detection failure in STTRAM. In chapter 6, design modifications in STTRAM for applicability to MBC are discussed. In chapter 7 SRAM cell custom design for MBC along with the post extraction results are discussed in detail. Chapter 8 gives us a preview

on the 3D stacking induced design challenges to SRAM. Finally, Chapter 9 summarizes this thesis and points to future directions where the work might be extended.

# CHAPTER 2

# PREVIOUS WORK

This chapter discusses some of the previous works in this domain. The discussions on the prior works begin with the STTRAM technology and documents efforts aimed at showing its utility in the embedded domain. Studies evaluating and characterizing the thermal distribution in STTRAM cell are discussed. At each such stage, we discuss the scope of future improvement and the challenges that exist. Temperature development in the cells and the arrays can lead to parametric failures. Thermal impact on such parametric failures is an area of key concern. For adoption of a technology a complete design infrastructure needs to be in place. Hence, design space exploration of this emerging technology has received widespread interest from the design and automation perspective. Evaluation of power-performance trade-offs possible in different applications is another such important area of study. Application-aware design demands a lot of in-depth understanding of the circuit realizations. This is true of maturing and mature technology alike. A brief literature survey is conducted on the application-aware design in SRAM. Besides, the role of packaging in determining the memory design is investigated. Case studies along this direction specifically for memories are presented for motivation of later part of the work.

## 2.1 Maturing of STTRAM Technology

Spin Transfer Torque based Random Access Memory is an interesting and important entity in the memory market. The theory for magnetic coupling existing between two

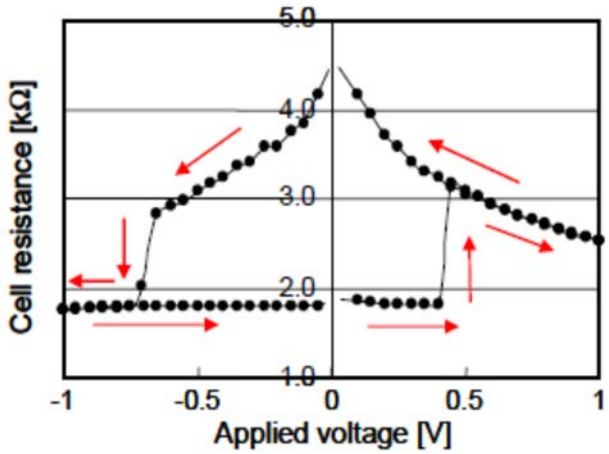



**Fig 2.1: STTRAM R-V characteristics [2]**

barrier separated ferromagnets was proposed by Slonczewski [9]. This existence of the magnetic coupling in ferromagnetic materials would make state storage and alteration possible. This opened the avenues for spin-torque assisted switching and propelled spin based MRAM as a contender in the embedded memory market [10, 11]. To make STTRAM a usable memory technology, a number of goals remained to be achieved. The programming current was of the order of mA and clearly needed to be brought down to avoid circuit complexity. Secondly, Tunneling Magneto Resistance (TMR), which is a ratio of the high and low state resistances, was 20-50%. This difference was clearly not enough for a reliable sensing, especially at the variation prone submicron technologies. Researchers devoted themselves to the solution of these fundamental shortcomings. Structures and materials of the MTJ were widely experimented with to improve these properties [12]. The solution suggested was in the form of a MgO tunneling barrier based STTRAM, which could increase the TMR to 150-200% while maintaining a current density of $1.1 \times 10^{6}$ A $cm^{-2}$ [13]. Once these problems were waived, the next step was to verify CMOS integration compatibility [14, 15]. In CMOS integration the MTJ is usually integrated at the higher level of the metal and is connected to the access device through

via. As per the projections in [10], STTRAM could go down to as small as 4 $F^2$ area with current prototypes having 6-20 $F^2$ area. Also under Time Dependent Dielectric Breakdown (TDDB) test environment under voltage and temperature stress, STTRAM shows a high write endurance (>$10^{16}$) [16]. With these developments, STTRAM has entered the mainstream memory market slowly. Currently it is used in avionic and space applications where high levels of radiation make charge based storage infeasible. It is also being applied to embedded systems in a variety of situations [17].

## 2.2. STTRAM Design Space Exploration and Need for Application-Aware Methodology

As the STTRAM technology matures, new potential areas of application of the technology are investigated. This requires tools and design methodologies for the technology to be in place. Hence the area of STTRAM-specific design methodology and tool design has seen a rise in research interest [18]. At the architectural level, trade-offs between power and performance have been studied for application to cache by Xiaoxia et al [19] while that between endurance and speed have been explored for energy-efficiency by Smullen et al [20]. Modeling of a self-consistent framework trying to solve the Non Equilibrium Green's Function (NEGF) and the Landau Lifshitz Gilbert (LLG) equation is the core to model the MTJ [21, 22]. The MTJ has been modeled at different levels – from the device level models to the behavioral models [23-25]. Based on the STTRAM device models, a large number of works aimed at design space explorations involving circuit parameters [26, 27]. Design methodologies to achieve a desired yield estimate in presence of variations for such circuits have been explored [28, 29]. These proposed design methodologies and tools aim to make this maturing technology usable from

different design perspectives – power, performance, yield etc. However, the formalization of the understanding of design alterations to be performed for application-specific systems is still missing at large. For example, depending on whether read operations might have precedence over write operations, the design can be tailored for achieving energy efficiency. In order to achieve that, the detailed STTRAM array level energy model has to be built while accurately understanding the constraints imposed by the read and write requirements.

### 2.3. The Issue of Endurance and Need for Thermal Distribution Evaluation

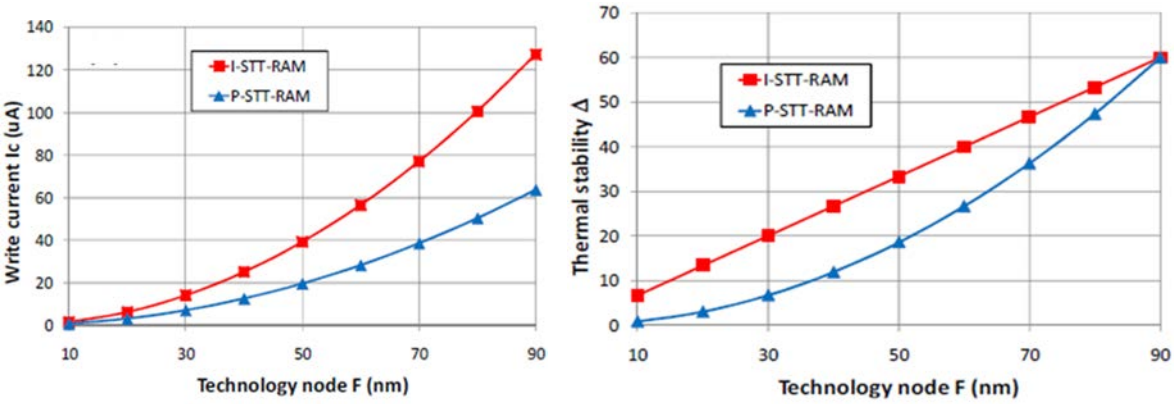
The pulse width and programming current for write in STTRAM are interrelated [14]. If the write time requirement is restricted, the programming current has to be exponentially increased in the precessional operational region for MTJ. However, the flip side to increased speed is an increased temperature. In presence of elevated temperature, the endurance of the cell is sacrificed. Hence, there is a direct trade-off between writing speed and data storage lifetime. The thermal stability factor is linked to the anisotropy of the storage node. It is defined as:

$$\Delta = \frac{K_U V}{k_B T}$$

where Ku is the anisotropy energy density and V is the free layer volume respectively. The thermal stability factor for non-volatile applications is expected to be greater than 60 [27]. A higher temperature hinders the thermal stability factor. Also, the switching characteristics are a function of temperature [30]. The initial magnetization vector distribution bears an impact of temperature. This introduces an additional asymmetry in

the P-AP and AP-P switching mechanism. Characterization of the STTRAM thermal distribution was performed and it was found that for a 1GB STTRAM to achieve a 10 year data retention and 1000 FIT read disturb error rate requires a thermal stability of 75 [16].

All the above mentioned facts point to the fact that thermal distribution evaluation is an important aspect of STTRAM circuit evaluation. Another interesting fact is that while most works focus on the MTJ behavior with temperature, little attention has been paid to the transistor behavior. However, for a complete analysis of the STTRAM cell and array, the MTJ and transistor impacts need to be coupled together. Towards that end, there is a need to incorporate some key understandings from the STTRAM switching studies [31, 32] and submicron technology transistor behavior [33].




**(a)** **(b)**

**Fig 2.2: Scaling projections in STTRAM (a) Write current scaling projections with technology [1](b)Thermal stability scaling projections with technlogy [1]**

## 2.4. Design Issues in STTRAM and Solutions

Having achieved a considerable impetus from the technology and materials aspect, the future of STTRAM will depend on the ease of CMOS circuit integration. It has been proven that the STTRAM complies with CMOS integration and this has been verified at submicron technology nodes [15]. Earlier works concentrate on achieving a standard

performance out of the STTRAM array of smaller size. A 10ns read and write pulse width based STTRAM array was demonstrated in [34]. Larger sized arrays were developed [35, 36] with different connection styles as well [37]. Faster read times (4ns) were achieved for larger sized arrays as the technology matured [38]. Writes in STTRAM are energy expensive. To improve on the write energy efficiency, early write terminations based on detection was suggested in [39]. The read operation is crucial from the reliable operation perspective in STTRAM. In presence of process variations impacting the MTJ and the access device, sensing becomes even more difficult. Techniques simultaneously utilizing the magnetic and the circuit properties are required for reliable sensing [40]. To this end, self-referencing schemes sacrificing some performance have been proposed [40-42]. Negative resistance shunting of STTRAM cell for read operation was also suggested for read reliability [43].

STTRAM is a zero leakage system. This is one of its big advantages and hence circuit designers do not have to use additional power saving schemes (gating techniques) as in SRAM or DRAM. However, STTRAM detection mechanism can be impacted by leakage. Leakage from unselected cell of selected column can contribute to offsets. This can lead to wrong judgment at the sense amplifier output [44]. This is more severe as in maturing technologies like STTRAM impact of process variation is quite acute. This problem needs to be thoroughly analyzed and solutions have to be suggested to improve read reliability.

### 2.5. An Embedded Application Platform - Memory Based Computing

The design methodology we proposed earlier is applicable to a wide range of applications. The effectiveness of the methodology can be showcased particularly for

read intensive embedded applications. Memory based computing (MBC) is an application platform to this end. Let us discuss a bit about MBC before moving on to the non-volatile memory based findings in this application.

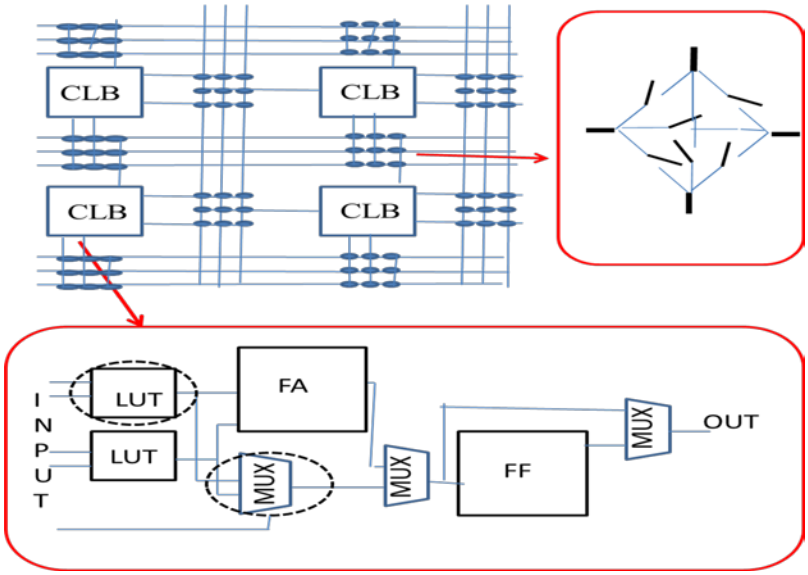



**Fig 2.3: FPGA structure**

Field programmable gate array (FPGA) is a spatial computing approach where computations are performed in computing units called configurable logic block (CLB). The CLBs consist of look up table (LUT) composed of smaller look up table modules and usually a full adder corresponding to the mathematical mode of operation. A D flip-flop is present at the end of the structure. The CLB can be synchronous or asynchronous in its mode of operation. These units are connected by a set of programmable interconnects. The programmability is achieved in the switch network of the interconnect fabric. These switches are typically composed of tri-state buffers. However, the interconnect overhead grows at a huge rate with the complexity of the function and the number of inputs. Thus interconnects become the prime source of delay and power dissipation in FPGA. At submicron technologies the interconnect delay does not scale as well as the CMOS delay.

Hence the problem aggravates and interconnect delay dominates over the rest and this becomes a bottleneck to performance betterment in FPGA in nanometer nodes.

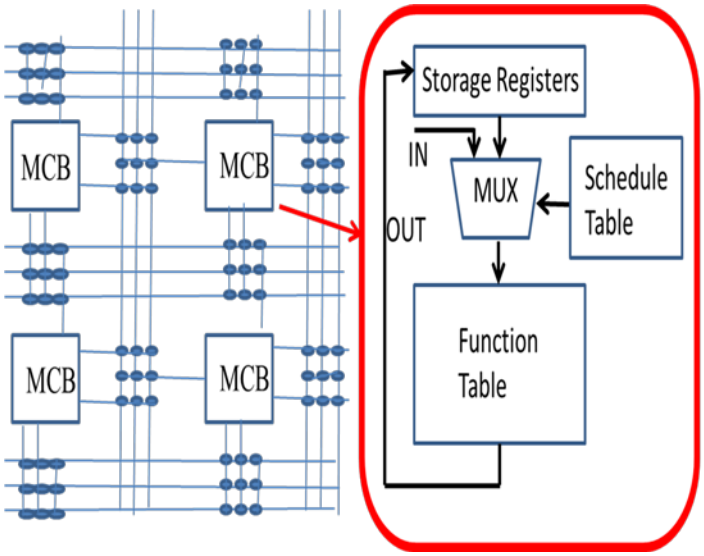



**Fig 2.4: The MBC system**

An approach that tries to bypass this additional interconnect delay is to have a larger 2D memory array for the look-up table and minimize interconnect between units. The units are made capable of performing multi-cycle operations and hence less number of such units is to be used. This means reduced interconnect delay dependency and consequently reduced interconnect delay sensitivity for the system.

Multi-input multi-output LUTs are used in MBC for storing the partitions of the target application. The unit computing element is called memory computing block (MCB). Information on the address, scheduling etc. are stored in a finite state machine called schedule table. The schedule table is implemented using flip flops. The mapping into the LUT (memory) as well as the schedule table is done during the application mapping phase. During the operational phase, the address selection for the lookup is done through a multiplexor tree network. The select signals to the MUX tree arrive from the

schedule table. Based on the signals, input is selected from the available choices- the external inputs or the results stored in the MLB from earlier computations. Based on the computed memory address, the lookup is performed. The lookup results are stored in a register file called the intermediate register. The results from previous cycles are stored in the dummy registers. These can be used towards multi-cycle operations.

The 2D memory is known as the function table. The MCB input is limited by the function table size. However, making the function table large is not a good prospect from the area and memory access time perspective. For MCB of partition inputs 12 and outputs of size 4, memory size of 2kB has been considered [45]. The data read out in the current cycle (i) is stored in intermediate registers for use in the next cycle (i+1). For use in cycles i+2 and beyond, the data is stored in dummy registers. However, the increased number of such registers means a more complicated MUX tree. This in turn means the schedule table has to supply an increased number of select signals which complicates its design. Also the data read out of the function table may be transmitted to neighboring MCBs as well.

MBC has received a lot of interest in the research fraternity in recent years [8]. As the 2D array is the heart of the MBC, it has been subjected to a lot of exploration. Different memory structures have been tried for power and performance benefits [7, 46]. The program once, read multiple times environment favors the usage of non-volatile memory. Hence, STTRAM based hybrid memory have been shown to be useful for such applications [47]. However, the design question as to how to design the STTRAM memory cell and array for this kind of an application to achieve power performance benefits had not been addressed.

MBC design is usually performed currently with six transistor SRAM cell. Read related parameters receive priority in SRAM design for MBC. To achieve scalability and read reliability, MBC specific transistor sizing was adopted with limited benefits [48]. A 6T SRAM topological modification in the SRAM was suggested to perform better for low-voltage and high-speed applications [49, 50]. However, some critical issues remained to be addressed for the structure which would hamper its power and performance. One of the major concerns was to stop the flow of current from the unselected cells in the selected column. This required architectural as well as circuit level changes for the cell organization.

### 2.6. New Packaging Environment-3D and the Design Constraints Imposed

3D integration of chips has become necessary to adhere to Moore's law. In this integration, dies are stacked on top of each other and inter-die connection is established through metallic through-silicon-via (TSV). Memory stacking is especially of high interest since the high number of TSVs provides a huge bandwidth to be utilized. The issues of memory stacking are very relevant to industry and memory alike.

High temperature is an unavoidable phenomenon for 3D systems. With limited heat escape routes (one heat sink and the package for multiple dies), the temperature rise in the dies away from the heat sink is a major cause of concern. As the processor is likely to dissipate a much larger (~90% of heat) [51], the processor is likely to be placed near the heat sink. The memory operation temperature is expected to be higher than in a single die.

There have been numerous efforts in trying to model the temperature rise in a 3D environment [52]. A lot of these efforts have stuck to the simple resistive heat flow model

for temperature estimation [53, 54]. The initial objective of these efforts is to understand the thermal dynamics in a 3D integrated system [54]. This analysis needs to be performed across multiple materials, stacking style and geometry to arrive at the optimal condition [55, 56]. The estimated temperature is used to suggest micro architectural and process technology based techniques to reduce the undesired impact of temperature [57-59].

3D cache stacking with SRAM is a reality today [60]. Hence, it is open to architectural as well as design considerations to extract the best performance out of SRAM cache [59]. However, SRAM is extremely sensitive to process and temperature variations [61] and the yield might suffer under the stacked scenario. In the stacked scenario, temperature also behaves as an added source of variation. Also, a consistent high temperature is favorable to aging mechanisms [62], resulting in further increases in bit error. Hence, before stacking it is necessary to get a good estimate of the design estimates to avoid the impact of failures due to the raised temperature. This understanding is the goal of our work.

The following chapters discuss in detail the research done in each of these areas. The next chapter begins with the STTRAM energy model development and proposal of a design methodology for energy-efficient STTRAM based design.

# CHAPTER 3

# STTRAM ENERGY MODEL: TOWARDS AN ENERGY EFFICIENT NON-VOLATILE MEMORY

Due to its non-volatile nature and very low standby power STTRAM arrays are suitable for embedded applications. Further, very small cell size can help to improve the integration density and hence, allow a larger memory capacity relative to the SRAM memory for the same area. A larger size for example reduces the L2 miss-rate and consequently reduces off-chip memory accesses, leading to better performance and lower energy consumption. One of the primary challenges in using STTRAM is a correct evaluation of the energy dissipation during read and write operations. Hence, detail analysis and co-optimization at both the circuit and architecture level are necessary to understand how to design the STTRAM cell (given a MTJ) to minimize the energy dissipation for an application. The first step towards that goal is the detailed realization of the STTRAM energy model at the array level.

In this chapter, we present a detail analysis of the energy dissipation of an STTRAM array and explore the cell design space constraints that should be taken care of while minimizing the array energy dissipation. For the results we use in 180nm TSMC technology [63]. In particular, we do the following:

1. A comprehensive analysis of the array energy is performed using a detail dynamic model of the STTRAM array.

2. A methodology to co-design the cell access transistor and operating voltage to minimize the energy dissipation of the array while maintaining cell quality.

## 3.1. Cell Metrics and Operations

The two important quality metrics for an MTJ are: (a) the TMR and (b) the switching current ($I_C$). At a particular operating temperature, both of these factors depend on the structure and material of the MTJ device. TMR is the readability metric and is defined as [64]:

$$TMR = \frac{R_H - R_L}{R_L} \qquad (3.1)$$

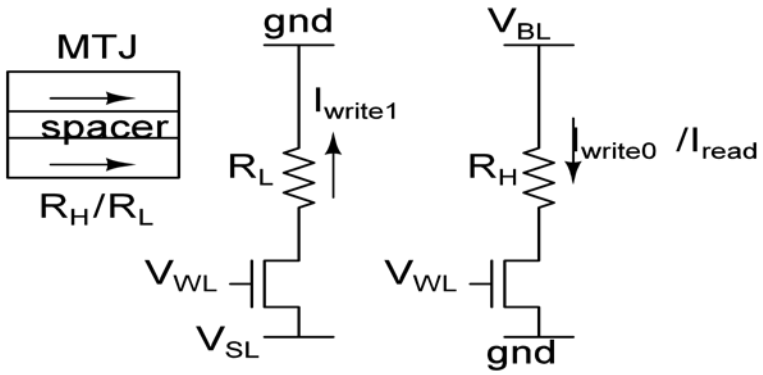



**Fig. 3.1: MTJ structure and read/write operations in STTRAM cell**

where $R_H$ and $R_L$ denote the resistance offered by the MTJ in anti-parallel and parallel orientations. A high TMR indicates a larger difference between resistances in the parallel and anti-parallel modes and facilitates reading. A high TMR is hence desirable.

However, for an STTRAM cell, equivalent quality metrics need to consider the effect of the access device. In this section, we analyze the effect of access device on these qualities and define the regions in the array operating voltage ($V_{WL}$) and access device width (W) plane that will satisfy the cell quality requirements (Fig. 3.1). In this analysis, we assume that, the word-line voltage during read/write operations and bit-line/source-

line voltage (for parallel/anti-parallel writes) during write operations are same (defined collectively as the array operating voltage, $V_{WL}$). The bit-line voltage during read operation ($V_{read}$) is different. This ensures that, only two voltage levels are required for the array.

### 3.1.1 Read operation: Cell TMR

During read the bit line is pre-charged with a small voltage ($V_{read}$) and current flows in the direction of bit line (BL) to source line (SL) (Fig. 3.2). The errors in read operation are classified under two categories, (a) false read and (b) read disturb. False read occurs when the output sensed in different from the state of the MTJ. Read disturb occurs when in trying to read the state of the MTJ is flipped. In this subsection, we study the read operation metrics, their relation to the read errors and validate our analysis with simulation results from STTRAM simulations using TSMC 180 nm.

False read is related to the quality of read operation. The quality of read operation of the cell depends on the difference between the current flowing through the cell while reading "1" (i.e. anti-parallel state) and reading "0" (i.e. parallel state). The higher this difference, the more robust is the circuit against false reads. The cell TMR (CTMR) can be defined as:

$$CTMR = \frac{I_{read\,0} - I_{read\,1}}{I_{read\,0}} \qquad (3.2)$$

where $I_{read0}$ and $I_{read1}$ are the read currents for the parallel or "0" and anti-parallel or "1" cases respectively. For a given access transistor, since the resistance in the anti-parallel ("1") state is higher than the resistance in the parallel state, $I_{read0} > I_{read1}$. From simple circuit analysis equation (3.2) can be interpreted as:

$$CTMR = \frac{R_H - R_L}{R_H + 1/kW(V_{WL} - V_{th})} \qquad (3.3)$$

where $R_H$ and $R_L$ are the resistances of the MTJ in the anti-parallel and parallel states; W is the width of the access device; $V_{th}$ is the threshold voltage of the access device; $V_{WL}$ is the word-line voltage (i.e. array operating voltage). For equation (3.3), the device length is incorporated in the constant k (assuming minimum device length for the technology). Since the voltage drop across the transistor is expected to be small we have neglected the 2nd order term of drain-to-source voltage across the transistor in equation (3.3). However, the circuit simulation based results presented in the paper do consider this effect. It is important to note that $R_H$ and $R_L$ actually vary with the voltage across MTJ for large voltage swing. In our case the voltage swing across MTJ permits variation in $R_H$ ~2-5%. Hence it is logical to assume constant $R_H$ and $R_L$ values. In order to distinguish the currents for reading "0" and "1" the CTMR needs to satisfy a minimum bound (say, $\eta_{CTMR}$) governed primarily by the robustness of the sensing circuit. From equation (3.3), we can obtain the region in the $V_{WL}$-W plane that will satisfy this minimum CTMR requirement:

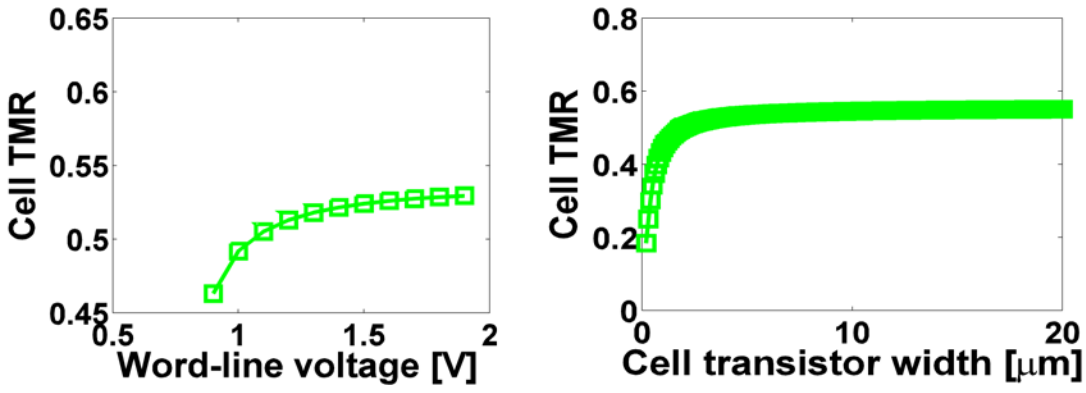
$$W(V_{WL} - V_{th}) \geq \frac{1}{k\left(R_H - R_L\right)}\left(\frac{1}{\eta_{CTMR}} - \frac{1}{TMR_0}\right)^{-1} \qquad (3.4)$$

where, $TMR_0$ represents the TMR of the MTJ given by equation (3.1) and is a property of the MTJ. From equation (3.4) we can clearly observe that, a higher word-line voltage and larger access device help to improve CTMR. Fig. 3.2 shows the effects of W and $V_{WL}$ on CTMR considering TMR value obtained from using $R_H$= 4.5K, $R_L$ = 2K at 180nm technology [2].

The read disturb occurs when the read current is higher than the MTJ switching current which cause the cell data to flip during reading. We observe that flipping of a "0" during read is not a possibility. This is because the read current flows from BL to SL. Hence read disturb is limited to flipping of a "1". The read current needs to be sufficiently smaller than the write switching current for the MTJ to reduce the probability of the read disturb (i.e. switching of MTJ state while reading it). On the other hand, a minimum value of read-current (say, $I_{min}$) is required to ensure that it can be sensed properly even under noise in the sensor. We consider these effects using the following criteria:

$$I_{\min} < I_{read1} = \frac{V_{read}}{R_H + 1/kW(V_{WL} - V_{th})} < \eta_{RM} I_C \qquad (3.5)$$

where $\eta_{RM}$ is the minimum ratio between switching current and read current. The ratio $\eta_{RM}$ quantifies a design margin for read such that even under variation (which might increase $I_{read1}$ beyond its expected value) read disturbs do not occur. Hence, we refer to $\eta_{RM}$ as the read margin. Equation (3.5) helps to determine the $V_{read}$ voltage that should be used while reading a cell.

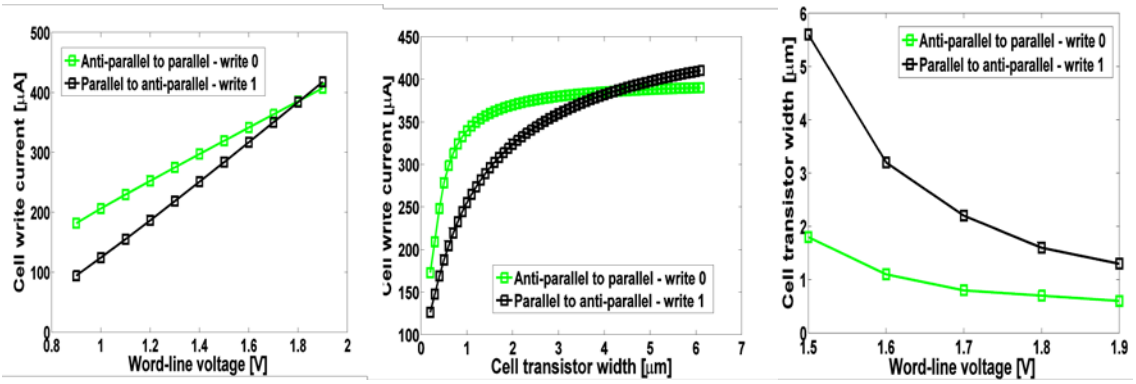



**(a)** **(b)**

**Fig. 3.2: Variation in CTMR with design parameters (a) cell word line voltage, and (b) transistor width.**

### 3.1.2 Write operation: Cell Switching Current

For an MTJ to switch the current through the MTJ must be greater than a critical current $I_C$ and a higher switching current is required for faster switching. Given the MTJ switching current, the cell switching current ($I_{cell}$) strongly depends on the access device and the direction of switching. During this operation $V_{WL}$ and $V_{BL}/V_{SL}$ are tied to the same voltage (depending on whether it is parallel/anti-parallel write).



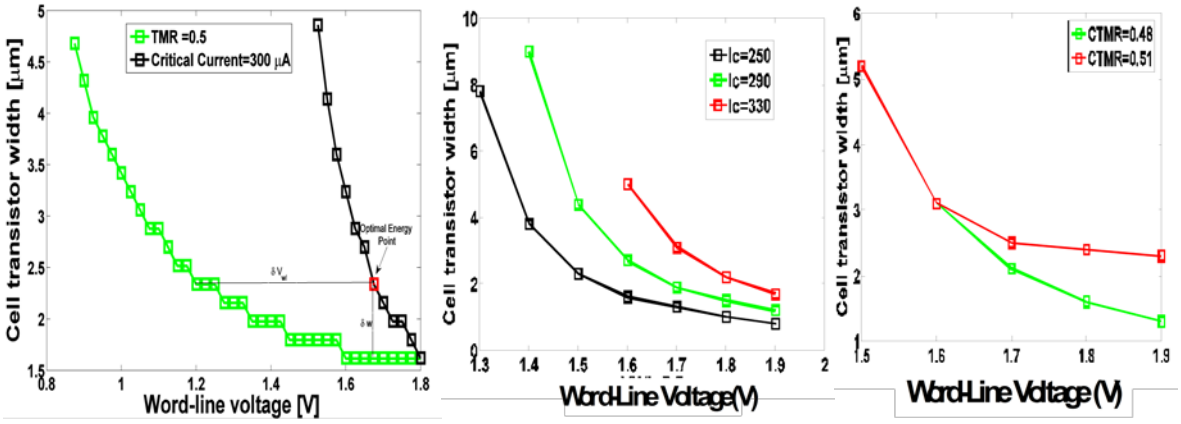

**(a)** **(b)** **(c)**

**Fig. 3.3: Effect of $V_{WL}$ and W on cell switching current: write current variation with (a) $V_{WL}$, (b) W, and (c) $V_{WL}$-W plane for $I_C$=0.3mA.**

Anti-parallel to Parallel Switching – Write "0": During anti-parallel to parallel switching, the current flows from bit-line to source-line. During this operation, the MTJ acts as a resistive load to the transistor. Hence, $V_{ds}$ of the access device is low and the device remains in the linear region. Further, during this operation, the initial cell resistance is $R_{AP}$ or $R_H$. We obtain the region in $V_{WL}$ – W plane that allows $I_{cell} > I_C$.

Parallel to Anti-parallel Switching – Write "1": During parallel to anti-parallel switching, the current flows from source-line to bit-line. The MTJ acts as a resistor at the source of the transistor thereby reducing its effective $V_{gs}$. The transistor remains in the saturation region (as $Vg=Vd=V_{WL}$). Further, during this operation, the initial cell

resistance is $R_P$ or $R_L$. We can obtain the region in the $V_{WL}$ – W plane that will allow $I_{cell}$ > $I_C$ for writing "1".




**(a)** **(b)** **(c)**

**Fig 3.4: $V_{WL}$-W plane with (a) CTMR=0.5, $I_C$=300uA, (b) variation in CTMR, and (c) variation in $I_C$**

Fig. 3.3(a) and 3.3(b) shows the current through the cell for different W and $V_{WL}$ for write "1" and write "0" conditions. Note, for write "0", increasing the width beyond a certain point does not have strong impact on the current as the device acts as a resistor and current is limited by $V_{WL}/R_H$. But, for write "1" increasing W always increases the current as the device is in saturation. For reasonable voltage ranges, for a given $V_{WL}$, the minimum device width required for write "1" is larger. This is because the increase in the source voltage reduces the effective Vgs of the transistor. However, for small $V_{WL}$ (i.e. when $V_{WL}/R_H \rightarrow I_C$) the higher width is required for write "0". The operating region in the $V_{WL}$-W plane to ensure a required cell switching current can be obtained by simultaneously considering the two write operations (Fig. 3.3(c)).

### 3.1.3 Design Space for Read and Write Operations

Proper choice of $V_{WL}$ and W are required to simultaneously satisfy both the cell TMR and switching current requirements [4]. Essentially, CTMR target provides a lower bound

on the allowable $V_{WL}$ and W values. From Fig. 3.2 we observe that a higher CTMR targets requires a higher $V_{WL}$ and/or W values. Hence, a higher CTMR target reduces the acceptable $V_{WL}$–W plane as illustrated in Fig. 3.4(c). Fig. 3.4(a) also shows the acceptable $V_{WL}$-W plane satisfying a critical current target ($I_C$=0.3mA). The device width for a given $V_{WL}$ is determined considering the method discussed earlier. Fig. 3.5 shows the acceptable regions in $V_{WL}$-W plane for a given $\eta_{CTMR}$ (=0.5).

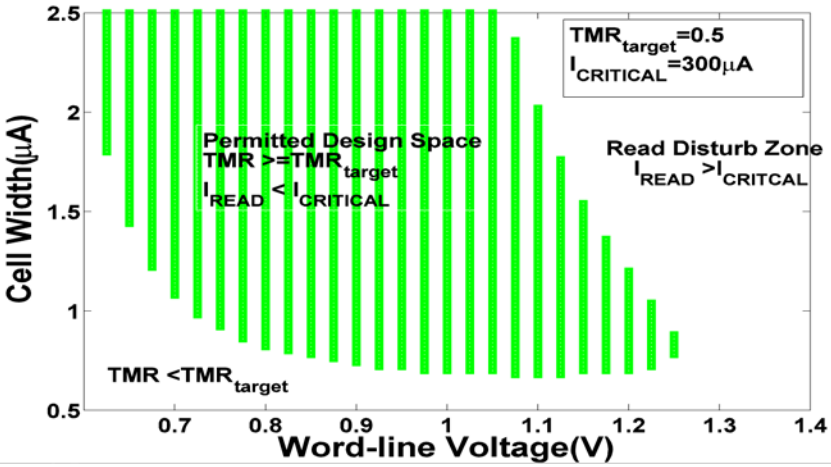



**Fig 3.5: Allowed design space for TMR>0.5, p=0.5 at Vread=0.6V**

As expected a higher critical current requires a higher $V_{WL}$ and/or W value and hence reduces the acceptable $V_{WL}$-W region. This is illustrated in Fig. 3.4(b). In summary, the $V_{WL}$ and W values for a STTRAM cell need to be chosen to ensure that CTMR and switching current requirements are satisfied. A trivial solution is to increase both W and $V_{WL}$ until read margin requirements are violated. But that will significantly increase the energy-dissipation of the array. Hence, the cell design (i.e. $V_{WL}$ and W choice) need to consider their effect on the array energy. To analyze this effect in the next section we present the overall array energy model.

## 3.2 Energy Dissipation of STTRAM Array

The energy dissipation of the array is contributed by the total energy dissipation during read and write operations. Note, since STTRAM is a non-volatile memory all the voltage levels can be at 0 for a stand-by array. Hence, standby energy dissipation can be neglected. Let us now consider the energy dissipation during write and read operation for an array with $N_{row}$ number of rows and $N_{col}$ number of columns.

### 3.2.1 Dynamic Array Model for Energy Estimation

The dynamic energy of the array depends on the bit line and word line capacitances. To compute these capacitances, we need to construct the dynamic model of a column while switching. The dynamic model of a STTRAM cell consists of per cell bit line ($C_{BL}$), the word-line metal capacitance ($C_{WL}$) and parasitic capacitance at the intermediate node ($C_{node}$) as shown in Fig. 3.6. $C_{node}$ is contributed by the overlap and junction capacitance of the transistor as well as interconnect capacitance due to the connection of MTJ and transistor. The parasitic capacitance also exists at the node connected to the source-line and increases linearly with an increase in the width of the cell transistor. The MTJ structure also consists of two ferromagnetic layers separated by a thin dielectric layer and hence acts as a capacitance between bit line and the intermediate node. If we consider the bit line is switching from 0 to $V_{WL}$, intermediate nodes of all the cells (selected and unselected) will be charged to $V_{WL}$. Hence, the total energy dissipated during the bit line switching need to consider intermediate capacitance of all the nodes. Similarly for source-line, source parasitic capacitance ($C_{source}$) of the all the cells connected to the source-line need to be considered.

We evaluate the per cell bit line (i.e. $C_{BL}$) and word line (i.e. $C_{WL}$) capacitance from layout considerations (Fig. 3.7). We estimate the per cell bit line length $L_{BL}$ and per cell

word line length $L_{WL}$. Bit line capacitance is per cell is $C_{BL}=L_{BL}*C_0$ and word line capacitance per cell is $C_{WL}=L_{WL}*C_0$ where $L_{WL}=W+L$, $L_{BL}=7L$. Here W, L and $C_0$ are the width, length and capacitance of the metal line per unit length (taken to be 0.2 fF/um). Based on this model the total switching capacitance for bit-line ($C_{BL\text{-}switch}$), source-line ($C_{SL\text{-}switch}$) and word-line ($C_{WL\text{-}switch}$) are estimated as:

$$C_{BL-switch} = N_{row}\left(\underbrace{L_{BL}C_0}_{C_{BL}} + C_{node}\right) = N_{row}\left(7LC_0 + C_{node}\right)$$

$$C_{SL-switch} = N_{row}\left(\underbrace{L_{BL}C_0}_{C_{BL}} + C_{source}\right) = N_{row}\left(7LC_0 + C_{source}\right) \qquad (3.6)$$

$$C_{WL-switch} = N_{col}\left(\underbrace{L_{WL}C_0}_{C_{WL}} + C_{gate}\right) = N_{col}\left((W+L)C_0 + C_g\right)$$

where, $N_{row}$ is the number of rows per column; $N_{col}$ is the number of columns per row; Cg is the gate capacitance of the cell transistor which depends linearly on the cell width and W is the cell transistor width. The array consists of $N_{row}\times N_{col}$ number of individual cells.

### 3.2.2 Total Energy Estimation of STTRAM Array

The write energy of the array is contributed by three components, namely, (a) bit-line switching energy ($E_{BL}$), (b) word line switching energy ($E_{WL}$), and (c) energy dissipated in the cell due to the flow of write current ($E_{cell}$).

Bit line Switching Energy: We consider the worst case switching of the bit line and source lines. This occurs when a write “1” is followed by a write “0” or vice versa. Under the valid assumption that interconnect capacitance for source line and bit-line are similar, the total bit line switching energy is given by:

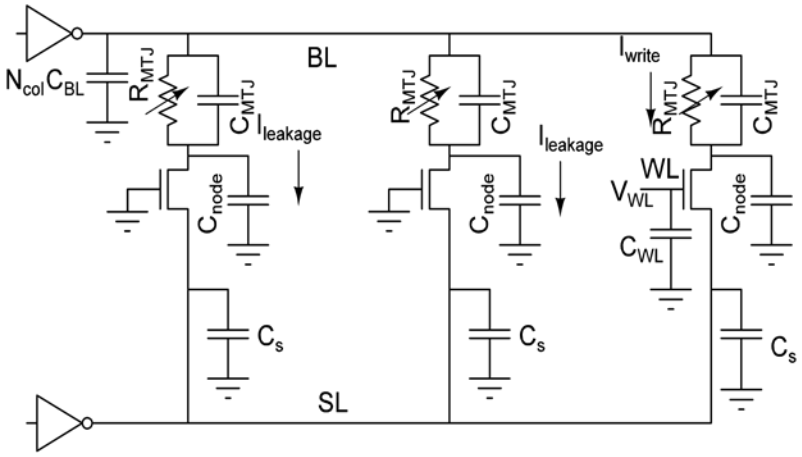



**Fig. 3.6: R-C models for array level energy computation**

$$E_{BL} = \left(C_{BL-switch} + C_{SL-switch}\right)V_{WL}^2 = N_{row}\left(2L_{BL}C_0 + C_{node} + C_{source}\right)V_{WL}^2 \quad (3.7)$$

Word line Switching Energy: Considering the array dynamic model, the word-line switching energy is given by:

$$E_{WL} = \left(C_{WL-switch}\right)V_{WL}^2 = N_{col}\left(L_{WL}C_0 + C_g\right)V_{WL}^2 \quad (3.8)$$

Energy due to Write Current: During writing we need to ensure the current flowing thorough the cell is equal to (or higher than) the required (worst-case) MTJ switching current for both write "0" and write "1" operation. As already mentioned we consider the higher of the MTJ switching current for write "0" and write "1" as our target MTJ switching current ($I_C$). The average cell switching current ($I_{sw}$) for write energy estimation (Fig. 3.8) is estimated as follows:

$$I_{sw} = 0.25\left[I_{sw1\to0} + I_{sw1\to1} + I_{sw0\to1} + I_{sw0\to0}\right]$$
$$E_{cell} = V_{WL}I_{sw}T_{write} \quad (3.9)$$

Array Write Energy: For a memory array with word-size of Nbit (i.e. Nbit number of columns are selected in each access) the total array energy is given by:

$$E_{write} = N_{bit}\left(E_{BL} + E_{cell}\right) + E_{WL} \qquad (3.10)$$

Read Energy of the Array: Following the previous discussion we can also obtain the read energy of the array which is given by:

$$\begin{aligned} E_{BL} &= N_{bit}\left(C_{BL-switch} + C_{SL-switch}\right)V_{read}^2 + \left(C_{WL-switch}\right)V_{WL}^2 \\ &+ N_{bit}V_{read}\,0.5\left(I_{read\,0} + I_{read\,1}\right)T_{read} \\ &= N_{bit}N_{row}\left(2L_{BL}C_0 + C_{node} + C_{source}\right)V_{read}^2 \\ &+ N_{col}\left(L_{WL}C_0 + C_g\right)V_{WL}^2 + N_{bit}V_{read}\,0.5\left(I_{read\,0} + I_{read\,1}\right)T_{read} \end{aligned} \qquad (3.11)$$

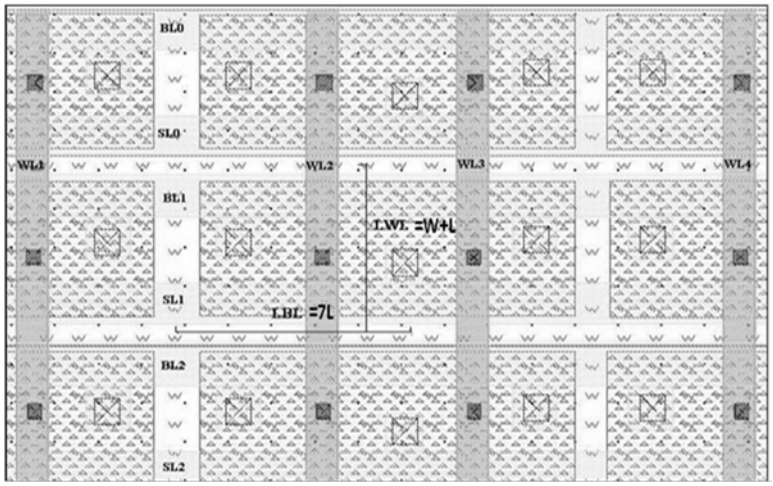



**Fig 3.7: Evaluation of word line and bit line capacitance from STTRAM cell layout**

In order to compute the energy model we consider equal word-line pulse width for both read and write operations. Note that the write pulse width is determined by the time required to write to a single cell i.e. the difference between the time instants when wordline is raised high and magnetization of the cell is switched. The read pulse width is determine by the difference of the time instants when the wordline is raised high and sufficient voltage difference is developed between the bitline and reference voltage of the sense amplifier for voltage sensing. For current sensing it will imply the time required to develop the sufficient difference between the bitline current and the reference current.

From [2] for a switching current of about 300uA we can expect a write pulse width of 10ns ( Twrite=10ns). Note the read access time of the cell can be smaller than the write access time. This method can be extended to capture this effect with proper read-write probability considerations and weights attached to Tread and Twrite can be used. In this work we assume equal cycles for read and writes. However it is possible that writes are accomplished in multiple cycles. To capture that effect an extension of the method with proper read-write probability considerations and weights attached to Tread and Twrite can be used.

### 3.3 Energy Aware Design Space Exploration

Let $I_{sw\_write0}$ is the switching current while writing "0" to a cell storing "1"; $I_{sw\_write1}$ is the switching current while writing "1" to a cell storing "0"; $I_C$ is the MTJ switching current; $\eta_{CTMR}$ is the minimum CTMR target; $TMR_0$ is the TMR of the MTJ device; $\eta_{RM}$ is the minimum read margin target (i.e. determines maximum read current value).

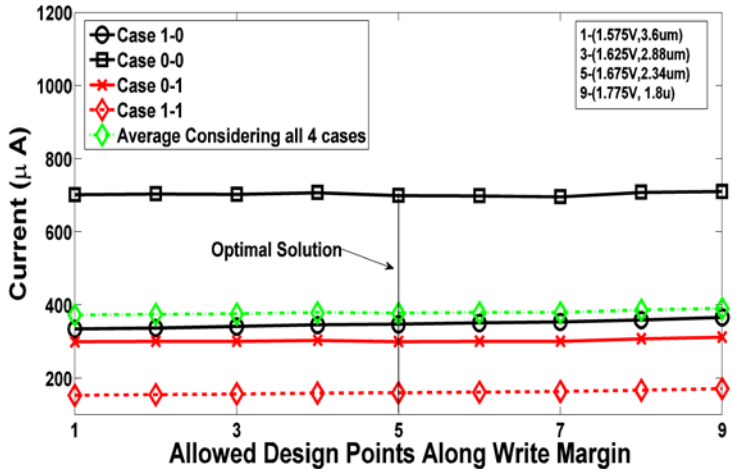



**Fig 3.8: Write current during 1→1, 1→0, 0→1, and 0→0 switching and their average value for all points in the $V_{WL}$-W plane that satisfies $I_C$=0.3mA**

First, we create a look-up-table (LUT) that provides the cell write current, read current and CTMR for $V_{WL}$ and W values satisfying the constraint. Next, given a $TMR_0$ and $\eta_{CTMR}$ we obtain the design space (given by $U_{TMR}$) for $V_{WL}$ and W to satisfy constraint (3.4) using the LUT as shown below:

$$U_{TMR} \equiv \{(V_{WL}, W): \ CTMR > \eta_{CTMR}\} \qquad (3.12)$$

Next, given $I_C$ we obtain the design space for $V_{WL}$ and W (given by $U_{IC}$) to satisfy writability constraint using the LUT as shown below:

$$U_{IC} \equiv \{(V_{WL}, W): \ I_{sw\_write0} \text{ and } I_{sw\_write1} > I_C\} \qquad (3.13)$$

Next, we obtain the design space ($U_{rd\_dstrb}$) that satisfies the read disturb condition using the LUT:

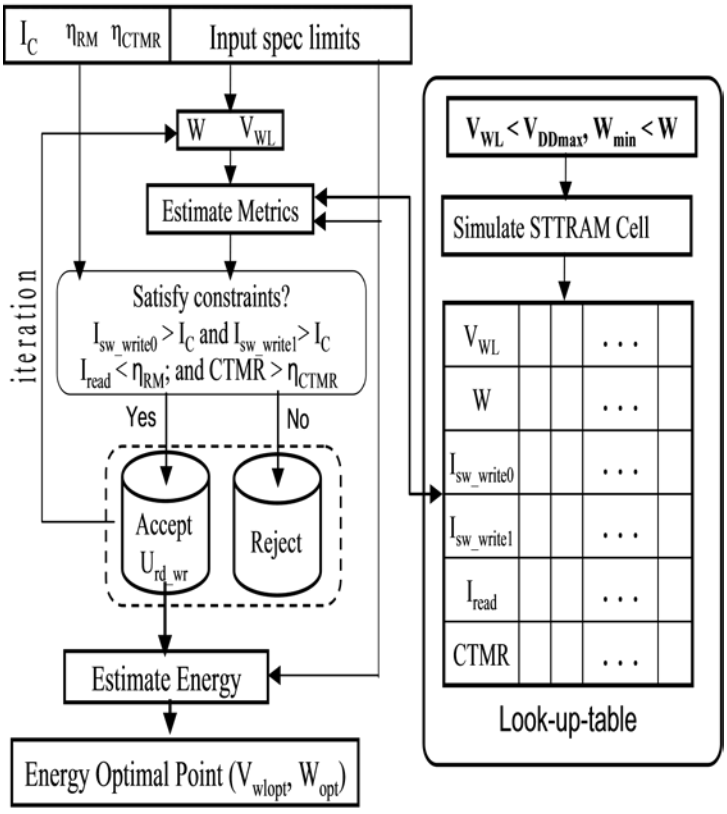



**Fig 3.9: Proposed methodology**

$$U_{rd\_dstrb} \equiv \{(V_{WL}, W):\ \mathrm{I}_{\min} < I_{read} < \eta_{RM}\ I_c\} \qquad (3.14)$$

Next, we obtain the feasible design space ($U_{rd\_wr}$) i.e. $V_{WL}$ and W that satisfy both constraints as shown below:

$$U_{rd\_wr} \equiv \left\{(V_{WL}, W):\ (V_{WL}, W) \in \left(U_{TMR} \cap U_{IC} \cap U_{rd\_dstrb}\right)\right\} \qquad (3.15)$$

Finally, we explore the $V_{WL}$-W space in $U_{rd\_wr}$ to obtain the minimum energy solution [using (3.12)-(3.15)] for $V_{WL}$ and W. The transistor width and the word line voltage are thus fixed giving us an energy efficient solution for the array meeting the criteria of TMR and write current. The methodology is shown in Fig. 3.9. From the energy surface for the allowed design space for particular set of constraints (the acceptance bin population in Fig. 3.9), the minimum energy solution is chosen. The factor '$\eta_{RM}$' is assumed to be 0.5. An exact estimation of '$\eta_{RM}$' for a reliability target can be performed using the statistical

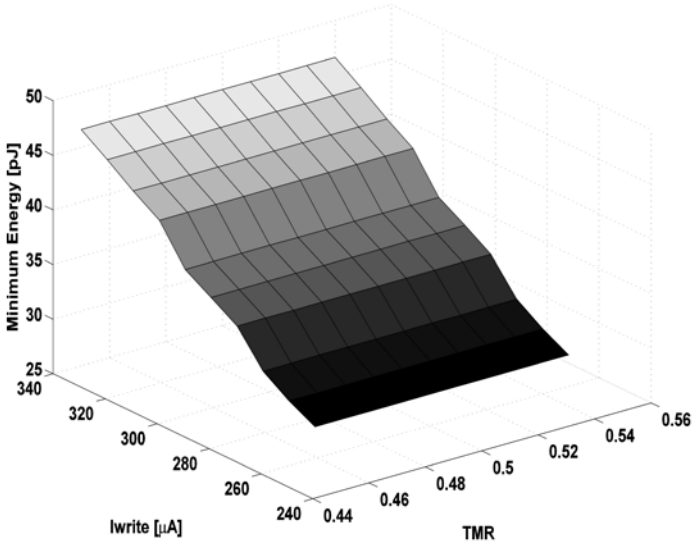



**Fig. 3.10: Increase in optimal energy for different switching current and CTMR target**

modeling technique proposed in [64]. Next, we consider different switching current and CTMR targets to obtain the minimum energy condition. Since a tighter constraints (i.e. higher CTMR and higher switching current) results in higher values for $V_{WL}$ and W, minimum achievable energy increases at tighter requirements (Fig. 3.10).

### 3.4 Performance Impact of Proposed Methodology

The performance of the STTRAM cell is primarily determined by the MTJ switching time which depends on the write current target. Since the proposed energy optimization method uses write current as a constraint it does not impact the MTJ switching time. However, the optimal solution calls for a new combination of operating voltages and transistor widths. Corresponding to each such combination the word line length changes and consequently there is a change in word line capacitance. The bit line length is not dependent on the transistor width and is not influenced by these changes. Hence, it is expected that the optimization of the energy will also impact the word-line delay and hence overall cell performance. For worst-case performance overhead estimation we consider the “1” to “0” (i.e. high resistance to low resistance transition) which requires a less write time for the same switching current as discussed in [2]. With increasing write

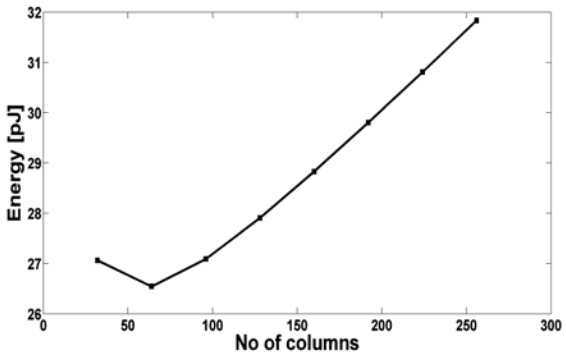



**Fig 3.11: Impact of array organization on energy**

current, the optimal transistor sizes increase, write pulse sizes decrease and the word line delay becomes a more significant part of the write time requirement. However even this is restricted to 6% which is not a significant contribution. For a read access requirement of 10ns, word line delay is also a constant 6% of the access time across switching currents.

### 3.5 Energy Variation with Array Organization

Increasing the column height increases the bit-line switching energy. Beyond a point, this switching energy dominates the total write energy. In other words, for long-bitline arrays the energy dissipated due to high switching current can be less than the energy dissipated in bit line switching. Consideration of different Nrow and Ncol values for a given array size show that bitline energy reduces with an reduction in the column height.Again, increasing Nrow increases bitline energy while reducing word line energy. The optimal energy point is also expected to vary with memory organization (Fig 3.11). However changing array configurations also have implications on the read constraint requirement which is beyond the scope of this work. Hence, in our work we restrict ourselves to a fixed array organization.

### 3.6 Energy Variation with Read-Write Ratio

In this section we consider the effect that different read and write ratios on the energy consumption of the STTRAM array. The total energy contribution of the array can be evaluated as:

$$E_{total} = \alpha E_{write} + (1-\alpha) E_{read} \qquad (3.16)$$

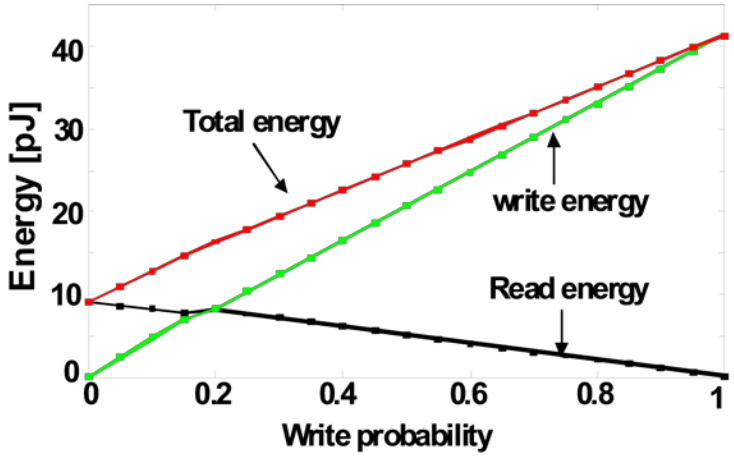



**Fig 3.12: Variation of energy with write probability**

where α is the probability of writing to the STTRAM cache. Fig. 3.12 shows that at very low write probabilities (<0.2) read energy is more than write energy. However, beyond write probability of 0.4 write energy starts dominating. This is due to the fact that the write current and the bit line voltages are larger than in the corresponding read case. Another interesting aspect is the total energy is 4X lower for read intensive operations compared to write intensive ones.

However for programs in the same benchmarks the read-write ratio varies appreciably in case of a cache application. Hence we do not consider optimization based on the read-write probability for cache. However, there is a scope for minimizing the total energy if we can modify the read-write ratio in L2 caches using STTRAM [65]. In that case, (3.16) can be treated as cost function to extend the proposed methodology to read/write biased applications.

Also this property could be useful in judging STTRAM applicability for different applications [66]. It however can be concluded that read dominant embedded applications could benefit with STTRAM use.

### 3.7 Exploring Opportunities for Skewed STTRAM Design in Cache

Read-write skewing in applications can be put to energy saving benefits by properly designing the STTRAM cell. The most obvious application in which the STTRAM can be thought of is at the lower levels of cache. So in this subsection we try to evaluate the opportunities of such fine tuning that exist there. The architectural simulations have been performed by Mitchelle Rasquinha of the CASL Lab.

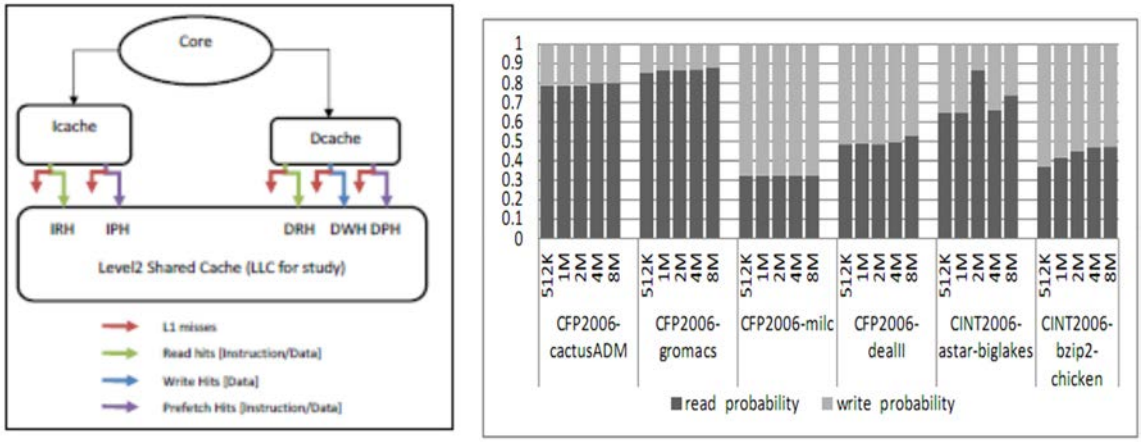
In this section, we present our experimental setup for capturing LLC access behavior and computing the read/write statistics. Fig. 3.13 illustrates a high level representation of a two level cache hierarchy with a single processor core and the accesses that correspond to read/write behaviors. All instruction read, data read, and data prefetch hits in the LLC correspond to cell read operations (3.17) while all instruction, prefetch and data misses as well as data write hits correspond to cell write operations (3.18). For our model we assume that the tag arrays of the caches do not use a STTRAM based cell design and hence the energy from accesses to the tag array is not accounted for.

$$\text{Total reads} = \text{IRH} + \text{DRH} + \text{IPH} + \text{DPH} \quad (3.17)$$

$$\text{Total writes} = \text{IRM} + \text{IPM} + \text{DWH} + \text{DRM} + \text{DWM} + \text{DPM} \quad (3.18)$$

Both expressions are multiplied by the corresponding array read and write energies Where E write is energy consumed during a write operation and E read is the energy consumed during a read operation. $E_{write}$ and $E_{read}$ are obtained by scaling (3.10) and (3.11) for a 32 bit word. A write occurs when there is a write from a program perspective or from a write due to a miss in the cache. Hence the raw miss rate amplifies the number of write operations. A second parameter that can adversely impact the number of write

operations is the cache pre-fetch policy. The result of a pre-fetch operation is a number of writes with the further impact that in some cases a pre-fetch may pollute the cache and be the cause for a cache miss, further increasing the number of writes. Traditionally hit rates are known to improve when the cache size is increased. With the use of STTRAM, which takes significantly less area than an SRAM equivalent, we have the option of realizing a much larger cache in the same area. Other design decisions or optimizations that reduce the number of write operations, even at the expense of increasing the raw number of read operations can prove to be beneficial when considering the energy consumption with an STTRAM based cache.




***IRH**: Instruction Read Hit; **IPH**: Instruction Prefetch Hit; **DRH:** Data Read Hit; **DWH**: Data Write Hit; **DPH**: Data Prefetch Hit;*

**Fig 3.13: Simulation setup and read-write ratio for different benchmarks**

We first analyze the read-write statistics of a set of benchmark applications. Fig. 3.13 shows the variation in read/write statistics for four SPEC CPU 2006 floating point applications and two SPEC CPU 2006 integer benchmark sets. All simulations were executed for 3 billion instructions (1 billion warm up period) using L2 cache sizes varying from 512KB to 8MB. The simulation environment for this section involved a

multicore version of zesto [67] compiled with gcc 4.1. The L2 is modeled as a 16 way set associative cache of varying sizes with an 8 cycle latency of access. We note that the relative probabilities of read vs. write operations are fairly stable for the MILC and the cactus ADM benchmarks and some variability can be seen for the bzip2 benchmarks reducing the write probability by approximately 0.1. A variation of 0.1 in the write probability is a reduction of approximately 10pj/access. As an alternative to changing the cache size, we may choose between a shared vs. private L2 organization. Shared caches are typically much larger in size and consume a significant percentage of the total energy of the core. The degree of sharing is the primary determinant in the choice of the private vs., shared caches. However, a high degree of sharing enables a single copy to be maintained in the LLC rather than being replicated in multiple private LLCs increasing the number of reads per cell with the consequent energy cost. In designs with private caches sharing however can cause increased coherence traffic and depending on the coherence protocol lead to invalidations thereby increasing the write traffic.

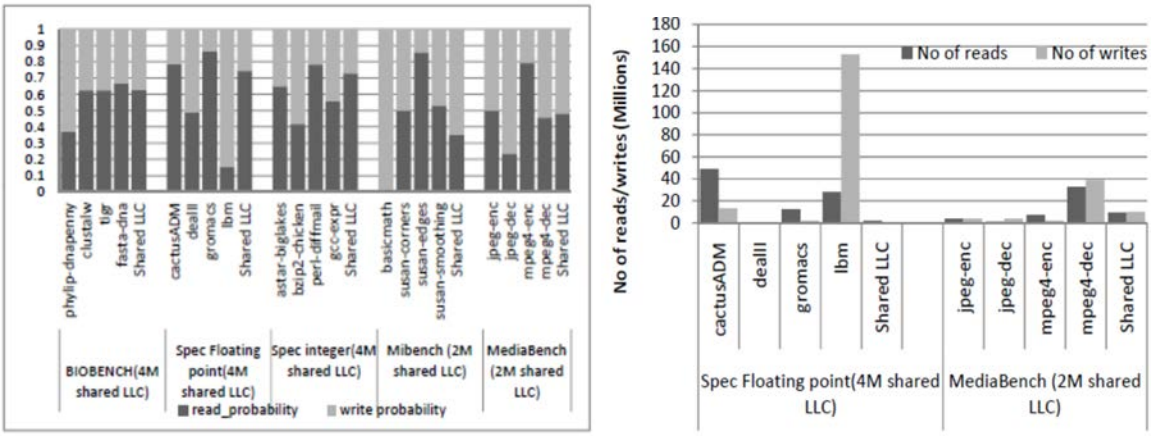



**Fig 3.14: Shared v/s Non-shared Cache Read-Write Ratio**

Hence design decisions such as these can play a significant role in the total energy consumed in the caches. Our analysis focuses on multi-programmed workloads where sharing is not within an application but rather through shared libraries. We evaluate four sets of benchmarks suites as shown in Fig. 3.14. The first 4 bars denote the write probability and energy/access from a single private cache of 1MB and the last bar indicates the energy/access from a single shared cache of 4MB (embedded applications are simulated on 2MB shared caches and 512 KB for the private caches). In simulating private caches, each benchmark group used a private L2 of one-fourth the size. A single cache line is 256 bits wide [8, 32bit words]. Thus the private cache model and the shared cache model use the same total size for LLC storage. The first set comprises of phylip, clustal, tigr and fasta from the biobench benchmark suite. The second two sets are from the SPEC CPU2006 benchmark suite. 436.CactusADM, 447.dealII, 437.gromacs and 470.lbm from the floating point applications and 473.astar 401.bzip2, 400.perlbench, 403.gcc from the integer applications. The last two sets are applications from MIBench [68] and MediaBench. In both the private and shared cases, each core is executing a single program from the set. Fig. 3.13 shows the read write probability and the the total number of cell reads and writes for a 3 billion instructions cycle simulation. The last bar in each set is for the shared cache case with the 4 multiprogrammed workloads. In all cases the LLCs are designed using the write statistics as detailed earlier. One write intensive benchmark can be seen from the plots in Fig. 3.14 is basicmath which is a benchmark from the automation and industrial control category of MIBench and performs simple mathematical calculations that often don’t have dedicated hardware support in embedded processors. Fig. 3.14(b) further shows how the absolute number of

writes is reduced with the use of a shared cache for the lbm benchmark. Reducing the number of writes will further reduce the energy per access. For write intensive workloads the saving in terms of energy may be significant as can be seen for the lbm workload in Fig 3.14(b).

While STTRAM presents opportunities for energy minimization it affects several traditional cache optimizations. For example, pre-fetching is a common optimization that has the negative consequence of amplifying the higher write energy of STTRAMs. Consequently, pre-fetch is no longer an obviously desirable feature. Turning off pre-fetch optimizations can increase the miss rate and therefore the execution time. However, this can be compensated for by increasing STTRAM cache size without a net increase in the miss rate and therefore without a net increase in the number of write operations. Finally the asymmetric read-write energy costs motivate compiler and architectural optimizations that will minimize write operations even at the expense of increased read operations. For example, replacement policies may bias the choice of line to evict based on the probability of this line be written rather than read.

## 3.8 Scalability of the Methodology

The study of technology scalability of the methodology is a must to demonstrate its effectiveness. Particularly interesting is to study the impact of the increased leakage energy on the optimal solution. We use PDK model based simulation at 65 nm to study the performance of the STTRAM cell. Works like [2] and projections from [1] suggest that STTRAM will scale with technology and the switching current will go down. Our works are based on reported values in [2]. We estimate the resistance and current at 65nm

assuming constant current density. Fig. 3.15 shows the optimal energy point at 65nm technology node. The critical current requirement is evaluated on the basis of scaling. The scalability study can be categorized under the following threads:

*Analysis of the trends followed by the optimal solution:* An interesting observation is that the w/l ratio at the optimal point at 65nm registers a decrease by 71.5% over the w/l ratio at 180nm. This shift in the optimal point over technology is significant. This is due to the fact that leakage becomes a more significant contributor from the array level sensing perspective. Hence the optimal point is being shifted more towards a smaller width solution to account for the increased leakage as technology scales. Thus it is expected that with scaling the optimal point will move towards a reduced width solution.

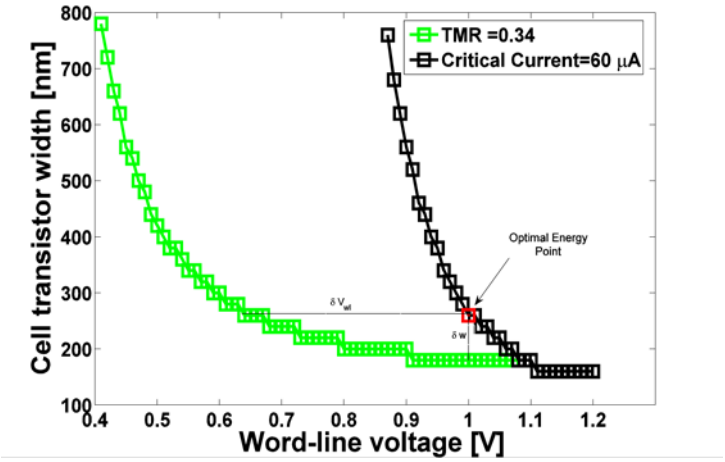



**Fig 3.15: 65 nm technology energy optimal solution**

*Impact of the scaled optimal solution on the read to write energy ratio:* Another interesting observation is shown in Fig. 3.16. The result illustrates the fact that the read–write energy ratio increases significantly at high read probabilities. The ratio is compatible at higher write ratios where write energy is the dominant component. In nanometer nodes, the critical current is expected to be less than 100 μA. However the

read voltage cannot scale proportionally with the word line voltage. Hence the read energy contribution towards the total energy increases.

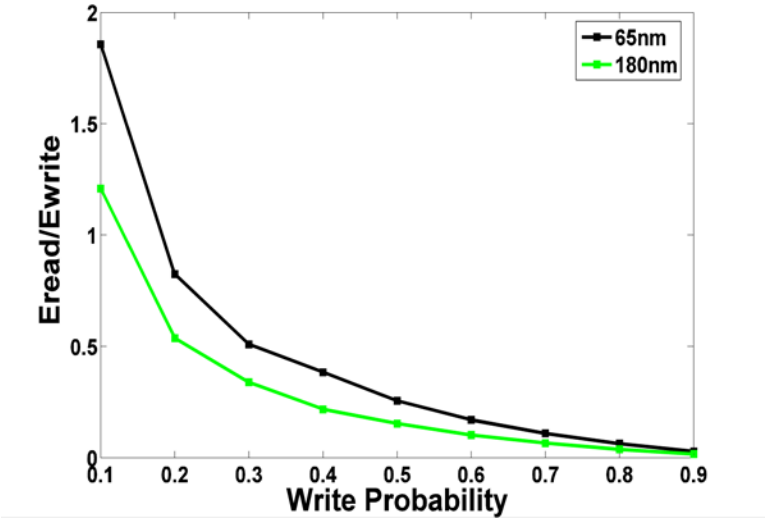



**Fig 3.16: Read-write energy ratios across technology**

## 3.9 Conclusion

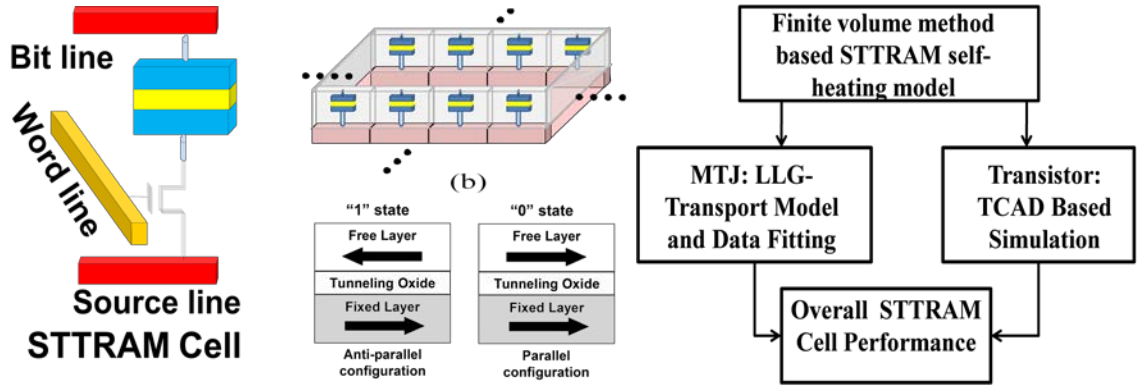
This chapter gives us a comprehensive understanding of the energy requirements of STTRAM. However, simply confirming the energy efficiency is not enough for deployment of STTRAM in embedded memory systems. For that, we require to also ensure that read and write reliability is sustained for such a memory technology. In the next chapter, we proceed to study read and write reliability in STTRAM from the perspective of thermal stability.

# CHAPTER 4

# IMPACT OF SELF-HEATING ON STTRAM: A STUDY ON THERMAL RELIABILITY

An STTRAM cell is written by applying a larger voltage difference between bit-line and source-line which cause a high *write current* to flow through the MTJ. When a write current greater than the critical switching current flows through the MTJ in the proper direction, the state of the MTJ switches from anti-parallel to parallel or vice-versa (~100μA for a write pulse width ~5ns-10ns). The proper direction of the write current is from bit-line to source-line for anti-parallel to parallel switching while source-line to bit-line for parallel to anti-parallel switching (Fig. 4.1).

The high write currents and small device volume can result in very high power density within the MTJ device and STTRAM cell. Consider an MTJ device of 100nm diameter and 10nm thickness (surface area $\propto$ 50nm×50nm, and volume $\propto$ 50nm×50nm×10nm) connected to bulk-silicon transistor through a metallic via. Considering write current of ~100μA, the power density within the MTJ volume can be easily estimated at ~$10^{12}$ W/cm$^3$ and at the surface is ~$10^6$ W/cm$^2$. This significantly high value of the power density can lead to localized temperature increase in MTJ. The temperature increase is further enhanced due to the fact that MTJs are embedded within ILDs which is poor conductor of heat. The metallic via conduct the heat from the MTJ to the transistor silicon resulting in increased temperature across the silicon substrate. Hence, high switching current in MTJ can seriously modulate the thermal profile STTRAM cell and increase in both MTJ and silicon temperature. We refer to this effect




**(a)** **(c)** **(d)**

**Fig 4.1: (a) STTRAM cell structure with MTJ, NMOS and controlling metal lines (Bit Line, Source Line and Word Line). (b)STTRAM array structure (c) MTJ in the "1" and "0" configurations. (Relative orientation of free and fixed layer determine resistance state) (d) Electro-thermal co-simulation framework for STTRAM. It uses FVM based self-heating solutions coupled with LLG-transport model for MTJ and TCAD simulations for silicon**

as self-heating in STTRAM. The increased temperature can significantly degrade operational reliability of an STTRAM cell. Decreased operational reliability induces read disturb (cell flipping during read), write (incorrect write operation) and detection (incorrect sensing of cell value) failures. Hence, simulation of the self-heating effect in STTRAM and analysis of its effect in cell reliability is important for the development of STTRAM technology. While the effects of temperature on MTJ device has been studied [30], the modeling and analysis of the self-heating effect has received limited attention in research.

This chapter models self-heating in STTRAM cell and analyzes its effect on operational reliability of the cell (Fig. 4.1(d)). It presents a detailed finite volume method (FVM) based model to estimate self-heating effect in STTRAM cell. We analyze the steady state and transient thermal behaviors of STTRAM cell using the FVM model considering the impact of critical magnetic-tunnel-junction (MTJ) parameters such as

switching current, resistance-area product, and the switching current and write pulse width interaction. We further study the temperature dependence of switching current and resistance of MTJ as well as properties of the NMOS transistor (considering 65nm high K metal gate device using drift-diffusion device simulation). The mixed-mode device-circuit simulation is used to analyze the effect of temperature rise on cell reliability. The interaction of temperature rise within the MTJ device and transistor and circuit behaviors of STTRAM cell such as read/write current under different data conditions are studied. Finally, the observations from above analysis were coupled to study the interaction of self-heating effect and cell reliability to estimate the effect of read/write data patterns on cell reliability.

### 4.1 STTRAM Cell: Impact of Temperature

The functional reliability of an STTRAM cell is defined by following metrics:

**Write margin:** Write margin is defined by the difference between write current and MTJ switching current. The switching current is a property of the MTJ. The write current depends on MTJ resistance and the current drive of the transistors. An increase in the switching current and/or reduction in the write current (due to increase in MTJ resistance and/or reduction in the transistor strength) degrades write margin.

**Read margin:** Read margin is defined as the difference between switching current and read current (current flowing through MTJ during read operation). The read current depends on MTJ resistance and transistor resistance. A lower switching current and/or higher read current degrades read margin.

**Detection accuracy:** Incorrect or false detection refers to the detection of a bit as “1” when stored bit is “0” and vice versa. In STTRAM the bit values are detected depending

on the difference in the read current for cell storing “0” and “1” (i.e. MTJ in anti-parallel or parallel states). Further, during sensing an STTRAM cell, the current flowing through the bit line is sensed. Chapter 5 points out that during reading a cell in a selected column, the unselected cells in that selected column contributes leakage current. This leakage acts as a circuit induced noise to the sensed current. Hence, variations in the MTJ resistances, transistor strength, and transistor leakage modulate the probability of false detection.

In this section, we discuss the effect of temperature on the reliability metrics of STTRAM cell. We study effects of temperature first on MTJ properties, next on transistor properties; and finally on cell properties.

### 4.1.1 Impact of Temperature on MTJ Properties

#### 4.1.1.1 Impact on MTJ Switching Current

The coupled quantum transport-magnetization STT simulator used for the study was developed by Dr. Sayeef Salahuddin. The results from the simulator have been used towards the study. The effect of temperature was included through a stochastic integration of the magnetization dynamics. Fig. 4.2 shows the schematic of the simulation methodology. The quantum transport is modeled using open boundary Schrodinger’s equation within the Non-Equilibrium Green’s Function (NEGF) formalism (Eqn. (4.1)). On the other hand, the magnetization dynamics is modeled using the Landau-Lifshitz-Gilbert (LLG) equation (Eqn. (4.2)) [69].

$$G_d = (E - H_d - \Sigma_1 - \Sigma_2)^{-1} \tag{4.1}$$

$$(1+\alpha^2)\frac{\partial m}{\partial t} = \gamma\left(mXH_{eff}\right) - \frac{\gamma\alpha}{m}\left(mXmXH_{eff}\right) + current\ torque \tag{4.2}$$

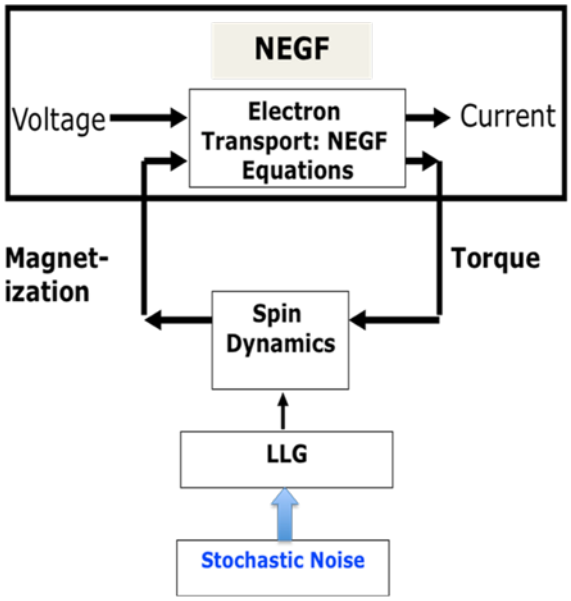



**Fig 4.2: MTJ device simulation framework: self-consistent solution of Landau-Lifshitz-Gilbert (LLG) and NEGF transport equation**

In Eqn. (4.2), $G_d$ is the device Green function, $H_d$ is the device Hamiltonian, E is the energy term in Schrodinger's equation and Σ's are the self-energy due to the contact. In Eqn. (4.2), m is the magnetization of the magnet, γ is called the gyromagnetic ratio, $H_{eff}$ is the effective magnetic field and α is the Gilbert damping parameter. Essentially the spin polarized current flowing in the device exerts a torque in the magnet that is calculated from NEGF and then used as a source term in the LLG equation. This changes the relative orientation of the magnetization. In turn, the specific orientation of the magnetization changes the way the spin polarized current flows through the device. Thus NEGF-LLG needs to be solved self-consistently. Detailed discussions on this self-consistent framework may be found in [31, 32, 70]. For this work, switching dynamics was simulated by self-consistent NEGF-stochastic LLG simulations and assuming 80% flipping as the switching threshold. We observe that an elevated temperature results in a faster but chaotic switching (Fig. 4.3(a) and 4.3(b)) [31, 32, 71]. This is because the spins

have higher energy to cross over the barrier. Consequently, the switching current requirement reduces at higher temperature (Fig. 4.3(c)). Thus at high temperatures, STTRAM can be susceptible to read disturbs whereas write failures at high temperatures can be suppressed.

4.1.1.2 Impact on MTJ Resistance

Majumder et. al. have experimentally demonstrated the effect of temperature on MTJ resistance [72]. It was observed that resistances in the parallel and anti-parallel modes for

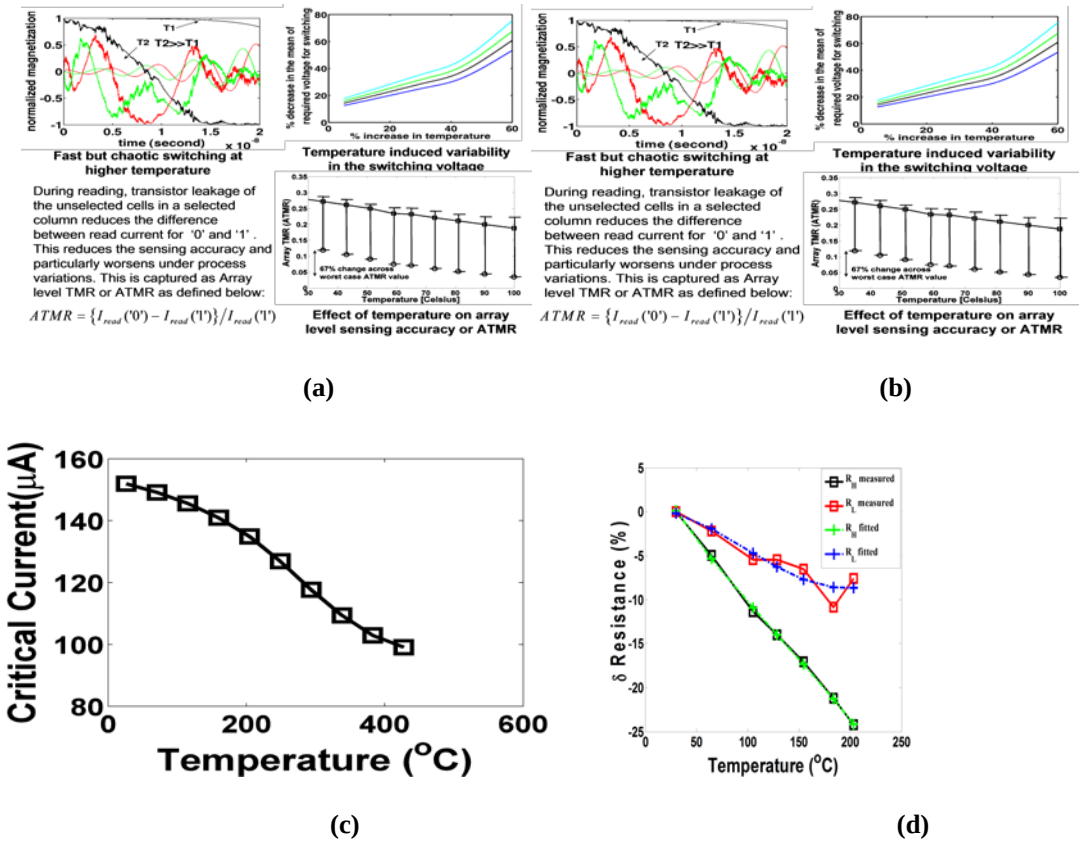



**Fig 4.3: (a) Faster but chaotic switching of the normalized magnetization at higher temperature (b) Temperature induced variability in switching voltage. Both (a) and (b) point to increased read disturb probability (c) Reduction in switching current requirement with temperature (d) Temperature drift characteristics of MTJ resistance. Drift characteristic curve fitted for the data reported in [10]. Both (c) and (d) point to reduced write failures with temperature**

MgO based MTJs reduce at higher temperature (Fig. 4.3(d)). This is in conjunction with the model assuming conductance having two components- spin-dependent and spin-independent. The spin dependent component is expected to follow the empirical dependence of $T^{-\alpha}$ [73].This reduction in resistance would increase write and read current through the cell. Hence read disturb probability would increase but that of write failure reduces. As the MTJ resistance reduces false detection probability will increase in presence of process variation.

We fit the observed data to a polynomial curve and use that to analyze the effect of self-heating first, on MTJ resistance and next on cell reliability.

#### 4.1.1.3 Impact of Temperature on NMOS Access Device

In this section we study the effect of temperature on the transistor properties. We perform this study considering a 65nm high K metal gate transistor. The drift-diffusion based mixed-mode device simulator (Medici [74]) is used to perform this study. In the following subsections, the transistor characteristics and their temperature dependence are characterized to evaluate the STTRAM properties. Table 4.1 shows the simulated transistor dimensions and properties. The Id-Vg characteristics for the transistor are shown in Fig. 4.4(a).

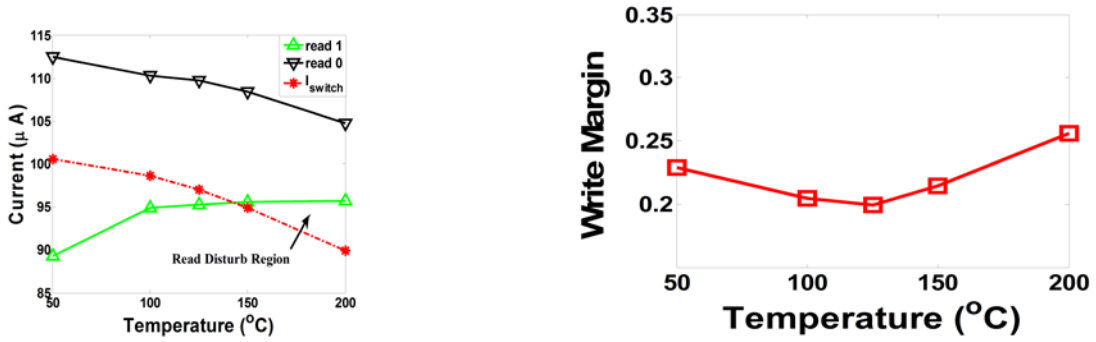
**Table 4.1: NMOS properties in STTRAM**

| | |
|---|---|
| DIBL | 44.44 mV/V |
| $I_{ON}$ | 532µA |
| $I_{OFF}$ | 4.36nA |
| $I_{ON}/I_{OFF}$ | 1.22e5 |
| $V_{THRESHOLD}$ | 0.26 V |
| Gate Length | 65nm |
| Work Function | 4.2eV |
| Oxide + High K Thickness | 3nm |

## 4.2. STTRAM Cell Level Impact

To understand the variation of the STTRAM performance metrics with temperature, we have to simulate the MTJ and NMOS temperature dependences in conjunction. We study the combined effect using mixed-mode device simulation [74]. We model the STTRAM cell using the NMOS device discussed in previous section and resistances to represent the MTJ. First, we consider the values of MTJ resistances ($R_H$ and $R_L$) and

critical switching current at room temperature. The width of the NMOS was chosen to ensure correct write operation at room temperature. This is achieved when the current flowing through the cell in both directions (i.e. bit-line to source-line and source-line to bit-line) is higher than the MTJ switching current at room temperature. To study the effect of temperature variation on the cell parameters, we vary the "simulation temperature" which modifies the NMOS properties. On the other hand, the effect of temperature on the MTJ is captured by modifying the value of the MTJ resistances ($R_H$ and $R_L$ as appropriate and following Fig. 4.3(d).

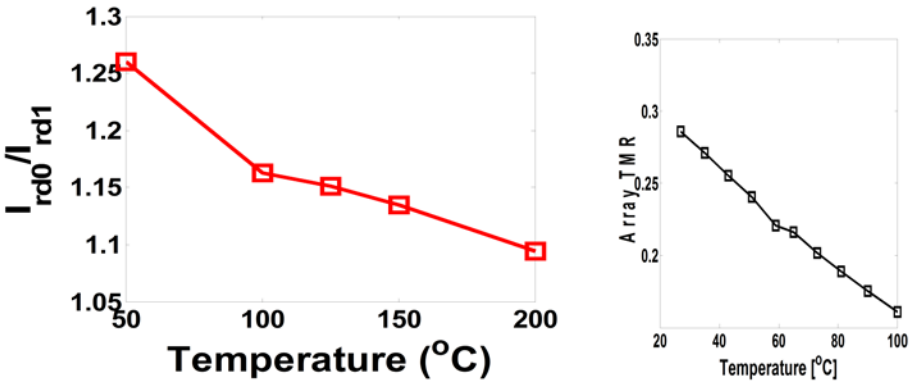



**(a)** **(b)**

**Fig 4.5: Evaluation of the impact of temperature on (a) Read disturb (b) Write margin**

### 4.2.1 Read Failure

Read disturb occurs when read current is larger than the switching current and hence, flips the bit content. Given the direction of read, only bit flip from "1" to "0" is likely. Whenever the read "1" current crosses the switching current, flipping occurs. Therefore, read margin can be defined as the difference between MTJ switching current and read "1" current. We plot the behavior of read "0" and "1" currents with temperature and compare it with the switching current at different temperature (Fig. 4.5(a)). Note that at higher temperature MTJ resistances in both parallel (read "0", $R_L$) and anti-parallel mode (read




**(a)** **(b)**

**Fig 4.6: Impact of temperature on (a) Read "0" current / Read "1" current (b) Array Level Distinguish-ability metric [4]**

"1", $R_H$) reduces, but $R_H$ reduces at much higher rate [Fig. 4.3(d)]. The transistor resistance increases with temperature [Fig. 4.4(b)]. Results indicate that the read "0" current reduces with temperature as increase in transistor resistance overshadows reduction in $R_L$. However, read "1" current can even increase at higher temperature as large reduction in $R_H$ can mask the increase in transistor resistance. The increase in read "1" current and decrease in the switching current of MTJ [Fig. 4.3(c)] results in a reduction in read margin.

### 4.2.2 Write Failure

Write failure occurs when the write current falls below the switching current. For a "0" to "1" flip, the circuit is in source degenerate mode (MTJ present at the source). For a "1" to "0" flip, MTJ acts as a resistive load at the drain. Thus for same $V_{dd}$ applied to both cases, write current is lesser amongst two flip conditions for "0" to "1" [4, 64]. Hence, we consider the write "0" to "1" as the more probable switching condition for write failure. We evaluate write margin for different temperature under this condition. Note that at higher temperature switching current reduces which tends to increase the write margin. The effect of increased temperature on write current is determined by two factors.

A reduced switching current of transistors at higher temperature tends to reduce write current while a lower MTJ resistance (~$R_L$ before MTJ switching) helps increase write current. The net effect of the above three factors determine the sensitivity of write margin with temperature. We observe that, the combined effect makes write margin less sensitive to temperature (Fig. 4.5(b)).

### 4.2.3 False Read

A higher NMOS resistance reduces the ratio of the difference between cell resistance (MTJ resistance + NMOS resistance) for bit "0" and "1" with respect to the average cell resistance. This can increase the probability of false detection. This is shown in Fig. 4.6(a) which plots the ratio of read current during reading "1" and reading "0" at higher temperature. Moreover, as explained earlier leakage from the unselected cells of the selected column reduces the detection accuracy further. As the leakage increases with temperature, the array level detection metric, array TMR or ATMR, degrades with temperature (Fig. 4.6(b)). The metric ATMR is given by: $ATMR=(I_{cell0}-I_{cell1})/(I_{cell0}+I_{leakage})$ [4].

## 4.3 Simulating the Self-Heating in STTRAM

In this section we characterize self-heating effect in STTRAM. We perform a finite volume method (FVM) based analysis of the STTRAM to characterize the cell thermal distribution.

### 4.3.1 Finite Volume Method Based Model

Finite volume methods have been widely used to simulate thermal systems. In this method, Fourier Conduction Equations are integrated over each control volume (grid cell)

to get algebraic equations for each cell. There is a tradeoff between the meshing resolution and solution time. For this work we use Gambit® for generation of the STTRAM cell mesh (Fig 4.7(a)). Figure 4.7(b) shows a cross section of the implemented cell structure while 4.7(c) shows the top view of the cell layout. We used non-uniform meshing across the STTRAM structure. This is to maintain a balance between the small mesh resolutions required for representing certain portions (eg. MTJ) while a coarse resolution is maintained for other parts to restrict the memory requirement and computation time. We use Fluent® finite volume solver for our purpose. A convection boundary condition is applied to the Si surface such that a current of 10uA flowing through an equivalent resistance of 1kΩ across a transistor of length=200nm, width=700nm, junction depth=10nm gives temperature rise of 52$^{o}$C across the bulk silicon [75, 76]. Next we consider a MTJ with an area of 100x100 nm$^2$ R-A product of such an MTJ is 30Ω-um$^2$. For a TMR of 42.8, it has high and low resistance values of 4.7kΩ and 3.29kΩ respectively. We note that the MTJ is compatible with 65nm CMOS technology.

### 4.3.2 Results of FVM Analysis: Steady State

Considering a critical current density of $10^6$A/cm$^2$, the critical current requirement for the MTJ is evaluated at 100uA. For a MTJ height of 10nm, power density of $2.5x10^{17}$W/m$^3$ results across the MTJ and power density of $5x10^{15}$W/m$^3$ across silicon. If current flows continuously through the structure the final temperature distribution reaches the steady state. Fig. 4.7(d) shows the thermal distribution across an isolated cell at steady state condition. The system is surrounded by interlayer dielectric ($SiO_2$) for which we assumed no heat inflow/outflow from the boundary. For the heat sink, the heat flux

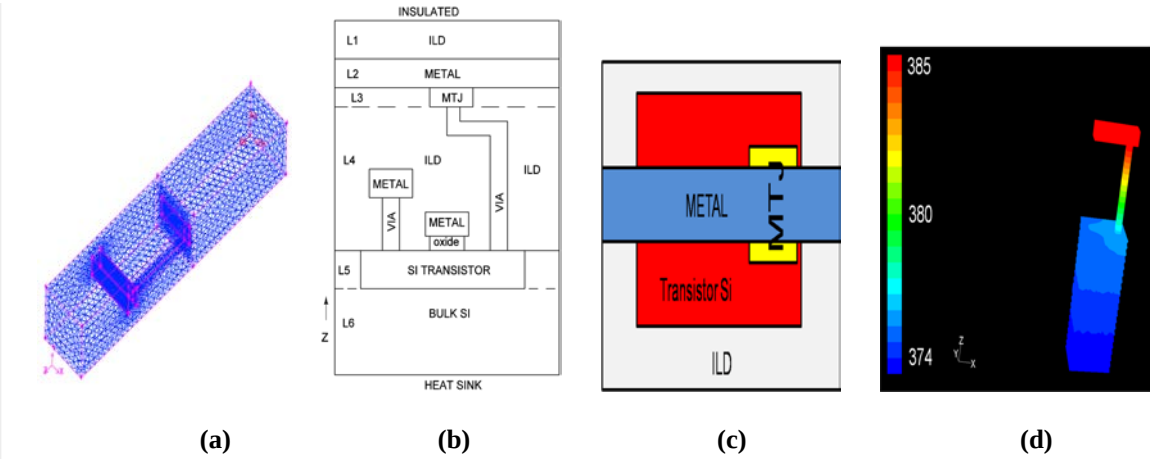



**Fig. 4.7: (a) Constructed mesh for the cell in Gambit® (b) Cross sectional view of a STTRAM cell (c)Top-view of the cell (d) Temperature distribution across simulated FVM model**

was assumed to be 100 W/m$^2$. The metal is assumed to conduct heat away using conductivity of A steady state solution estimates a final temperature of 112$^o$C inside MTJ [76, 77].

### 4.3.3 Results of FVM Analysis: Transient

We are interested in the transient thermal response of the MTJ and Si. To do this, the FVM model is subjected to current pulses of 100uA with 200ns time period and 50% duty cycle and temperature is observed. The result is shown in Fig. 4.8(a). If left for large

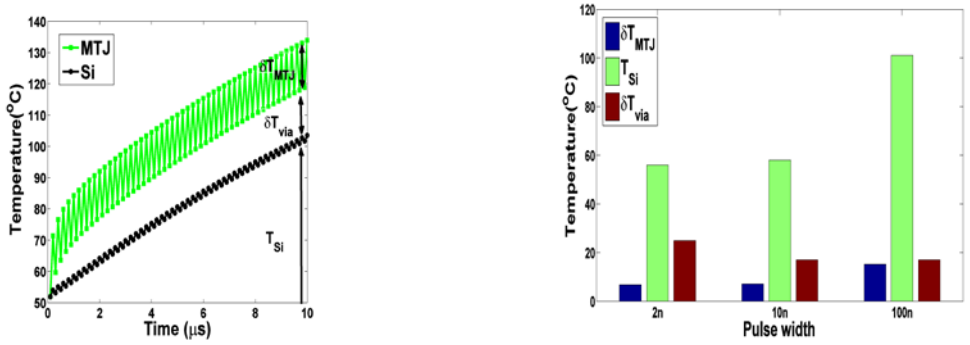


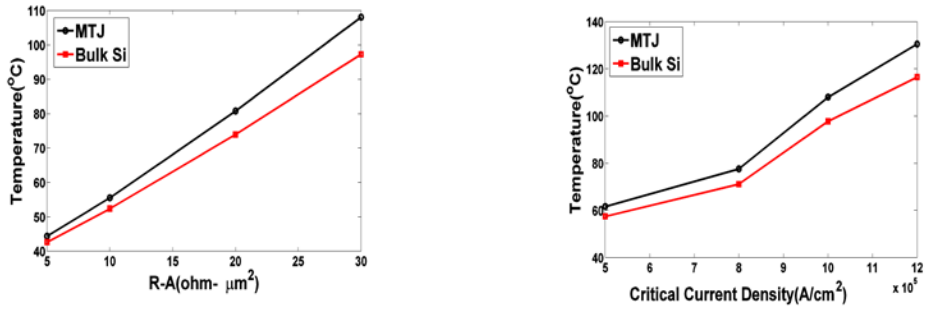

**Fig 4.8: (a) Temperature rise for applied write pulse across MTJ and bulk (b) Effect of different pulse width on temperature**

time it converges to the case in Fig 4.7(d). The notable feature is that the MTJ rises much faster than the Si. This is because of the much higher thermal capacitance that bulk Si offers in comparison to the MTJ.

Also the MTJ shows a fluctuation in temperature by $7^{o}$C within a time period between the on and off cycles. We take this along with the temperature drop across the metal via and silicon temperature to characterize the thermal transient. We consider the thermal distribution across the STTRAM for cases of write pulse widths of 2ns, 10ns and 100ns considering switching currents of 480uA, 300uA and 160uA (obtained by scaling [2]) respectively. The results for performing 100 consecutive write cycles are shown in Fig 4.8(b). The notable observations are that the difference across the metal via i.e. the MTJ and the silicon is approximately 9-$10^{o}$C and is the highest for 2ns. The overall temperature rise is greater for 100ns case.

### 4.3.4 Results of FVM Analysis: Effects of Material Properties

The temperature distribution across the STTRAM cell depends on the critical current density and material. The material decides the R-A product. To begin with we consider a material with a variety of R-A product. For the same area and switching current, increasing resistance twofold will result in two times the power dissipation. This translates to a proportional temperature increase. The result is shown in Fig. 4.9(a). Similar deductions can be made regarding the critical current density. A reduction in critical density means reduced current for same area and R-A product. This is supported by the observations made in Fig 4.9(b). A common inference which can be deduced from the above observations is that reducing critical current density and R-A product are desirable to reduce the thermal dissipation across the cell. A comparative study of the




**(a)** **(b)**

**Fig 4.9: Temperature variation with (a) R-A product (b) Critical current density**

plots reveals that the temperature reduces at a faster rate with the critical current density. This is understandable as the temperature is expected to vary in a quadratic manner with current density.

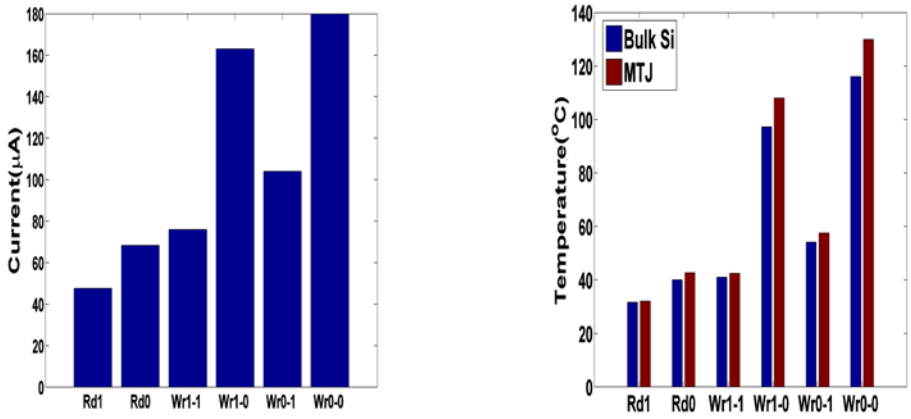
## 4.4 Other Impacts of Self-Heating

In this section we apply the self-heating results obtained in the previous section to the device thermal sensitivity results obtained earlier. We begin with the several read and write currents estimated from circuit simulations (i.e. applying a constant bit-line, source-line, and word-line voltage depending on the operating condition). The temperature corresponding to the initial set of currents is evaluated from the FVM model. Note MTJ and silicon temperature will be different for a given read or write condition. The estimated set of temperatures is used in mixed-mode circuit simulations with the NMOS device and MTJ resistance to re-estimates the read and write currents. The silicon temperature is used as "simulation temperature" to capture the effect of self-heating in NMOS properties. The effect of self-heating on the MTJ is captured by modifying the value of the MTJ resistances. We perform the above analysis considering self-heating for different read or write condition (i.e. read and write currents depending on the bit values and patterns). For a given read/write condition, we consider the steady-state MTJ and

NMOS temperature cell reliability analysis (i.e. the pulse width is large or the same operation is repeated continuously for a large number of cycles). The analysis considering the steady-state condition helps evaluate the worst-case cell reliability. In the following subsections we summarize our observations.

### 4.4.1 Different Write and Read Patterns

Though read current in STTRAM is sufficiently smaller than write current, it will dissipate considerable amount of power across the STTRAM. Consider the average read current with 100ns pulse width is 50uA. Further, consider write pulse width of 100ns as well. The thermal analysis for different read and write condition are performed under this assumption. Fig 4.10(a) shows that for same pulse width condition, the mean value of the temperature is less by 25-30$^{o}$C for reads. This difference however strictly depends on the read margin. Though we have studied their effect separately, read and writes occur to the same cell and the access pattern is governed by the memory application. Hence the thermal distribution of a cell in reality will be a function of the access pattern. The worst case corner arises for consecutive writes to the cell. Another interesting aspect is that




**(a)** **(b)**

**Fig 4.10: (a) Read-write currents (b) Temperatures for the cases**

even among the writes, there can be writes intended for flipping (0-1 or 1-0) and redundant writes (0-0 or 1-1). Each write condition offers a certain initial MTJ resistance

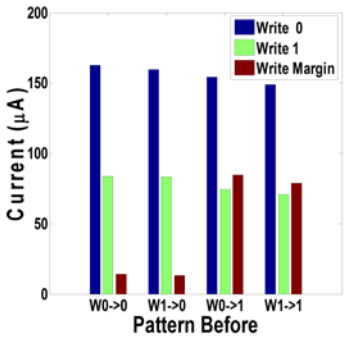



**Fig 4.11: Access pattern history dependence of write current**

($R_H$["1"] or $R_L$["0"]) and circuit configuration (write 0: BL->high SL->low and write 1: BL->low and SL->high). Simulation results show that the write 0-0 and write 1-0 give rise to the maximum temperature (Fig. 4.10(b)).

### 4.4.2 Effect of Past Access History

We have observed that the past set of operations determines the MTJ and silicon temperature. Therefore, the operational reliability of STTRAM cell during read or write mode will depend on read-write history of the cell [77]. In this section we study the effect of history of past operations on the reliability of a current read or write operation.

#### 4.4.2.1 Effect on Write

In this section we address the issue as to whether previous access patterns influence write disturb. We first evaluate the MTJ and NMOS temperature at the beginning of a current write operation considering self-heating due to different past read/write operations. The estimated temperature is used to estimate write current (using mixed-mode simulations as discussed) and MTJ switching current (using Fig. 4.3(c)) at the

beginning of current operation. The estimated write current is compared with the switching current to estimate write margin. In Fig. 4.11 the x-axis indicates the previous set of operations executed. The write current for writing “0”, the write current for writing “1”, and the write margin are shown in blue, green and red respectively. Write failure can occur when there is a write with flip requirement (“0” to “1” or “1” to “0”). Therefore, the write margin is measured as the difference between write current corresponding to flipping operation and the switching current. For example, the first set of bars with x-axis label w0→0 represents the write currents for writing “0”, writing “1”, and write margin. The label w0→0 represents previous operation was writing “0” to “0”. When the previous history is w0→0, the next write “0” represents a redundant write. The write “0” in this case, is therefore important for only energy analysis. The write margin (red bar) in this case is computed considering write “1” current (for ones with beginning state “0”- w0→0, w1→0) and write “0” current (for ones with beginning state “1”- w0→1, w1→1) and MTJ switching current. The sets of bars for other operations can be interpreted similarly.

We observe that for all cases of past operation, the write “0” current is always larger than write “1” current. This is attributed to the bi-directional switching condition in STTRAM. The NMOS size is determined considering write “1” conditions (MTJ appears in source of NMOS) which leads to larger than require size for write “0” condition (MTJ acts as a load at the drain of NMOS). A write failure while writing “0” can only occur when past conditions are w0→1 or w1→1. We observe that write “0” current is minimum when previous pattern is w1→1 (Fig. 4.11). This is because w1→1 leads to minimum initial temperature and hence, maximum value of MTJ resistance in anti-parallel ($R_H$)

condition and higher MTJ switching current (Fig. 4.2). Hence, Fig. 4.11 shows that self-heating and past history result in a 4% reduction in write margin for write “0”. Likewise we observe that write margin for write “1” is 3% smaller when past operation is w1→0 compared to w0→0. This is because w0→0 results in higher initial temperature due to self-heating.

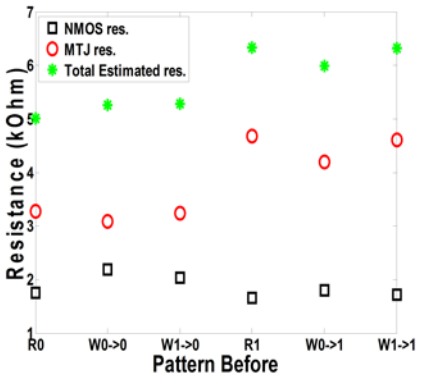



**Fig. 4.12: Access history dependence of read**

4.4.2.2 Effect on Read

We next study the effect of previous pattern on the detection reliability (Fig 4.12). Fig. 4.12 has a similar x-axis as Fig 4.11. In the y direction it plots the MTJ, NMOS and combined resistances. The read “0” cases are the ones with prior history of R0 (read “0”) and w0→0 and w1→0. The two resistances (NMOS and MTJ) are closest to each other following a redundant write “0” (w0→0). This is due to the maximum initial temperature for both MTJ and NMOS (Fig. 4.10) which results in lower MTJ resistance $R_L$ (Fig. 4.3) and higher device resistance (Fig. 4.3). The maximum difference between MTJ and NMOS resistance during read “0” is observed for previous pattern of R0 (smaller MTJ and NMOS temperature Fig. 4.10). We further observe that cell resistance for read “0”

can vary by 5.4% due to prior history and self-heating. For read “1”, previous history of w0→1 results in least cell resistance and minimum difference between NMOS and MTJ resistance ($R_H$). This is because prior history of w0→1 results in maximum initial temperature for read “1” condition (Fig. 4.10) and hence, smaller MTJ resistance and higher NMOS resistance. The prior history of read “1” (R1) leads to maximum cell resistance and maximum difference between MTJ and NMOS resistance (due to smaller temperature considering self-heating). Therefore, we observe that cell resistance for read “1” can vary by 6% due to self-heating. It is imperative that detection accuracy degrades when the difference between MTJ resistance and NMOS resistance reduces. Hence, false

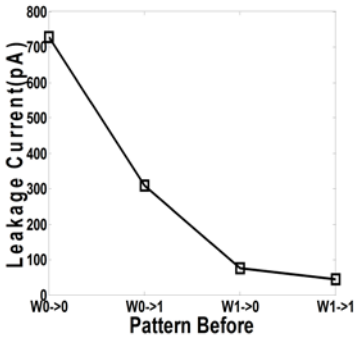



**Fig. 4.13: Access history dependence of leakage**

detection for read “0” and read “1” are most probable with past history of w0→0 and w0→1, respectively.

The cell level distinguish-ability depends on the ratio of cell current while reading “0” and “1” (i.e. ratio of cell resistances). From Fig. 4.12, we conclude that self-heating can results in appreciable variation in cell level distinguish-ability. The ratio of cell resistance (MTJ + NMOS) while reading “1” and reading “0” can vary from 1.26 to 1.13. The distinguish-ability can be further varied due to the variation in the leakage currents of the

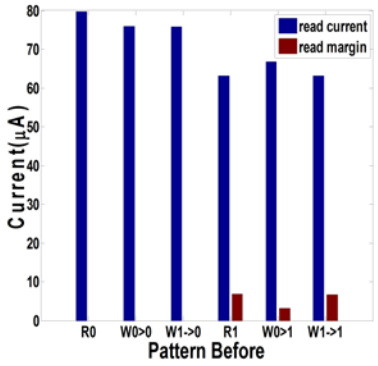



**Fig. 4.14: Access history dependence of read disturb**

unselected cells. Depending on the prior read/write history of the unselected cells, we may have different leakage currents as shown in Figure 4.13. The maximum leakage is expected for redundant "0" write as the temperature is found to be highest for this case.

The self-heating and prior read/write history modulate both cell resistance (read current for a given bit-line to source-line voltage) and switching current, and hence, read margin. To find the read margin we use the computed read currents and switching currents. We note that read disturb is only possible when reading "1" as flipping "0" is not possible by current flowing from BL to SL. Fig. 4.14 shows the read currents and read margins with the x-axis indicating the previous set of operations. We observe that the read margin is minimum when read "1" follows w0→1. This is attributed to the fact that w0→1 results in higher temperature compared to other two cases of prior operations (i.e. R1 and w1→1). The higher temperature results in lower cell resistance and hence, higher read current (Fig. 4.12, 4.14). This is coupled with lower switching current at higher temperature (Fig. 4.3). Therefore, the read margin i.e. difference between switching current and read current is smallest among different cases.

4.5.2.3 Impact on Bit Stability

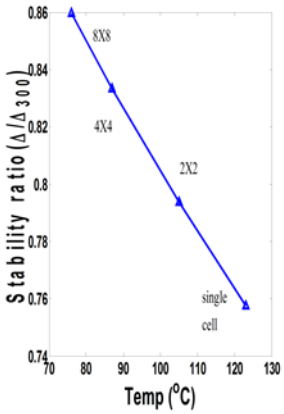



**Fig 4.15: Bit stability ratio as a function of the different activities**

For nonvolatile applications bit-cell stability is usually a crucial factor and is preferred to be over 60 [27] . The thermal stability factor (Δ) is defined as [27]:

$$\Delta = \frac{K_u V}{K_B T}$$

where $K_u$ is the anisotropy energy density and V is the free layer volume. Thermal stability degrades at a high rate at advanced technology nodes [1]. With our assumptions at 65nm technology for example, we evaluate the rate of change of the stability factor with respect to a base case temperature of 52$^o$C (the base temperature used in this study for evaluation). As the Joule heating corresponding to different activity rates are different, so will be the temperature and consequently stability ratios. Hence, the stability ratio will depend on the rate of activity (Fig 4.15).

## 4.6 Conclusion

We have presented a detailed analysis of the self-heating effect in STTRAM cell and its impact on operational reliability. The self-heating has been studied with detailed FVM

simulations and the results have been used for mixed mode device simulations. We studied the effect of material properties, read-write access patterns on the thermal profile. We have first observed that operational reliability parameters, such as read/write margin and detection accuracy are strong functions of temperature. Next, we have shown that high write current density, small MTJ volume, and surrounding dielectrics can result in high local power density and hence, self-heating in STTRAM cell. Finally, we observed that due to self-heating and inherent temperature dependence of cell parameters, there exist a correlation between read-write history of a cell and its operational reliability. Our cell level study suggests that self-heating can have strong impact on the reliable STTRAM operation and hence, needs careful analysis. The next work needs to investigate the detection properties of STTRAM in presence of self-heating effect and process variation and suggest improvement.

# CHAPTER 5

# READING AND SENSING ACCURACY OF STTRAM

For cache applications or other read-heavy applications, reliable read is the prime concern along with energy efficiency. Though standby leakage is non-existent in STTRAM, leakage during reading operation exists. This leakage if not accounted for can cause read failures in presence of variations and impede STTRAM scaling in the sub 90nm regime. Hence an accurate estimation of the read related challenges needs to be made and solutions suggested for a seamless read operation.

## 5.1 Detection Challenge in STTRAM

The circuit induced challenges to reliability and write-current scaling of Spin-Torque-Transfer Random Access Memory (STTRAM) needs to be evaluated. We show that at sub 90nm nodes, increased transistor leakage increases the probability of incorrect sensing requiring higher read current. But higher read current can increase the read disturb failure, particularly with reduced write current. Using the ITRS predicted

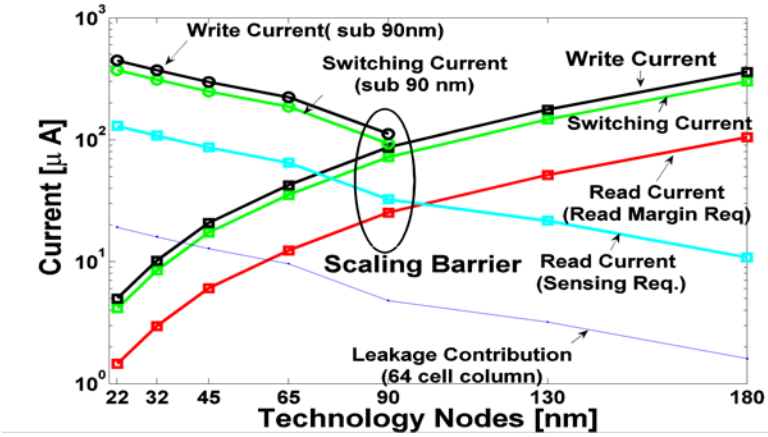



**Fig 5.1: STTRAM scaling challenge**

transistor parameters and MTJ scaling predictions from [78], we project that in sub-65nm nodes the device leakage can limit the minimum read current and hence, the write current scaling (Fig. 5.1). Thus reliable sensing in presence of increased leakage is a potential barrier to STTRAM scaling for sub-90nm nodes [79](Fig. 5.1). To satisfy the conflicting requirements read margin and sensing accuracy, we propose a source line biasing scheme.

Simulations in predictive 65nm node shows that the proposed solution simultaneously reduce the sensing errors and improve read margin. The continuous scaling of STTRAM in sub-90nm nodes requires reduction of the write-current to reduce write energy and write latency. The variation in transistor and MTJ parameters results in statistical variations in MTJ switching current and cell read currents (i.e. current flowing through the cell during read operation). If due to variation, cell read current increases beyond the MTJ switching current read disturb (flipping of the cell content while reading) failures can occur [64] [Fig. 5.2(a)]. To reduce read disturb failures, the cell read current needs to be sufficiently smaller than the switching current (the difference is defined as read

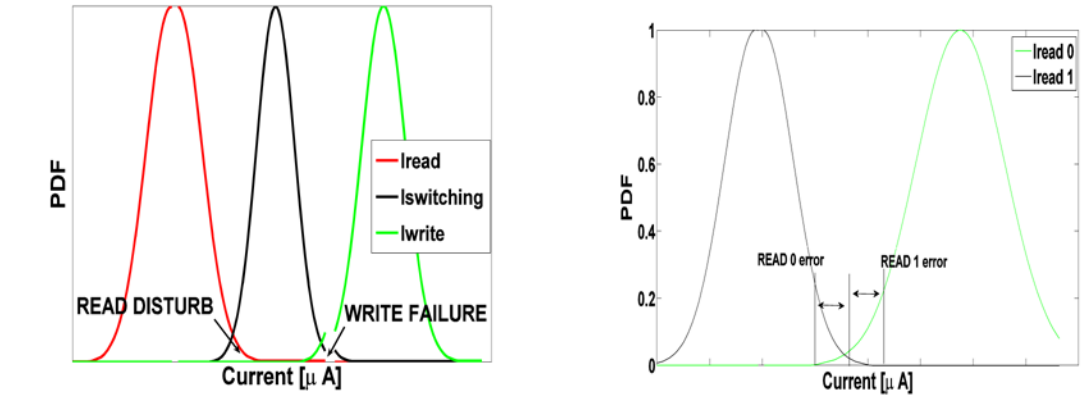
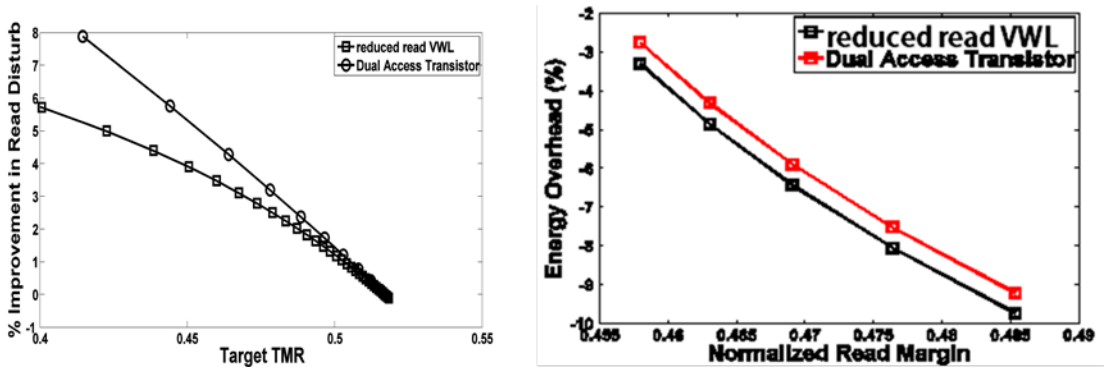



**Fig 5.2: (a) Read disturb and write failures, and (b) Incorrect sensing in STTRAM**




**(a)** **(b)**

**Fig. 5.3: Reduced read $V_{WL}$ and dual access transistor schemes (a) Read disturb and (b) Energy consideration**

margin). Hence, the read margin requirement requires that read current needs to be scaled along with the write current. Several previous works have investigated methods to dynamically reduce the read current [80, 81]. However, the interaction of the read current and sensing accuracy have not been analyzed.

In this work, we analyze the impact of the transistor leakage on the read current and sensing accuracy. We show that leakage current of the unselected cells in a selected column adds a circuit induced noise to the array read current (the total current flowing through the selected column = cell read current + total leakage of the unselected cells). The process induced variations in the leakage current results in significant variations in the array read current leading to incorrect sensing [Fig. 5.2(b)]. Hence, maintaining sensing accuracy under this condition requires higher cell read current and imposes a constraint on the read current scaling. As the leakage current increases with technology scaling, the minimum read current required for reliable sensing also increases. We propose a robust reading scheme using dual source line bias (DSLB) for STTRAM array to improve the sensing accuracy with minimal impact on read margin and performance. The proposed technique uses a common voltage level (supply voltage) for bit line and

word line during read and write operation. However, the source-line is biased at a positive voltage during reading to reduce the leakage current of the unselected cells and improve the sensing accuracy. The non-zero source line bias also reduces the strength of the access transistors during reading to improve read margin. The impact of this approach on robustness, energy, and scalability of the STTRAM array is analyzed considering both transistor and MTJ variability.

## 5.2 Simulation Environment

Simulations are performed with MTJ of R-A product 30 $\Omega/\mu m^2$ at $65^oC$. The sizes of the MTJ devices have been taken as $50X90nm^2$ which requires approximately 90μA of switching current assuming current density of $10^6$ $A/cm^2$. The high and low resistance states are represented by 11.1k Ω and 6.67k Ω respectively. The transistors are simulated using BSIM4 predictive models at 65 nm technology [33, 82]. A supply voltage of 1V was used for simulation. Both writing "1" and writing "0" conditions are simulated to obtain the corresponding transistor widths.

## 5.3 Alternative Techniques: Dual Word line Voltage for Read Margin

The read margin can be improved by reducing the read current by dynamically reducing VBL or increasing $R_{FET}$ during read operation. The $V_{BL}$ control is a simpler approach and it does not impact the distiguishability. However, simulations show that for a 65nm cell lower than 100mV of bit line voltage is required to achieve 50% normalized read margin. Bit line swing over this small voltage can result in high change in sensing delay and low robustness of the sense amplifier [83]. $R_{FET}$ can be dynamically modified either by reducing the word line voltage (referred to as reduced read $V_{WL}$) or access transistor width during reading [80, 81]. The dynamic modification of W (referred to as

dual access transistor approach) can be implemented using two access transistor per cell - both will be ‘on’ for writing and only one will be ‘on’ for reading. But this approach will increase the overall cell area and the switching energy of word line (metal capacitance of two word lines instead of one) and bit line (increased parasitic capacitances of two devices). Hence, although both approaches improve read margin [Fig. 5.3(a)] we consider the reduced read $V_{WL}$ approach as it lower cell area, higher array density and lower energy for a target read margin [Fig. 5.3(b)].

However, as reduced read $V_{WL}$ approach requires three voltage levels ($V_{WL}$=$V_{DD}$ and $V_{BL}$ or $V_{SL}$=$V_{DD}$ for write, reduced $V_{WL}$ for read, and small $V_{BL}$ for read) instead of two levels required for only $V_{BL}$ control. We propose to eliminate this additional level and operate with two voltage levels – the nominal supply ($V_{DD}$) and a lower voltage ($V_{read}$). The word line and the bit line have the same voltage level both during read and write operation. This common voltage is $V_{DD}$ during write operation to ensure high switching current and reduced to a lower voltage level ($V_{read}$ <$V_{DD}$) to maintain read margin (source-line is at 0V) during reading. We refer to this approach as dual word line voltage approach (DWLV) to distinguish it from reduced read $V_{WL}$ approach.

### 5.4 Proposed Technique: Dual Source Line Bias (DSLB)

The primary challenge with DWLV approach is the leakage current through the unselected cells during reading which degrades the sensing accuracy. To simultaneously address the distinguishability and read margin challenges, we propose dual-source-line-bias (DSLB) technique. The voltage configuration during write in DSLB and DWLV approach are same. But during reading, instead of applying 0V at source-line and Vread at wordline/bitline, we propose to use VDD for bitline/wordline and apply a positive bias

Vsb (= VDD – Vread) at the source line (see Fig. 5.4). Note Vsb (= VDD-Vread) maintains the same gate-to-source (i.e. gate overdrive = Vread) and drain-to-source (bitline to source-line = Vread) voltage for the accessed cell as in the case of DWLV with common voltage Vread. This ensures the read current (hence, read performance) and the read margin of the accessed cell with DSLB and DWLV are similar. However, a higher source-line results in a negative Vgs for the unselected cells thereby significantly reducing their leakage. This reduces the $I_{leakage}/I_{cell}$ ratio and improves ATMR. Hence, DSLB uses two voltage levels (VDD and Vsb) similar to DWLV; has read margin/performance comparable to DWLV; but provides better sensing accuracy (i.e. higher ATMR). At this point we proceed to study the trade-offs involved in the choice of the source bias. The next subsection talks on the source bias choices permissible based upon the power and failure requirements.

### 5.4.1 Choice of Source Bias

The choice of the source bias depends on its impact on ATMR. For a given VDD, a

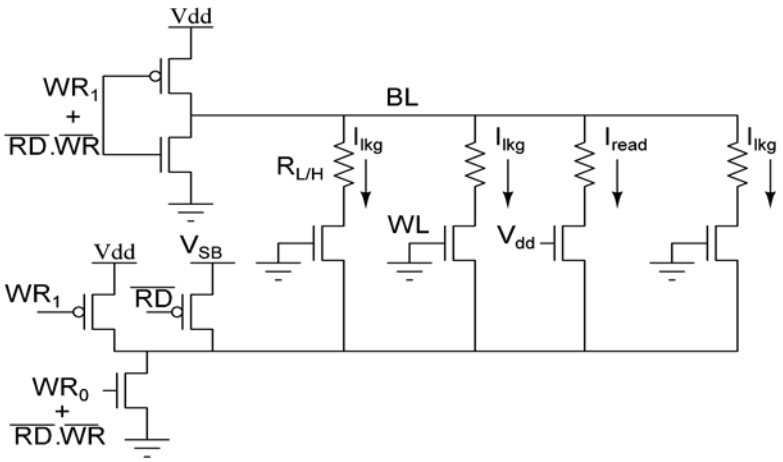
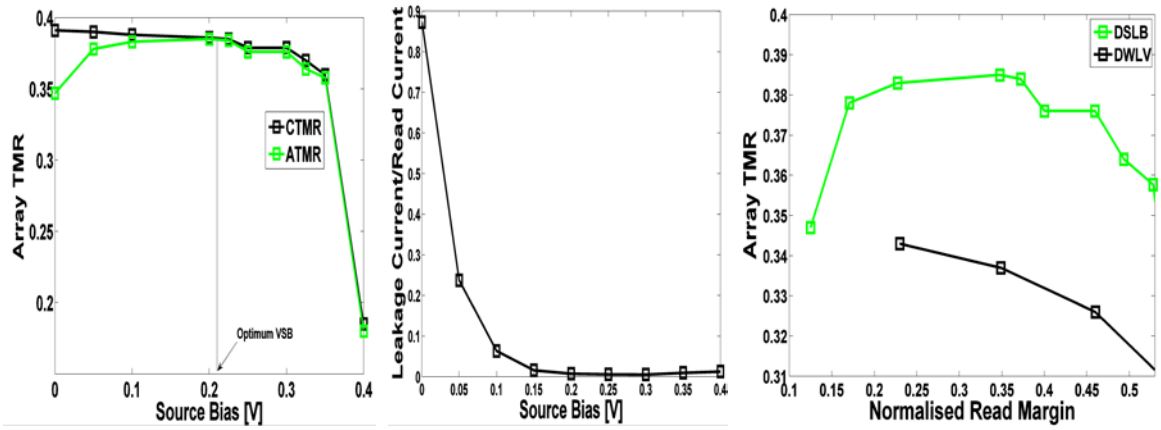



**Fig 5.4: Dual source –line-bias scheme showing leakage**

higher source bias reduces leakage. However, it also increases the Vth of the transistor of the selected cell (negative body bias) and reduces its gate over-drive (i.e. lower Vread) both of which reduces cell read current and CTMR. Fig 5.5(a) shows that there is an optimum VSB for maximum ATMR. At too low VSB the leakage contribution is considerable which reduces ATMR [Fig 5.5(b)]. At too high VSB, read current reduces significantly resulting in very low CTMR. Fig. 5.5(c) shows that for a given read margin (i.e. cell read current) DSLB scheme is able to provide almost 50% improvements in the ATMR compared to DWLV scheme by reducing the leakage. Similarly, for a target ATMR a higher cell read current is required in the conventional case which reduces the read margin. A variation in the bias value around the optimal point has minimal impact on the ATMR or read margin [Fig. 5.5(a)].

## 5.5 Distinguishability under Variation

We study the effect of the transistor threshold voltage variation and MTJ thickness variation on distiguishability (i.e. ATMR) with DSLB. First, the effect of the dielectric

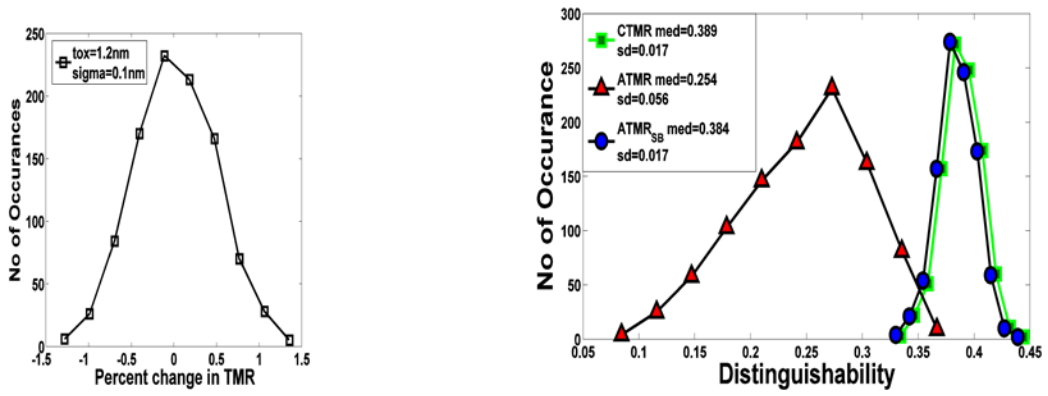



**(a)** **(b)** **(c)**

**Fig 5.5: DSLB: (a)ATMR at Vdd=0.8V (b) Leakage to read current ratio(c) Read margin-ATMR contour**

thickness variation is studied for MTJ dimensions 50X90 $nm^2$ with 1.2nm oxide thickness using [84]. Fig. 5.6(a) shows that the variation in TMR due to dielectric thickness variation is very small. The variation in resistance with variation in dielectric thickness does not affect the leakage significantly as it is practically independent of the drain voltage. We simulated the DSLB scheme with the worst case 1% variation in dielectric thickness to include the MTJ variation effects. Next, we consider the Vth variations in the transistors due to effects such as random dopant fluctuations. We perform Monte-Carlo simulations for a 256 column STTRAM array at 65nm technology with nominal Vth=340mV and σVth=30 mV. We consider both DWLV and the proposed DSLB techniques. The DWLV and DSLB cells are designed for same nominal read current (i.e. same nominal read margin).

As expected, the Vth variations in transistor result in statistical variations in CTMR




**(a)** **(b)**

**Fig 5.6: Effect of variations on DSLB: (a) TMR variation with oxide thickness, (b) ATMR distribution**

[Fig. 5.6(b)]. Further, due to the exponential dependence of leakage on threshold voltage, Vth variations in the transistors results in large variations in ATMR [Fig. 5.6(b)]. We noticed a 42.5% and 39.8% deviation of the mean of ATMR from the device TMR and

the mean CTMR, respectively for the DWLV array [Fig. 5.6(b)]. However, due to the reduced leakage current, ATMR distribution almost approaches the CTMR distribution with DSLB [Fig. 5.6(b)]. With increasing threshold voltage variation both the median and the minimum ATMR degrades significantly as the leakage variations increase [Fig. 5.7(a)]. This results in significant increase in median ATMR and reduction in the ATMR spread [Fig. 5.7(a)]. The increase in temperature also exponentially increases the leakage current and degrades the ATMR [Fig. 5.7(b)]. DSLB significantly improves the ATMR and reduces its spread at high temperature.

## 5.6 Sensing Accuracy and Performance

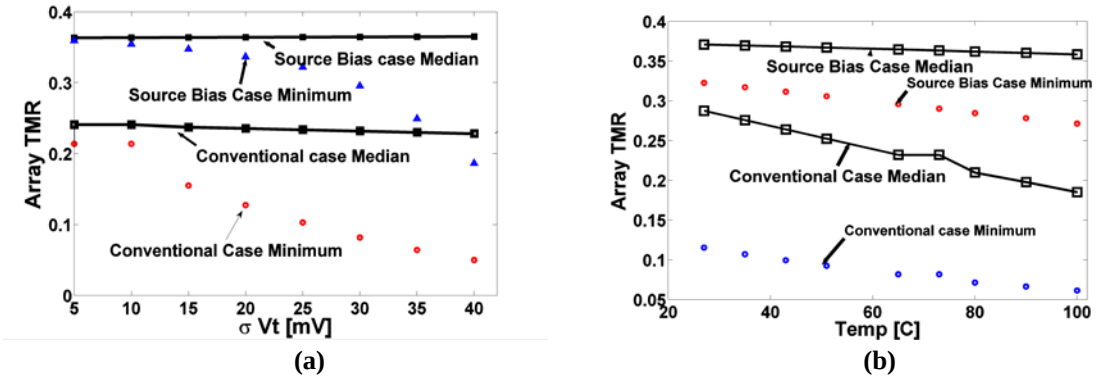



**Fig 5.7: Effect of variations on DSLB: (a) Effect of threshold variation, and (b) Effect of temperature variation**

In this section we analyze the impact of DSLB on read performance considering the voltage sensing based scheme proposed in [36]. In this scheme, during reading the bit line voltage drops at faster rate for state “0” (parallel or low-resistance) than “1” (anti-parallel or high resistance) [Fig. 5.8, DWLV without process variations]. The reference voltage at different time points is determined from the average of these deterministic bit line voltages for “1” and “0” cases. The voltage difference between the bit line voltage and the reference voltage at the time of sensing (i.e. read access time) is referred to as the bit-

difference. For correct sensing, bit-difference after the read access time needs to be higher than the offset voltage (minimum input difference required for correct sensing) of the sense amplifier. If process variation is considered, due to variation in both the read current and the leakage of the unselected cells, there can be significant spread in the bit line voltage drop (hence, bit-difference) at a given time [Fig. 5.7(b)]. Hence, there is a finite probability, defined as the probability of sensing error, that the bit-difference after the read access time for read "0" or read "1" is less than the offset voltage of the sense amplifier.

If process variation is neglected, the bit line voltage with DWLV drops marginally

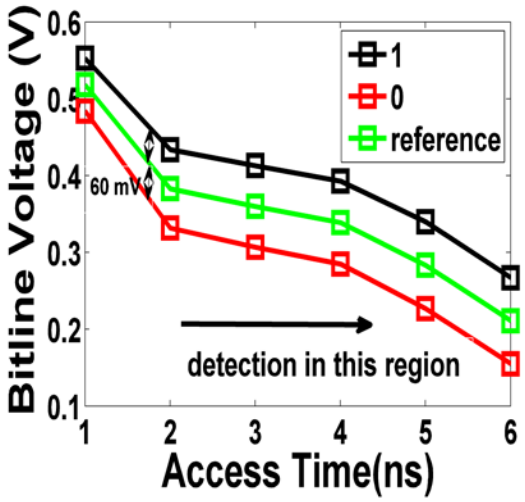
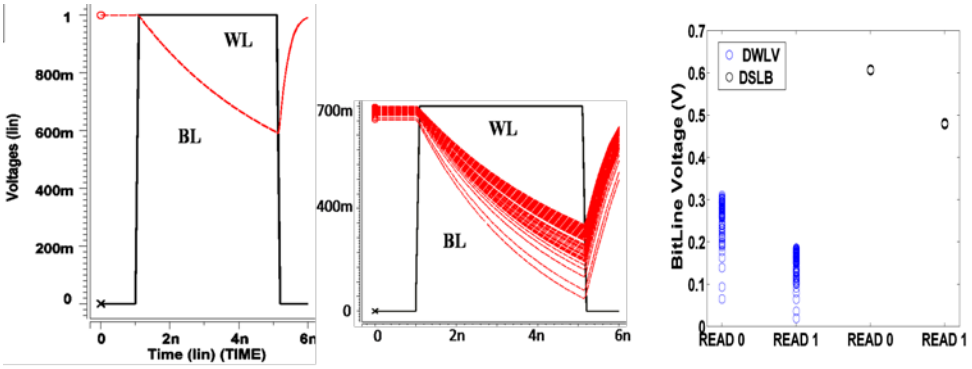



**Fig 5.8: Bit line values with no variation**

faster than that with DSLB as reverse body bias increases the Vth of the access transistor in DSLB. Note gate overdrive of the access transistor is Vread in both DWLV and DSLB to ensure similar read margin. In other words, if process variation is neglected, the performance of DSLB can be marginally worse. However, when process variation is considered the DWLV case has a large variation in the bit line voltage drop at a given time during both read "0" and read "1" [Fig. 5.9(b)]. However, even under process

variation, the bit line voltage drop for DSLB experience negligible variations as the leakage contribution is significantly reduced [Fig. 5.9(a), (c)]. This results in a very low sensing error.

## 5.7 Results with DSLB




**(a)** **(b)** **(c)**

**Fig 5.9: Bit line Values with (a) considering inter-die variation for DSLB (b) considering inter-die variation for DWLV (c) bit line distribution (DSLB gives almost no spread)**

### 5.7.1 DSLB: Sensing Error

We estimate the probability of sensing error for DWLV and DSLB cases considering different read access time (i.e. sensing time). For simulations we consider ±60mV sense amplifier offset and σVth=30mV. Fig 5.10(a) shows between 2-2.2 ns, both the scheme has high error probability and DSLB has higher error probability than DWLV (as the voltage drop in the DSLB scheme is slower). Beyond that, error probability of DSLB falls off sharply but that for DWLV remains appreciable due to high leakage contribution. In other words, for a target error probability (<0.5) DSLB requires much less read access time than the DWLV i.e. read performance of DSLB is better than DWLV for reliable read operation. Further, a lower magnitude of bit line voltages for DWLV also implies a significantly increased delay for the sense-amplifier.

### 5.7.2 DSLB: Read Margin

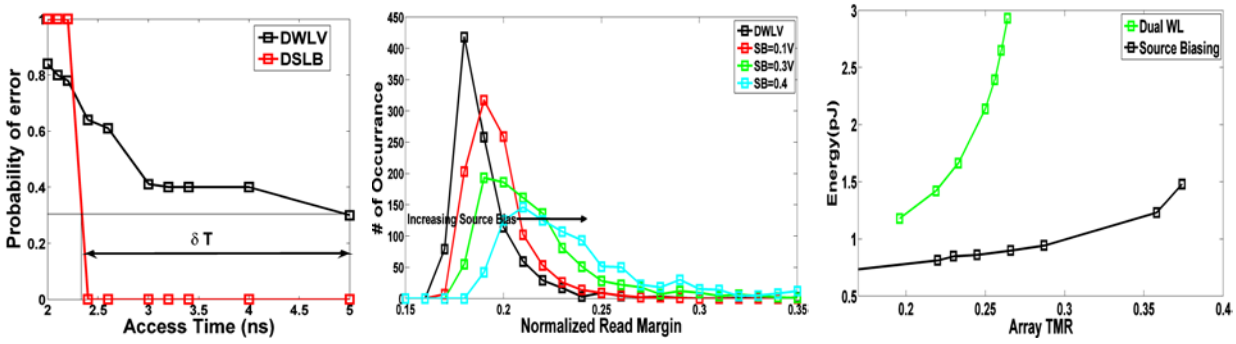
With source bias, threshold voltage of the transistor rises. Hence under variation the mean value of the read current reduces with source biasing (Fig. 5.10(b)). This leads to betterment of read margin. As source bias allows operation at reduced read currents nullifying effect of leakage, this read margin can be sustained.

### 5.7.3 DSLB: Energy Impact

Conventional read operation will not allow us to operate below a certain write current at sub 90nm nodes. Source biasing lowers the sensing constraint on read current. And allows us operate at lower write currents. Write energy is the dominant component of energy in STTRAM, being up to 10X times the read energy. For a given ATMR read currents in both cases were found and the write current is extrapolated from them using the read margin values. Fig 5.10(c) shows that the source biasing solution has considerable energy benefits.

## 5.8 Conclusion

In this chapter we have tried to understand the read related reliability challenges of




**(a)** **(b)** **(c)**

**Fig 5.10: Effect of DSLB on: (a) Sensing error and read performance, (b) Read margin, and (c) Energy**

STTRAM for usage in read heavy applications. Now that ways to improve readability has been suggested, the STTRAM design methodology can be used for different applications. To a first order, read heavy applications would seem to profit from the methodology. Reconfigurable frameworks are immensely read heavy and one has to write at the mapping phase. The writes are concentrated over a small phase of time. We propose to study the benefits of adopting the methodology in a reconfigurable application domain in the next chapter.

# CHAPTER 6

# STTRAM CIRCUIT/ARCHITECTURE CO-OPTIMIZATION TECHNIQUES FOR MBC

## 6.1. Introduction

Spin transfer torque random access memory (STTRAM) is an impressive candidate for Memory based computing (MBC) [47]. STTRAM benefits from its ultra-low leakage and dense layout. For the following reasons an STT-based MBC system is lucrative: (1) the non-volatile nature of STTRAM cells alleviates the requirement of reconfiguration at power-up, (2) high-density integration, (3) high read speed and low read power, (4) near-zero standby leakage, and (5) radiation hardness.

For this evaluation, the energy model developed in Section 3.3 is used. As pointed out earlier, distinct read and write ratios point to different STTRAM design solutions. Since MBC lies on the extreme end of the read spectrum, it is expected that the solution will be a custom one. The energy benefit obtained by using a custom design methodology is evaluated. Additionally, an energy distinction exists between reading “1” and reading “0”. Content mapping, which is based on this finding, can further boost energy efficiency. The STTRAM based MBC exploration is a collaborative work. It has been done in collaboration with Somnath Paul et al, from Case Western Reserve University. The STTRAM design related innovation is my contribution. The mapping algorithm and system evaluation are their contributions [45, 85].

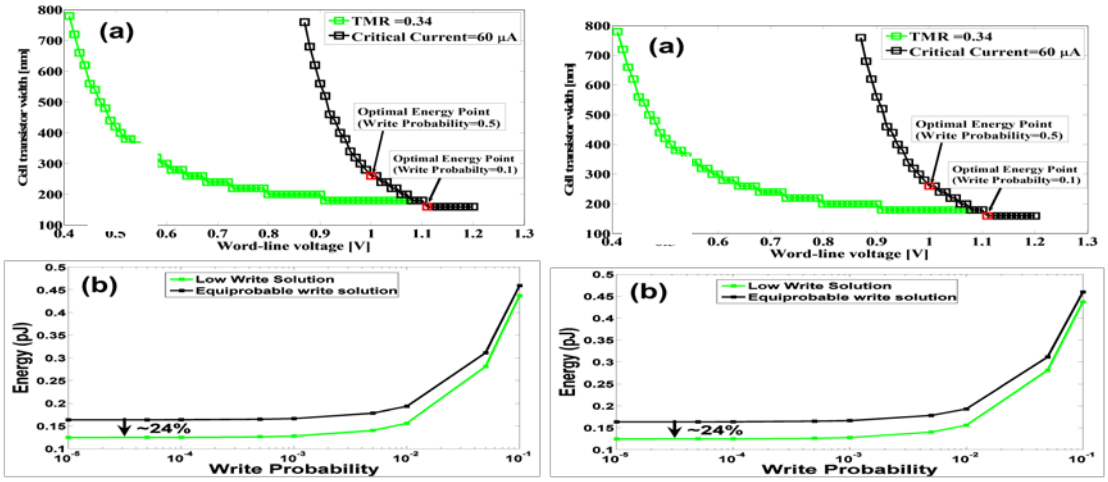
## 6.2. Energy Optimality in STTRAM

The design space for STTRAM is constrained by the readability and writability conditions i.e., the tunnel magneto-resistance (TMR) ratio and the write current requirement. Two STTRAM parameters (the access MOSFET width (W) and its word line voltage ($V_{WL}$)) can be used to navigate the design space of TMR and write current. To minimize the energy dissipation, choice of the energy optimal point in the W-$V_{WL}$ plane is proposed. This is shown in Fig. 6.1. The total energy is evaluated considering the write/read current through the MTJ-transistor structure and the switching energy associated with the word line and bit line. A key aspect of the solution is its dependence on the write probability.

Two solutions, corresponding to write probabilities of 0.5 and 0.1 respectively, are shown in Fig. 6.1(a). These two solutions are different because different write probabilities result in different read and write energies. A larger write probability means a solution with larger width and smaller $V_{WL}$ because write has a quadratic dependence on $V_{WL}$. From the read perspective, a solution with a lower width is preferred because of its low leakage power dissipation. For MBC with a read-dominant access pattern, the W-$V_{WL}$ configuration corresponding to equiprobable condition is not an optimal choice since it dissipates high read energy, as shown in Fig. 6.1(b).

Hence, we chose the optimal-energy point corresponding to low write probabilities, which provides a much lower read energy at the expense of increased write energy compared to the equiprobable case. As shown in Fig. 6.1 (b), a 24% saving in total energy for the write probability of $10^{-5}$ for an 8-bit 64x64 memory array with a read access time of 400ps is achieved.

## 6.3. Energy Distinction between Read “0” and “1”



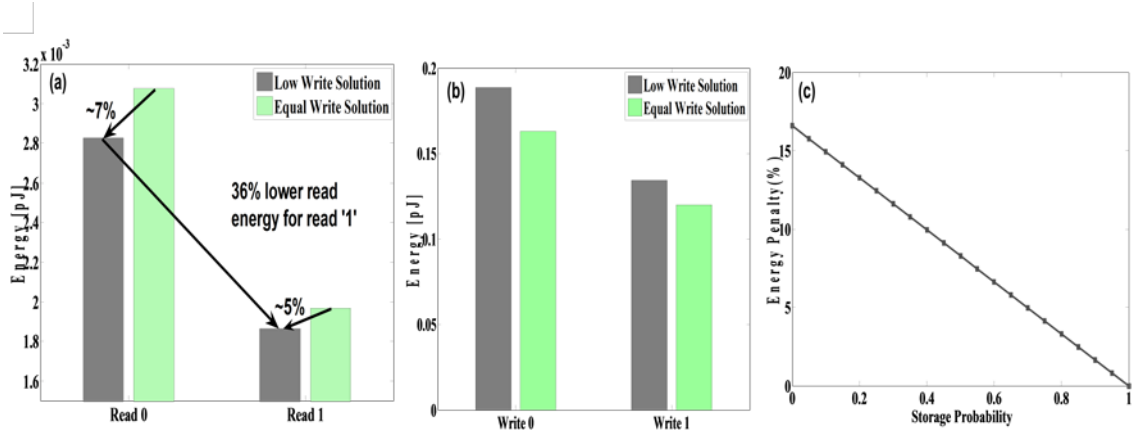

**(a)** **(b)**

**Fig. 6.1 (a) Design of STTRAM cell for MBC framework to achieve optimal read energy; (b) Read energy with varying write probability.**

From the STTRAM read operation, it can be concluded that a larger current flows in the circuit corresponding to Read "0" than Read "1". This disparity exists because resistance of State "0" is lower than State "1". This is shown in Fig. 6.2. Hence, there exists an energy difference between reading "0" and "1", which can be exploited to our benefit. There is 36% difference between energy dissipated in the Read "1" and Read "0" operations, as shown in Fig. 6.2(a). Additionally, write "1" also consumes less energy than does write "0", as shown in Fig. 6.2(b). However, for the MBC application intended, the system is heavily biased towards read. The write probability lies in the range of $10^{-3}$-$10^{-5}$. Hence, for MBC application, we conclude that the system should be biased to handle more Read "1"s than Read"0"s. With the proper design of STTRAM for MBC, considerable energy savings can be obtained.

Due to higher read power during a Read "0" operation, it is intended that the STTRAM array contain more Logic "1" than Logic "0". Considering this asymmetry, we have developed a preferential mapping approach to skew the LUTs to contain more Logic "1" than Logic "0". The preferential application mapping scheme amplifies the energy

savings by storing more Logic “1” than Logic “0” in the schedule and function tables. The greedy heuristic for skewing the Logic “0” to Logic “1” ratio in the LUTs is presented in [44]. The heuristic is used to exploit the lower read “1” power in the STTRAM based MBC framework. The effectiveness of the preferential mapping approach was validated for a set of standard benchmark circuits. For the selected benchmark circuits, the preferential mapping heuristic was observed to achieve about 49% increase in the Logic “1” count stored in the LUTs.




**(a)** **(b)** **(c)**

**Fig. 6.2 (a) Read for a cell storing logic “0” and “1”; (b) Write energy for a cell storing logic “0” and “1”; (c) Increase in read energy with increasing probability of storing “1”.**

From these discussions, we conclude that as the number of “1”s stored in the array increases, the energy advantage will keep on increasing. A study on STTRAM array energy with varying probability of “1” storage is shown in Fig. 6.2(c). It points to the fact that a solution with all zero storage will result in a 16% energy-access overhead compared to the case when all ones are stored in the array.

## 6.4. Results

Simulations are performed at the 65nm CMOS technology node. The resistance area product for the MTJ was 30Ω-μm$^2$. The sizes of the MTJ devices have been taken as

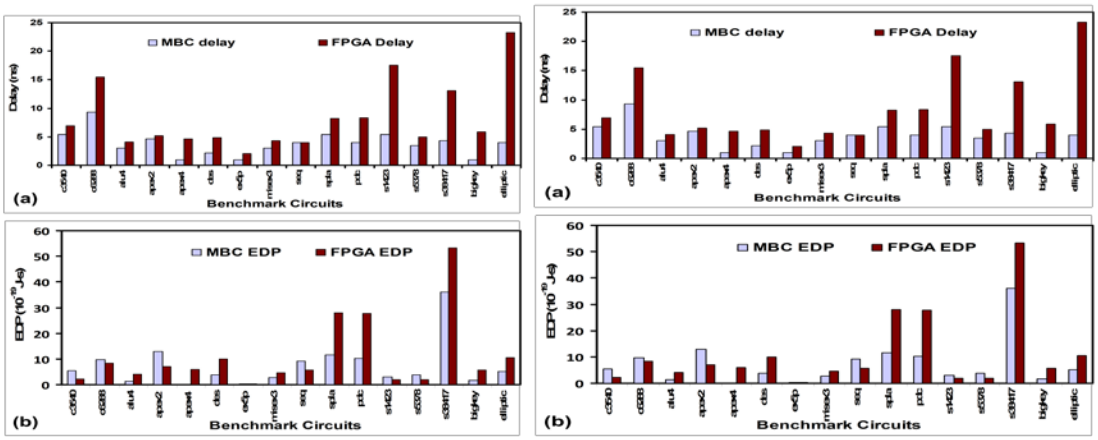
$50x90nm^2$. The MTJ requires approximately 60μA of switching current, assuming a current density of $10^6 A/cm^2$. The high-resistance and low-resistance states are represented by 11.1kΩ and 6.67kΩ respectively. The transistor was designed to drive the required switching current under both Write "0" and Write "1" conditions.

To obtain the solution for varying write probabilities, first a host of simulations with the high and low resistance are performed for a range of $V_{WL}$ and W. The solution space has to be extracted from the generated design space considering the constraints on minimum TMR and switching current. In this work, we consider minimum TMR and switching current requirements of 0.34 and 60μA respectively. Corresponding to this extracted feasible design space of $V_{WL}$ and W, we evaluate the read, write, active leakage, and total energy of STTRAM array. We select write probabilities of 0.5 and 0.1 to evaluate the design points. The read and write energies for storing "0" and storing "1" are computed for these design points.

The delay and energy requirement for the CMOS elements of the MCB were obtained through SPICE simulations using BSIM4 predictive models at 65nm technology [86]. A supply voltage of 1V was used for simulation. The cycle time for a given benchmark in the MBC framework depends on the delay through the programmable interconnects. Hence, the interconnect delays for both MBC and FPGA frameworks were obtained from the VPR toolset [87]. A 65nm FPGA model was used to simulate the performance of the programmable interconnects.

For standard benchmark circuits, the MBC framework improves performance by 45.4%, as shown in Fig. 6.3(a). The EDP values between the two frameworks (MBC and FPGA) are compared in Fig. 6.3(b). The non-volatile MBC framework achieves a 5%

improvement in EDP over the CMOS FPGA framework. The performance and EDP computation includes the cell optimization for read operation. The EDP improvement is further enhanced through the preferential mapping step, which skews the LUT contents to have more Logic "1" than Logic "0"s. As a result of this preferential mapping, the average EDP improvement was calculated to increase from 5% to 6.64%.




**(a)** **(b)**

**Fig. 6.3(a) Improvement in delay for STTRAM MBC over conventional SRAM-based FPGA; (b) Energy Delay Product (EDP) of STTRAM MBC and conventional SRAM-based FPGA.**

## 6.5 . Conclusion

An STTRAM design is proposed, which is skewed towards read, to boost the energy savings and performance in MBC. This skewing results in 45.4% performance improvement and 5% EDP improvement over equivalent CMOS FPGA solutions. A preferential mapping approach can further boost the EDP improvement to 6.64%. The application aware design techniques can be extended to the volatile memory domain as well. To demonstrate that, the innovations possible in SRAM design for MBC are studied in detail in the next chapter.

# CHAPTER 7

# ENABLING POWER SAVING IN MBC WITH SRAM

Reconfigurable computing platforms use hardware and software simultaneously thereby offering advantages of reduced design cost, configurability and faster time to design. The programmable hardware is set in accordance with the software algorithm for mapping. FPGA is the most prevalent spatial computing platform exhibiting reconfigurability. For mapping in FPGA, the function to be implemented is broken up into several multiple input-output functions whose connectivity is controlled through software. However, the reconfigurability comes at the cost of 3X speed reduction and 2X power increment over equivalent ASIC implementations. The primary component hindering speed and power scaling is the programmable interconnect structure. The dense resistive interconnect structure leads to delay degradation. With technology scaling, the interconnect delay becomes an increasing fraction of the total delay. Thereby the speed-power degradation in FPGA keeps growing with scaling. Hence, there is scope for improvement with the traditional FPGA structure.

## 7.1. MBC as a Solution and the Importance of Memory

A solution to this problem might be reduced interconnects. If individual computing blocks are made computationally more efficient, the dependence on the interconnect structure will be reduced to a large extent. Another solution could be to enable multi-cycle calculations in a block. Thereby, the spatial computing requirement will be reduced further. This will result in lowering the contribution of interconnect delay further. Memory based computing (MBC) is a spatial computing platform utilizing time-

multiplexing in the blocks. The computation is performed in unit blocks called memory logic block (MLB) which can communicate with each other. The lookup is based on a 2D memory array called look up table (LUT). The computation or parts of the computation can be performed via multi-cycle operations in these MLBs.

A prototype implementation of the unit cell of MBC has been done with 2Kb of on-chip memory (word length is considered to be 8 bits). We will discuss the system in detail as well as the individual details and later in the chapter will show some results for the individual blocks. A single MLB is shown in Fig 7.1 consisting of one 8 bit intermediate register and one 8 bit dummy register. We will explain their roles later.

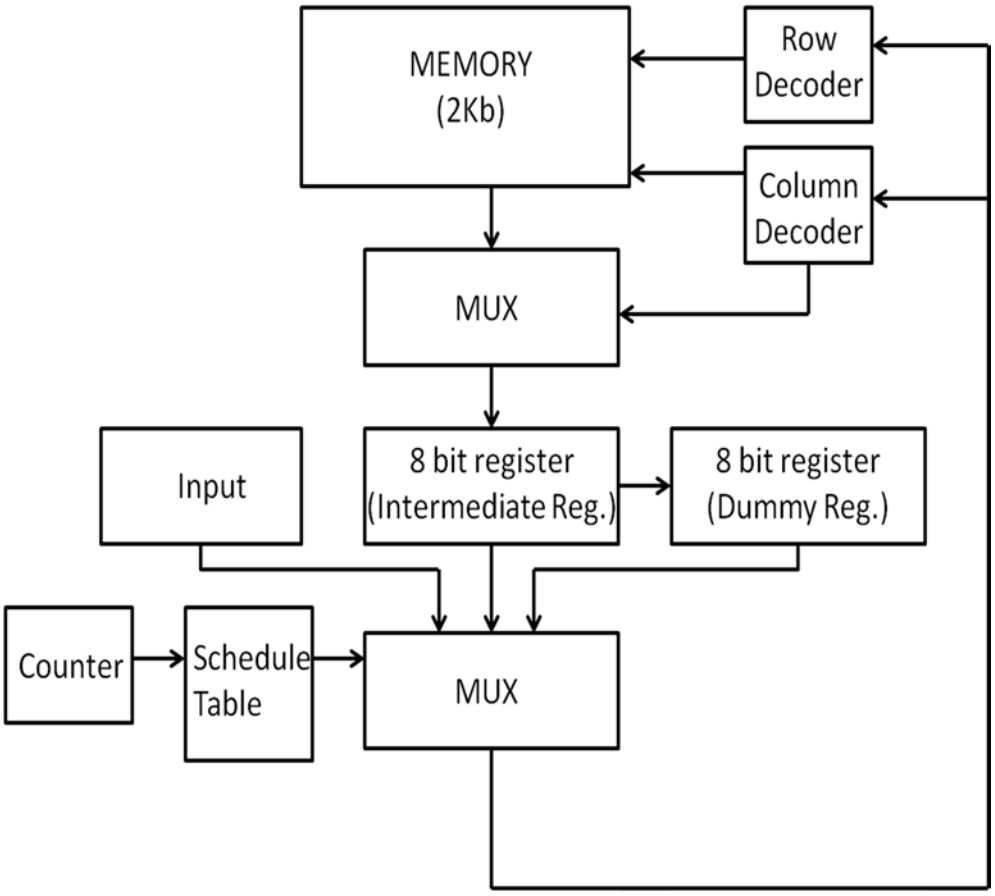



**Fig 7.1: Prototype Memory Logic Block (MLB) for a MBC System**

The schedule table is the heart of the operation in MBC. It is essentially a finite state machine. In our implementation it consists of values stored in registers which are selected according to the state of a counter. The counter is fed with the clock. The selected output from the schedule table serves as the input to the final multiplexer which helps to choose between the intermediate register, the dummy register or the external input. This output goes to the decoders for reading from the memory.

As memory plays a crucial role in determining the delay and power consumption in the lookup based MBC system, changes in memory circuit or architecture are due to be reflected in the performance-power of such a system. Proper choice and design of memory cell can result in reduced EDP for the system. We have already evaluated the applicability of NVRAM based LUT design for EDP reduction. This chapter discusses how the SRAM design can be optimized for the LUT leading to EDP reduction for MBC.

### 7.2. SRAM Design Options: Limitations of Relative Transistor Sizing

In prior literature, there have been efforts aimed at modifying the transistor width ratios towards achieving an improved read scalability [48]. However, the transistor sizing based approach has its own limitations. Read stability requires us to reduce the access device (AX) drive strength and increase the pull down (PD) device drive strength. However, read performance requires the drive strength of both access and pull down devices to increase. Hence, there is a conflict between the sizing requirement for the access device from the read margin and the read speed requirements (seen in Fig. 7.2). This conflict is embedded from the topology and can only be lifted with some insightful topological modification.

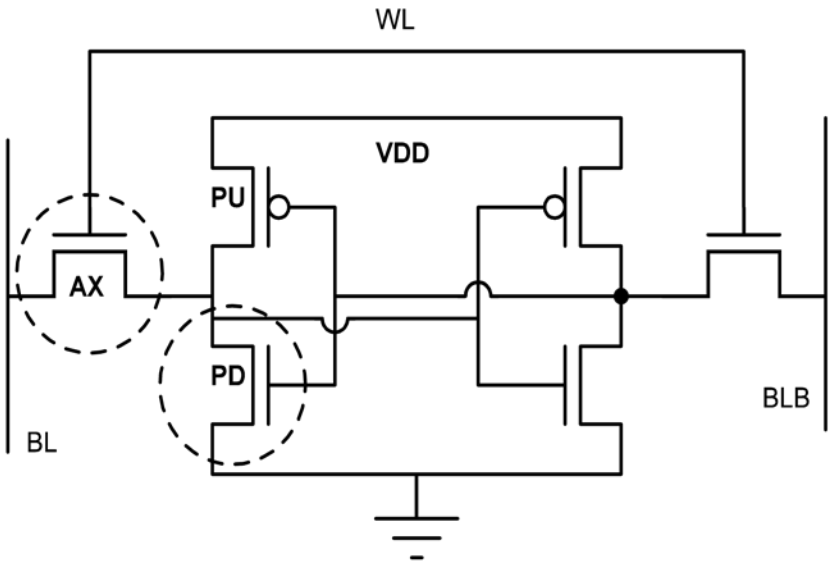



**Fig 7. 2: Read conflict in 6T SRAM**

### 7.3. Choice of Topology for SRAM Design in Low Power Embedded Application

A modification of the 6 transistor SRAM cell was suggested for high-speed, low power applications [49, 50]. One of the main aims of the cell was to reduce the cell power supply for operation. In other words, at scaled power supplies the read stability of the cell is better and this improved read stability comes without sacrificing read speed. As the internal storage node is not connected to the bit line the possibility of read disturbs is eliminated (Fig. 7.3). However, the cell is single-ended and single ended writing to the cell is a tougher challenge. Write is a tougher challenge in terms of reliability and energy consumption. However, the problem with write can be tackled with standard write-assist techniques. In MBC, the system is extremely read heavy (~99.9% of the operations are reads). This is because usual pattern in the reconfigurable platforms is of one-time programming and multiple reads (ratios between the reads and the writes are very highly skewed). In such a read heavy system, the one-time write penalty is very small compared to the read performance and power benefits.

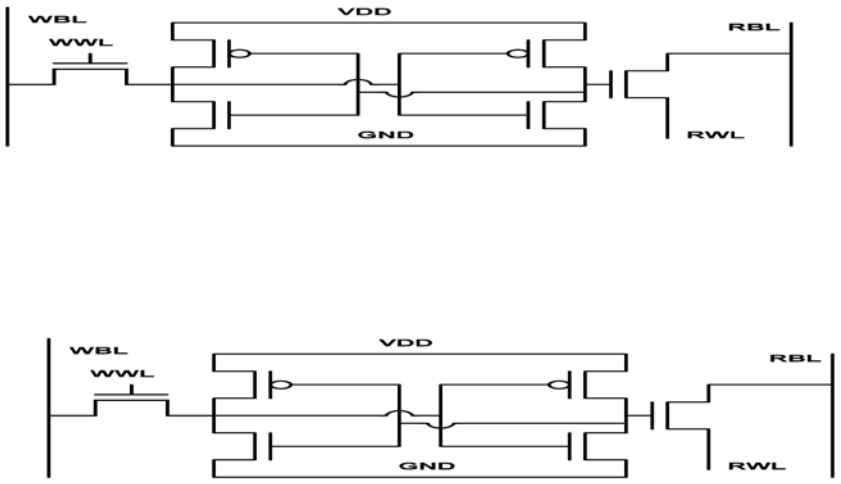
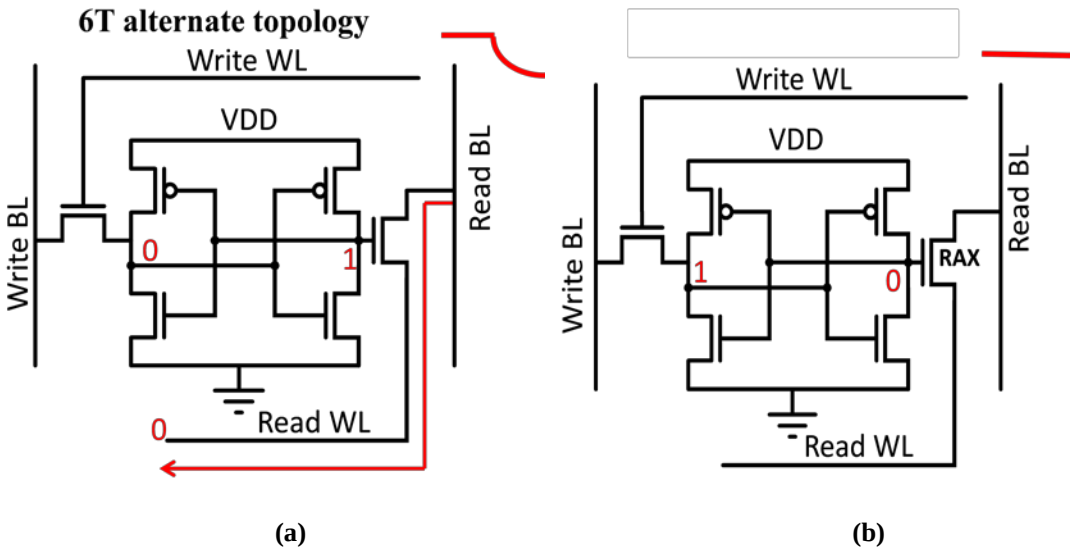



**Fig 7.3: The alternate cell**

If a particular word line is to be read, the RWL is made low as shown in Fig 7.4 (RWL is derived from RE and WL). In presence of read signal and word line selected, read access transistor has its source set to “0”. If the transistor stores a “1” at the storage node driving the read access transistor, the bit line discharges through the read access transistor (Fig 7.4(a)). If it stores a “0” at the storage node driving the read access transistor, the transistor does not conduct and the bit line does not discharge (Fig 7.4(b)). However, there are several modifications and unaddressed challenges for this topology. These need to be addressed before the SRAM cell can be applied for the MBC system.

The cell can be operated across a voltage range of 1.2V to 0.7V. However, with reduced drivability of the access transistor at lower supply voltages falls off rapidly. If performance can be sacrificed for power, we can operate the cell at low voltages. The advantage is that the read disturb concern is minimized in this topology. Thus, even though the cell performance would reduce at lower voltages, it would continue to remain free from read disturb.




**(a)** **(b)**

**Fig 7.4: Reading (a) "1" and (b) "0"**

## 7.4. Discussions on Read Stability

Read stability is measured by read margin. The measurement is performed by sweeping one of the storage nodes and observing the other. The bit line is kept floating and the word lines are on. The side of the largest square inscribed within the lobes of the butterfly curve indicates the maximum noise immunity. In the discussed structure, the write word lines are deactivated for test and the read margin equals the hold margin. Hold margin is the static noise margin in hold mode i.e. when the word lines are shut off. The measured read margin for the structure is 270mV, much higher than the equivalent 6T SRAM (Fig 7.5(a)). This stability advantage becomes a key advantage at scaled voltages where noise margin drops significantly (Fig 7.5(b)).

Read stability results are discussed on a set of 1000 Monte Carlo simulations with $V_{th}$ of 350mV and σ of 17mV. For a 6 transistor SRAM cell of equivalent size, the read margin at 1.2 V shows a mean of 270 mV and a standard deviation of 17 mV. The read margin at 0.7 V shows a mean read margin of 197 mV and a standard deviation of 14mV.

It is to be noted that the spread increases. For the bit cell, this component of the variation is insignificant as there is no scope of charge development or escape from the storage nodes.

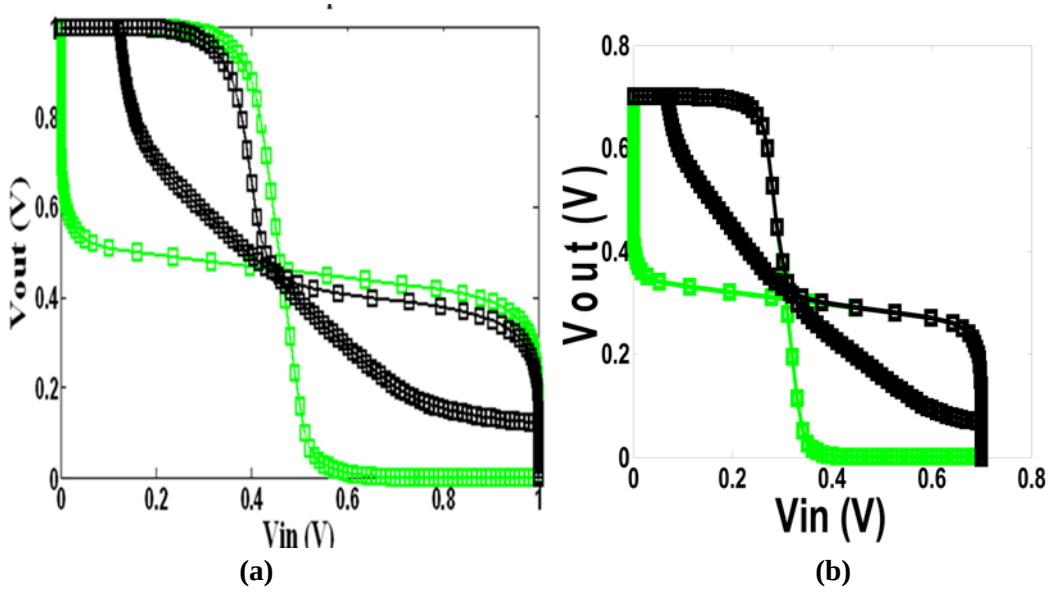
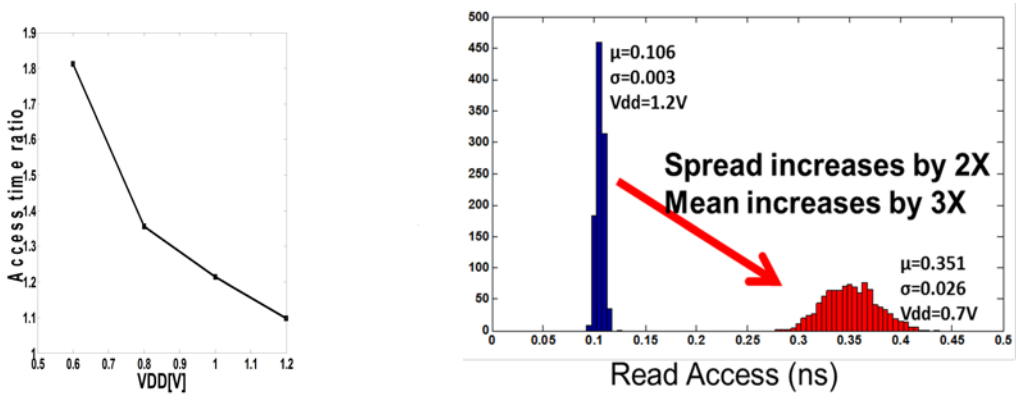



**Fig 7.5: Read static noise margin (a) at 1V (b) at 0.7V for the standard 6T and alternate version**

### 7.5. Discussions on Read Performance

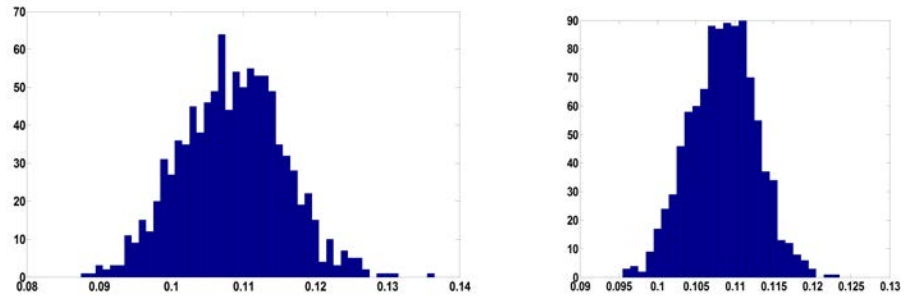
For similar voltage drop, the alternate topology outperforms the standard prototype in terms of access time (Fig 7.6(a)). The access time ratio is close to 1 for a voltage of 1.2V and increases to almost 1.8 for a voltage of 0.6V. The current in case of 6T SRAM read flows through the access and the pull down devices. With reduced power supply, the current is limited by the sinking capacity of the pull down device. For the topologically modified cell, the current flows through the read access device. Hence the impact of scaling on delay for similar drops is less severe. Alternatively the 6T structure offers stacked transistor resistance which is higher than the single transistor in the alternate cell read path.




**(a)** **(b)**

**Fig 7.6: (a) Relative performance of the bit cell (b) Access time variation**

Let us proceed to study the impact of variation on the read performance of both these circuits. For read access time, it is observed for both cells that reducing the power supply from 1.2 V to 0.7V increases the mean by more than 3X while the spread increases by about 2X as shown in Fig 7.6(b). With increased sizing of the read access transistor the RDF component of variation is expected to reduce resulting in lower access time spread. To illustrate the point, simulations are performed with reduced variations across the access device. Results show almost 1.7X reduction in performance variability (Fig 7.7). The upsizing would correspond to 2.89. Hence, upsizing the read access device can help in increasing current drive and reducing variability at the cost of area.



**Read Access Time (ns)** **Read Access Time (ns)**

**(a)** **(b)**

**Fig 7.7: Read performance variability improves with reduced variation in the read access device. (b) shows reduced variability in access device 1.75 reduction in the variation.**

## 7.6. Discussions on Write Performance: Use of Assist Techniques to Reduce Failures

Writes to the system are difficult. This is due to the fact that writes are single-ended. To this end the identification of the worst case consideration is required. This occurs when a “1” is written to a storage node containing a “0” (Fig 7.8). In a standard two sided write, the discharge of the node storing “1” is considered primarily. This occurs earlier and once the voltage level reaches below the trip point of the other inverter, the data is written. Of course, in the mean while the other node is also being subjected to writing “1”. For writes, with word line overdrive the worst case write delay is restricted to 0.2ns. For write, the figures indicate that compared to two ended writes (Fig 7.9(a)), single ended writes end up in ~10X more time with failure at the process corners (0.07% failure rate observed) (Fig 7.9(b)). With circuit assist techniques the failures are eliminated while the writes are about ~5X slower than double ended SRAMs (Fig 7.9(c)).

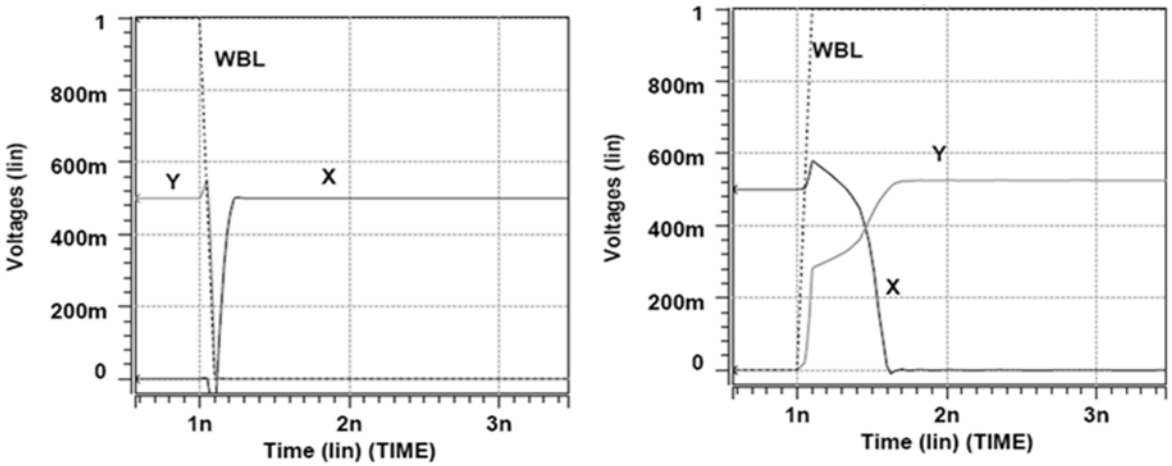
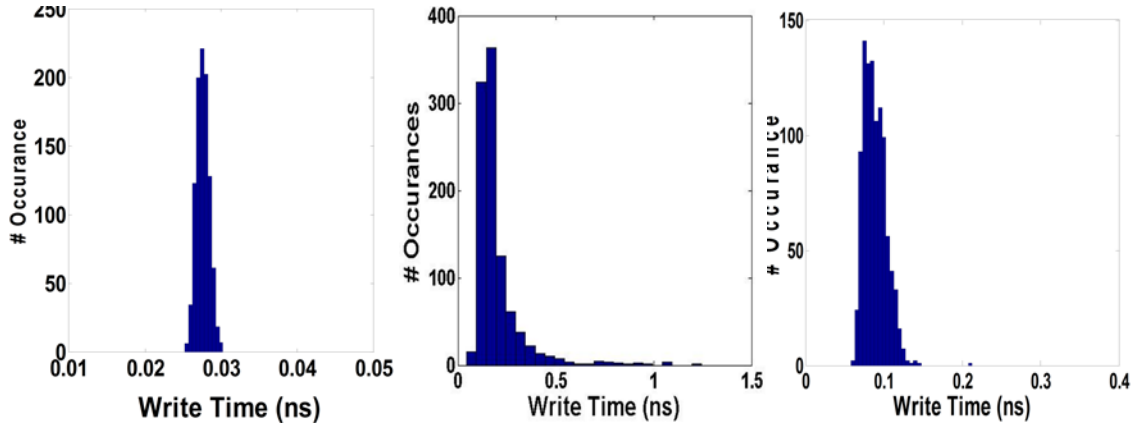



**Fig 7.8: (a) Writing “0” to the cell (b) Writing “1” to the cell**

In this topology the positive feedback loop is to be activated from a single side. Hence, the slower process of writing “1” is to be considered. This is a slower process because the access device beyond a point moves into linear region while the pull down

device competes to drain out the charge. Only sizing could result in a number of failures at the process corners. Hence, we use write assist technique in the form of boosted word line to drive the write word line. Driving the word line at 1.5V lets us accomplish the task of writing to the cell at the worst case process corner for the worst condition at 250 MHz




**(a)** **(b)** **(c)**

**Fig 7.9: (a) Double ended write (b) Single ended write with assist only (0.07% failure) (c) Single ended write with assist and sizing**

Write assist mechanisms are required for single-ended write to avoid write failures and improve write time. However, this forbids the use of bit line interleaving architecture as then the cells which are not chosen stand a chance of being written to. To avoid this, divided word line architecture was used for the cells where the words consisting of 8 bits form a segment. This segment is addressed by the divided write word line. For the proposed design, we intend to write the 8 words of 8 bit each at one go. Hence, the final array is a 64 x 32 bit array i.e. a 2kb array. This array has 8 words each of 8 bit in a divided word line arrangement.

### 7.7. Discussions on Power Requirements for the Cell

In differential structures, there is always a switching involved either for the bit line or its complement. The dynamic component of the switching power (in case of high

or rail-to-rail swing) is large. In the case of the cell, for reading “0” no dynamic power is spent as there is no switching involved. However, for reading “1” power is spent. But on an average the power consumed is lower. Also, depending upon whether the pulsed scheme is used, the bit line might not fully discharge. This will bring down the power consumed further. It is verified that at voltages below 1V, the read power for the proposed cell with a pulsed period for access remains almost 40% less throughout the interval. Having more “0”s instead of “1”s can result in significant power savings as the bit line switching is prevented in this case. A preferential storage structure could amplify the energy savings by storing more number of “0” than “1”.

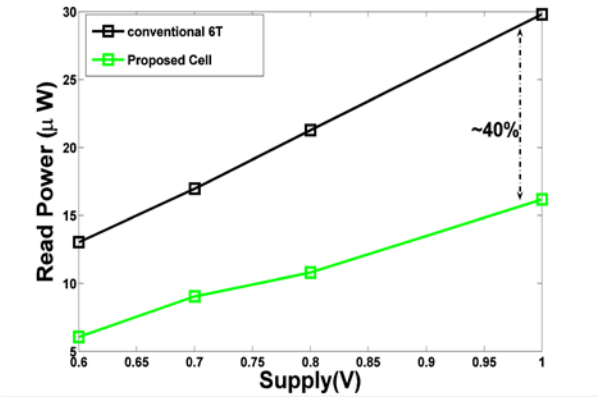



**Fig 7.10: Power savings using the proposed structure**

The alternate cell requires additional word line switching for the word line resulting in a write energy overhead of: $\Delta E/E = (V_{WL}/V_{cell})^2 - 1 = 1.78$. Considering that the target reconfigurable framework has a read dominated access pattern (~99.9% of the operations are read), the total energy savings of this structure is given by: $\Delta E = 0.999\times0.4 - 0.001\times1.78 = 0.397$, where 0.4 denotes the 40% improvement in read energy from Fig. 7.10.

## 7.8. Constraint Based Sizing of the Cell

The cell is meant to serve the read purposes heavily. The read stability of the cell is ensured topologically. Hence, sizing wise the read stability criterion does not need to be taken care of. We have minimum sized PMOS pull-ups and NMOS pull-downs. The access device for read is meant to be large to sink in large current such that the read access time is reduced and ensuring a high drop across the bit line. Also the write transistors have to be big to improve drivability especially for single ended writing. This means larger transistors are to be used for the writing purpose.

However, for read there is an additional overhead. Assuming that the read-word line is driven by an inverter of finite strength, it can sink in only a finite amount of current. This current depends on the number of cells attached to the read word line and the data pattern stored in them (Fig 7.11).

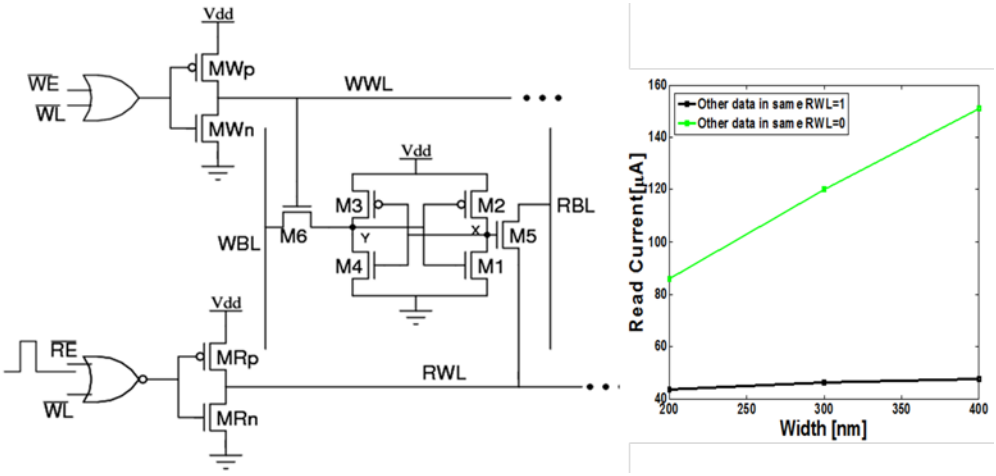



**Fig 7.11: Sizing the transistor for read: Considerations**

For a single such cell with 8 bits an 8X inverter driving the word line the current could be as high as 120μA (for the rest of the bits at "0") or could be as low as 46μA (for the rest of the bits at "1"). This observation is explained as the total current in the latter case is $N_{bit}$ x $I_{cell}$. Under this amount of current, however, there will be a substantial $V_{DS}$

drop across the NMOS device in the sink leading to a reduction of $V_{GS}$ across the read access transistor. This reduces the current consequently as the dependence of current on $V_{GS}$ is quadratic. Verma et al identified the problem and used a charge pump based overdrive of the NMOS to increase its drivability [88]. This is definitely a requirement for larger word size. However, for small word length, as in our case (8 bits) we do not need this additional circuit modification.

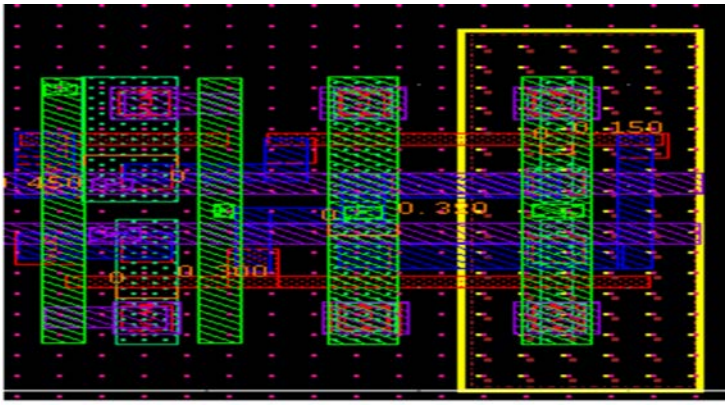


**Fig 7.12: Layout of the SRAM cell in IBM 130nm CMOSRF8SF process**

Also this sets a limit on the maximum width used for the read access device. Upsizing beyond that dimension does not help as then it is limited by the sinking capability of the driver. Also read is particularly heavily dependent on the access transistor. So to minimize impact of variation, it is a good idea to upsize it. However, upsizing it too much would add considerable capacitance to the read word line resulting in high dynamic power dissipation. The sizing for the cell is thus taken as 1(Pull Up): 1.875 (Read Access): 2 (Pull Down): 2.875 (Write Access). The pull up devices are taken to be minimum width devices for the technology (160nm).

We try to keep the cell dimensions as close to a type 4 cell layout in the CMOSRF8SF as possible. The dimension of the cell is 2.1μm X 3.1μm. The area overhead compared to a 6T cell is <1%.

### 7.9. The Issue of Single-Ended Sensing and Requirement of Gating

A single ended sensing scheme is required to detect the bit line drop. In the differential scheme, the voltage difference between the bit lines is detected. Any noise between the bit lines in the differential case is common mode. However, the noise in single ended sensing is not common mode. Hence, a larger swing across the bit line is required for detection of the same logic level. Normally, for differential purposes a sense amp differential of 50mV is used. For the single ended sense amp this swing is maintained at 100 mV.

The techniques used for single ended sensing are using dummy cells to mitigate the process impact. But that has the issue of adopting suitable memory architecture along with it. The other way to sense is via dynamic logic, however as we will see it is difficult to get a rail to rail swing and hence, this option is not lucrative as well. Thus a differential sense amplifier with an externally supplied reference voltage was chosen (1.1V for 1.2V level logic) (Fig 7.13(a)). The sense amplifier delay was estimated at 254 ps.

Also the sense amplifiers for this architecture have to be matched to the memory cell pitch (Fig 7.13(b)). Also, power dissipation in the sense amplifiers is high and occurs during every clock cycle irrespective of whether it is being used to read. An average current of 9 μA leading to 1.9μW/sense amp power output is produced. For a large

number of SAs, as in this case, this power can become large. To avoid it, the sense amp enable signal has been gated with the block select signal.

For the circuit, a current latch sense amplifier was used for its superior isolation and lower dynamic power. The sizing of its pull up, pull down, NMOS differential pairs and tail transistor are in the ratio of 1:1:2:3 respectively. The additional area overhead due to the extra sense amplifiers is estimated at 7.6%.

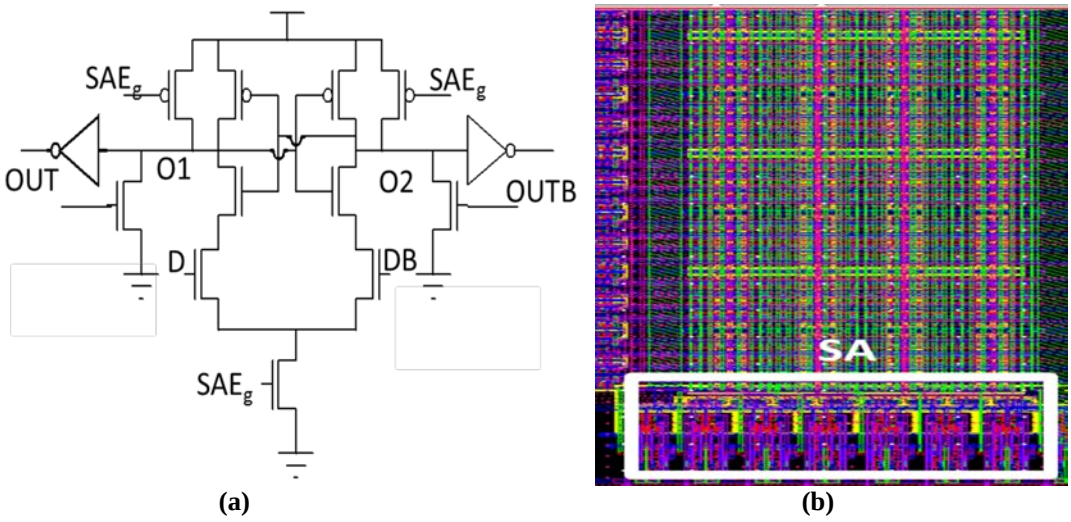



**Fig 7.13: (a) Sense amplifier structure (b) Sense amplifier layout**

## 7.10. Other Peripheral Designs

The decoder structure for this system consists of a 5:32 NAND decoder with pre-decoding for word line. The block level decoder was for 3:8 conversions. For the word line decoder, progressive sizing of the devices were done. Also the word line decoder had to fit into the array pitch. This required custom layout and routing of the standard cells to achieve that pitch. The decoder delay was observed at 600 ps.

## 7.11. Power Savings Using Mapping

The power savings in the memory arises from the fact that there is no discharge in the bit line for reading zero. Also depending upon the number of “0” stored in the bit line, a large number of cells might have a negative $V_{GS}$ condition suppressing leakage to a large extent. Hence, there is a huge potential in power savings. To evaluate it, we calculate the average power dissipation for reading a “1” and a “0”. Though a read “1” in this structure is slightly more energy expensive, the read “0” more than compensates for it. Overall, a 40% savings in power is suggested from simulation results. With scaling of the cell supply, the leakage power component contribution from unselected cells is minimized. This results in even larger power savings albeit at lower operating frequencies.

Another important factor to note is that the power savings are data dependent. A greater number of “0” can result in a larger power savings. Since the mapping is done once, there are ways to reconfigure the algorithms and software mapping to enable preferential mapping. This has been suggested by Paul et al to further improve on the energy savings front [45, 89].

### 7.12. Presence of Read Sneak Path: Power Loss and Solution

Functionally there are no problems with this circuit. However, the circuit suffers a major disadvantage when it comes to a pathological power dissipation scenario. This is illustrated in Fig 7.14. When the read access transistor storing a “1” conducts, the bit line discharges. The discharge goes on until the bit line reaches $V_{mem}$ -$V_{th}$. Beyond this level, the unselected cells have a sneak path to discharge leading to restricting the bit line voltage level to this value. A large sneak current now flows from the unselected cells via the chosen access transistor to the NMOS sink.

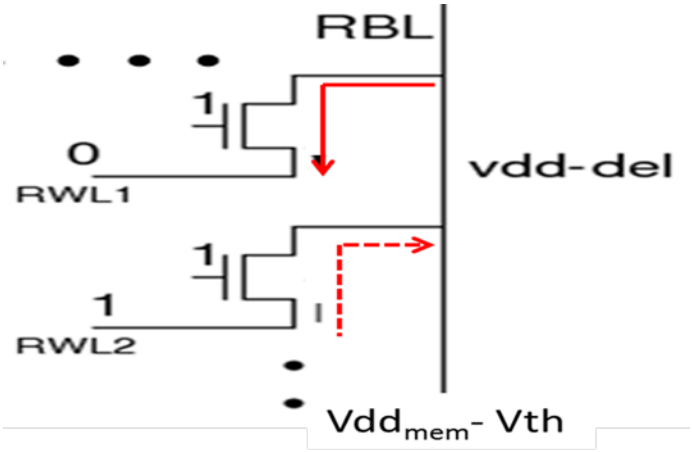
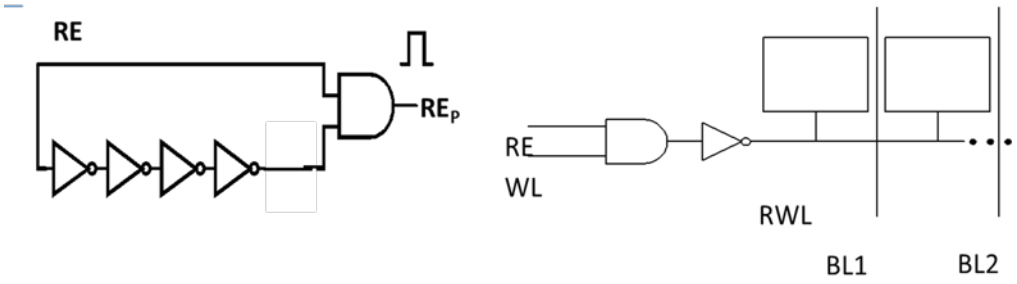



**Fig 7.14: Existence of the read sneak path**

To avoid this, a pulsed read solution is used. Again, the pulsed read solution comes with its own power and area overhead. Pulsing in SRAM has generally been associated with improving the read stability [90]. Here pulsing is used to avoid excess sneak current from flowing in the system. A very widely used strategy for pulse generation is to use delay chains. In this particular implementation, delay chains with externally controlled supply are used to control the pulse width. For the worst case (when a single cell is discharging at its full capacity) a 400 ps pulse width is required. The choice of the pulse width is motivated by the power crossover point of the alternate cell and the ordinary 6T cell. This is generated from the clock using the circuit shown in Fig 7.15(a). The pulse is incorporated into the circuit operation using the circuit in Fig 7.15(b). The circuit was successfully simulated to generate pulses of the given width using a 7 inverter chain and a delay chain voltage of 0.8 V. The implementation area overhead was <1% of the system. The power overhead was less than 4%.

A key problem for pulse generation schemes is process and temperature variation. To encounter the impact of process and temperature variations, the controlling parameter in our hand is the delay chain power supply. The structure was simulated across process

corners and temperature corners and the delay chain supply for each was noted. Hence, if the pulse is smaller than expected width resulting in functional failures or larger than the width resulting in excess power dissipation, we can fine tune and adjust it to our benefit.




**(a)** **(b)**

**Fig 7.15: (a) Pulse generation logic (b) Read pulse propagation to RWL**

## 7.13. Total System Performance and Power

The implemented MBC system was found to operate at 250 MHz at 1.2 V memory supply and 1.5V supply for the logic. The implementation uses a SRAM cell which is supposed to dissipate less power in a read-heavy application. The chip diagram for the system is shown in Fig 7.16.

The pulsed read generation circuit consists of the clock delayed with respect to itself. The width of the pulse is controlled by an externally supplied voltage, which controls the delay of the inverter chain. From the simulations we find, a delay chain element voltage supply of 0.8 V is required to produce a 500ps pulse. The pulse and the reading scheme is illustrated in Fig. 7.17. The schedule table output table is shown in Fig 7.18. As the schedule table output varies with the clock different operations which are programmed one after another are executed. These operations could be between the IR content, DR content or external inputs.

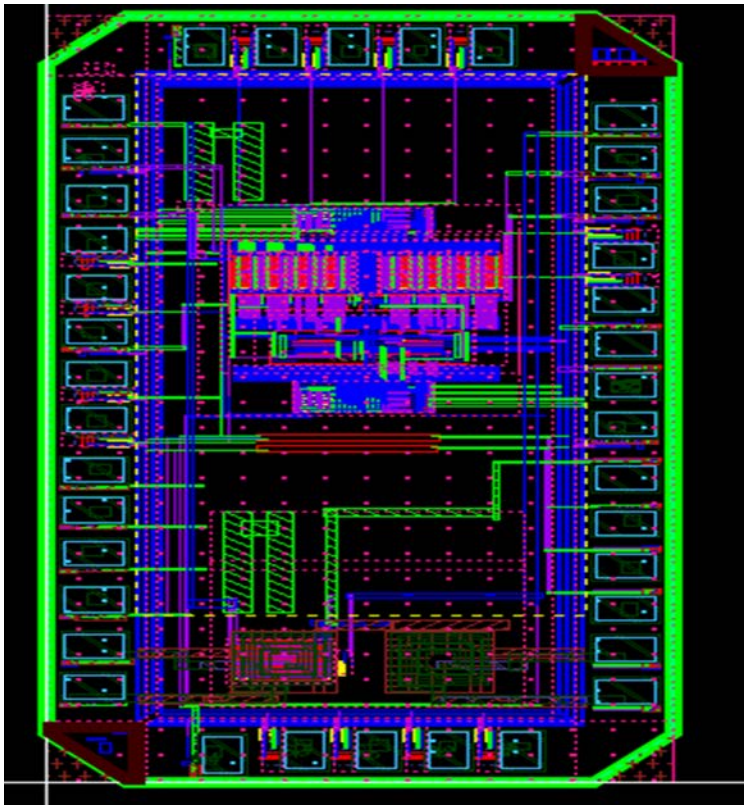


**Fig.7.16: Full chip diagram**

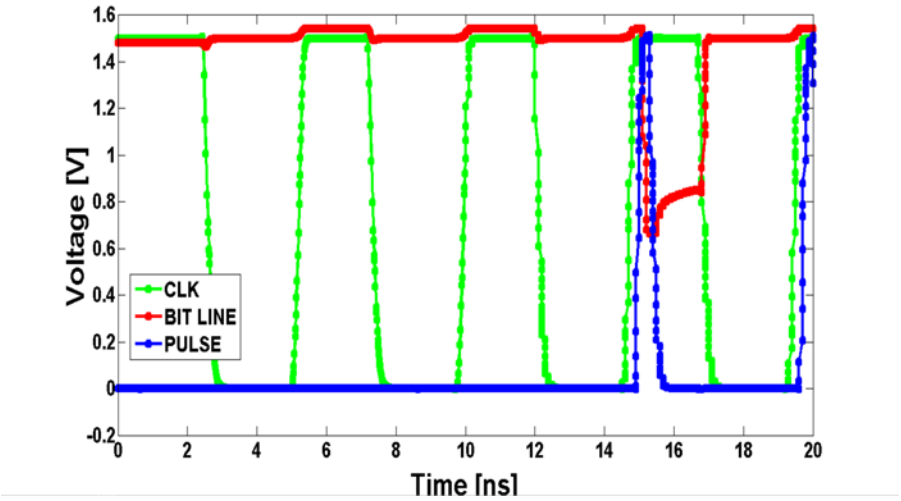



**Fig 7.17: Pulsed reading from memory**

The intermediate as well as the final values of the computation are stored in the

intermediate and dummy registers. These are 8 bit registers. For implementation simplicity, we restrict the number of dummy registers to one. Let us study an operation. For example, if we took a case of repetitive addition of two 4 bit numbers (say 1111 and 0001 in our case), the IR would store 00000001 starting from the point when the read operations begin. DR is initially expected to store 00010000 and then 00000001. We plot the significant bits of these registers to show the results of the discussed addition from our design (Fig 7.19).

A prototype memory based computing platform (consisting of a MLB computing element) has been designed at IBM 130nm CMRF8SF process. The implementation uses a pulsed read scheme to reduce the power dissipation across the memory. The complexity of the circuit has been minimized by restricting the number of dummy registers.

The design particulars are tabulated:

**Table 7.1: System Specifications**

| | |
|---|---|
| Technology | IBM 130nm |
| Transistor Count | ~35000 |
| Chip Area | 1mmX2mm |
| Nominal Operating Voltage | 1.2V (Memory) 1.5V (Logic) |
| Expected Performance | 250MHz |
| Key Blocks | Memory, Pulse Generation, Schedule Table, Intermediate Register, Dummy Register |
| Key Circuit Techniques | Voltage scalability, pulsed word-line read, Power variance under mapping modification. |

## 7.14. Pulsed Read

The necessity of pulsed read has been already discussed in an earlier section. Pulsed read is used to restrict excessive power dissipation due to possible read sneak paths. However, due to the sneak paths functionality is not compromised. The introduction of

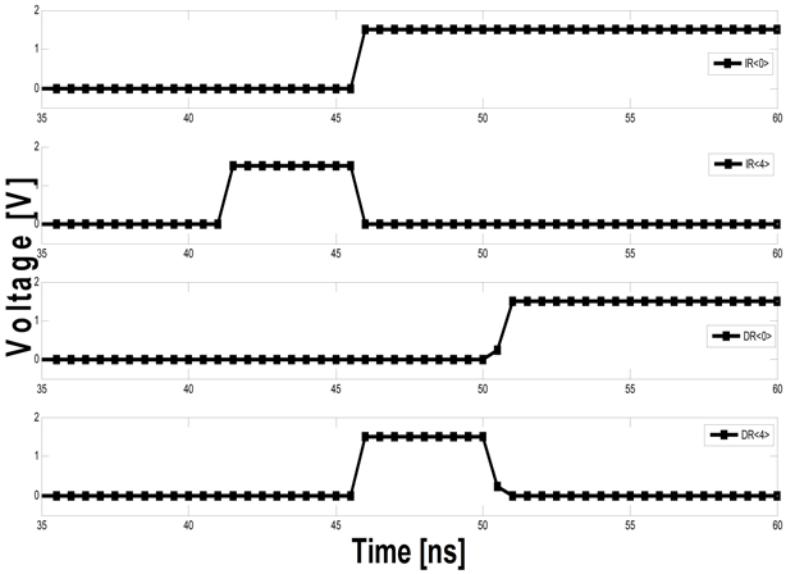



**Fig 7.19: Selected bits (4and 0) from Intermediate and Dummy Registers**

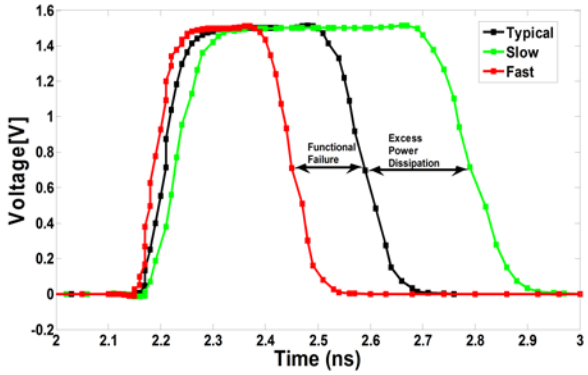
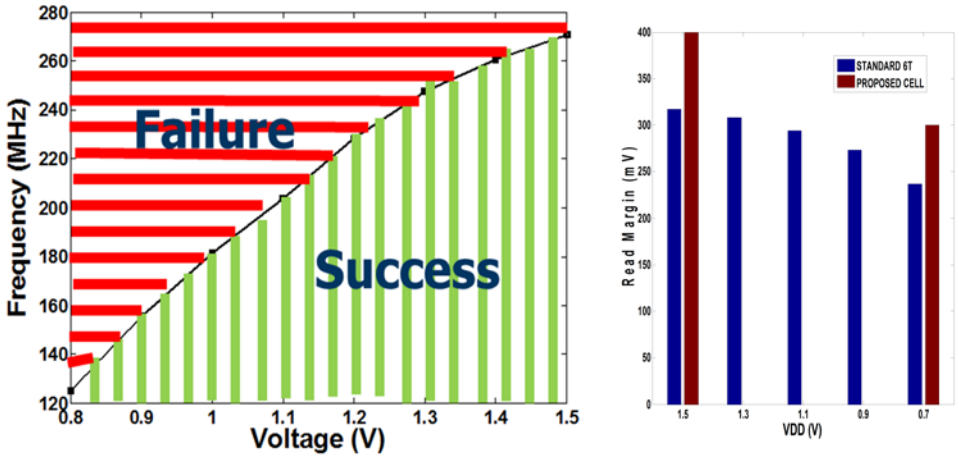



**Fig 7.20: Variation of pulse width across process corners**

pulsed read though brings along with it the chance of functional failure. In this section, we study why this occurs and discuss the remedies used to prevent failures.

The pulse width is determined from the power requirement. The pulse is expected to last for the voltage drop across the bit line to be sufficient for detection but at the same time not to let it reach a $V_{th}$ drop. However, with process imperfections the pulse width ratio is supposed to vary (Fig. 7.20). At the +3σ corner, this can lead to increased power dissipation. However, it is at the -3σ corner that the pulse width may be less than the required access times leading to false detection. To ensure proper functionality, we simulate the circuit at the slow corner to find the false detection probability.

To test the effectiveness of the pulsed read scheme, the data is written on to the memory. It is then read out without using the pulsed read scheme at a particular frequency to confirm functionality. Next, the same data is read out using pulsed read scheme for a particular pulse width (i.e. particular delay chain supply voltage). In case of process induced fluctuations, there could be failures leading to inconsistency of the two sets of data. The delay chain supply voltage is raised and the process repeated until the data are consistent.




(a) **(b)**

**Fig 7.21: (a) Voltage versus frequency plot for the system (b) Read margin with voltage**

## 7.15. Voltage Frequency Characteristics

To test the voltage frequency characteristics of the cell, data is written one time into the cell. The written data pattern is read out at different voltage and frequencies and noted for failure and success. The voltage frequency plane is divided into zones indicating successful operations and failures. This curve called the shmoo plot for voltage versus frequency. The shmoo plots for the alternate memory structure as well as the regular memory structure are shown in Fig 7.21(a). However, only looking at the shmoo plots for the voltage versus access time might give us an incomplete idea. The read disturb probability increases with scaled cell supply (Fig. 7.21(b)). The degrading read margin for standard 6T cells is a source of concern for reliable read operation. For the proposed cell, the SNM is higher and provides a better scalability opportunity.

## 7.16. Voltage Power Characteristics

Clearly the main objective of introducing the cell is power savings at scaled voltages. The cell saves on dynamic power while reading “0” as there is no discharge of the bit

line. Also at scaled voltages the memory leakage is reduced due to negative gate to source voltage of the cells storing “0”. To test the claim, the pattern written contains all “1”s is written and then read out. The current is noted for the read cycle. Similar steps are repeated for reading “0”. The write power is also noted for the above mentioned cases.

At a power supply of 1.5V (1.3 V for memory), a maximum operating frequency of

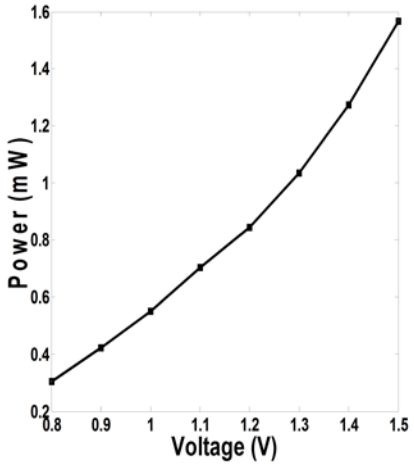
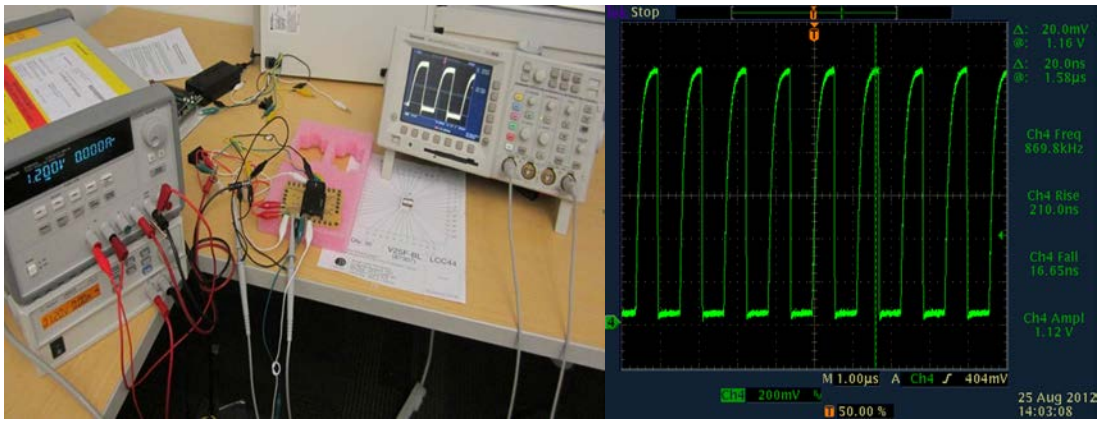



**Fig 7.22: Power versus voltage plot for the system**

250 MHz is obtained. The power consumed while read operation is found to be 1.6 mW 30% of the power arises from memory (for the worst case of reading all “1”) (Fig. 7.22). Scaling down the memory voltage to 1.0V and logic to 1.2V the power consumption reduces to 0.85mW where memory contributes 27% of the power. Clearly this suggests a great scope of saving power at scaled voltages for the proposed cell. This is because the standard 6T cells are not voltage scalable beyond a point from read failure perspective.

### 7.17. Hardware Data

The initial results for the fabricated chip are included in the chapter. Fig 7.23(a) shows the board level testing of the chip. The primary tests involved testing the internally generated clock used to drive the system. The clock signal is generated based on a 51




**(a)** **(b)**

**Fig 7.23: (a) Testing of the clock functionality of the chip (b) the clock at 222.4 MHz at 1.5V (the waveform shows a divided by 256 waveform obtained at the monitoring pin)**

stage ring oscillator chain. The observed waveform at the oscilloscope at a clock supply voltage of 1.5V is presented in Fig. 7.23(b). It corresponds to a frequency of 222.4 MHz at a voltage supply of 1.5V. This can be the maximum operational frequency of the system under test if the clock voltage is restricted to 1.5V. The observed voltage frequency characterization plot for the ring oscillator chain is shown in Fig. 7.24.

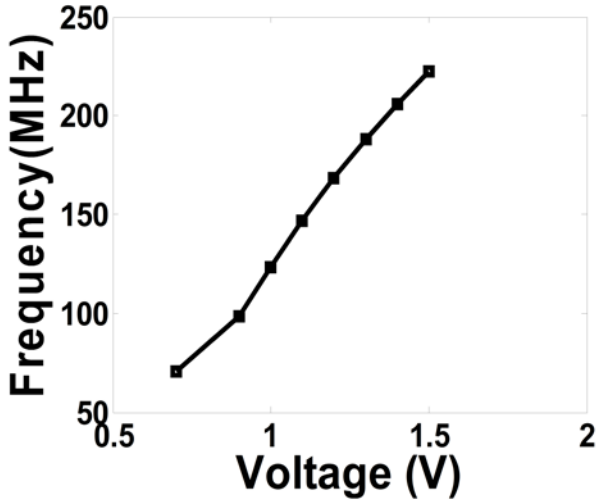



**Fig 7.24: Voltage-frequency characteristics of the clock**

An array of cells was included to illustrate the behavior of the SRAM cells under read and write conditions.

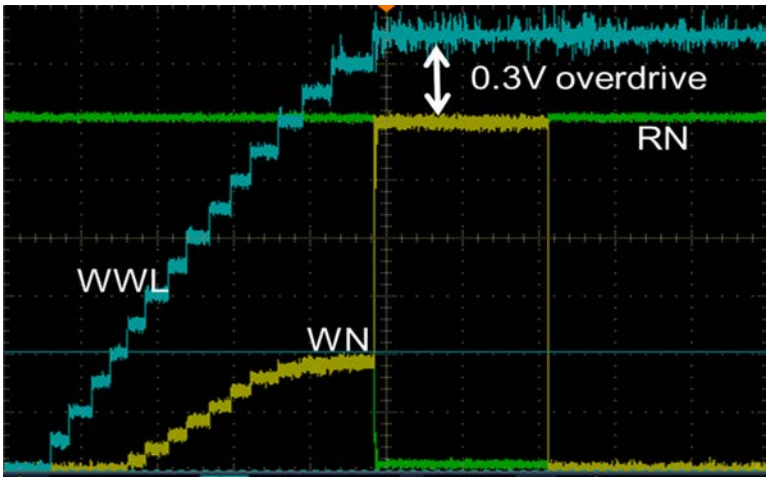
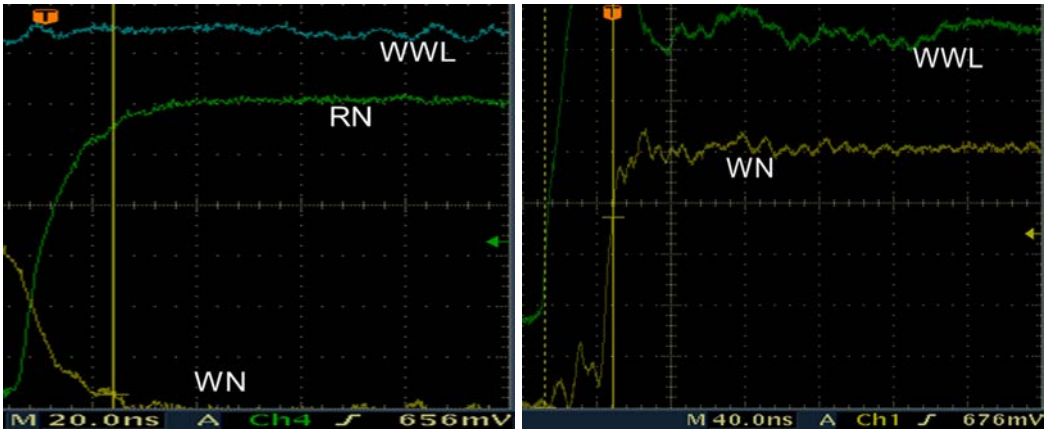



**Fig 7.25: Overdrive requirement for the alternate cell**

Writing is the difficult part of the cell functionality. Single ended writing necessitates word line over driving for the cell. This is substantiated with the cell write hardware data. Fig. 7.25 shows word line voltage and the storage node voltages with time. The cell voltage is at 1.2V. For writing, the word line voltage is steadily ramped up. Writing to the cell does not occur until an overdrive of 0.3V is forced on the cell. At that point flipping occurs. Also simulations have suggested that writing a “1” to a node storing “0” is the more difficult of the writing operations. This is because, under this condition, the strong pull down and the write access transistor are in contention. This disparity results in different write times for write “1” and write “0”. The difference is clearly illustrated in Fig. 7.26.

To verify reading operation, the total current drawn via the bit lines during read “1” condition was observed. The value of the current for different values of the cell supply voltage is plotted in Fig. 7.27. The current is found to be quite sensitive to the cell voltage




**(a)** **(b)**

**Fig 7.26: (a) Write "0" from "1" (b) Write "1" from "0". (a) is about 1.5X faster than (b)**

(37.5μA/V). This is expected as the cell voltage decides the gate overdrive of the read transistor.

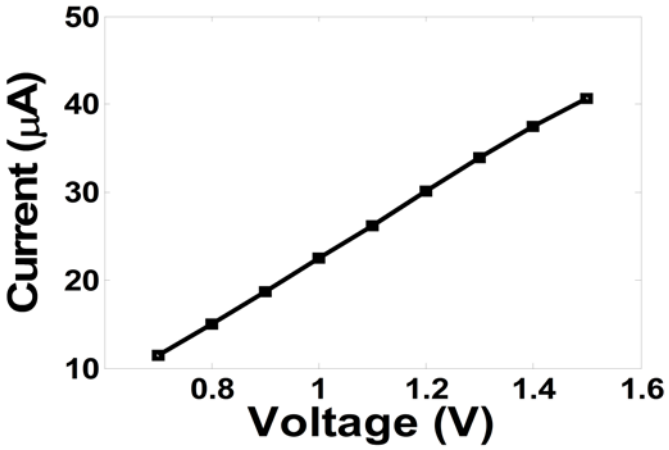



**Fig 7.27: Read Current versus cell voltage**

## 7.18. Conclusion

In this chapter, we have studied how knowledge about the application characteristics can be used for design maneuvering of SRAM. The modified SRAM topology assists us to achieve low power, reliable reading. The energy benefits can be further boosted by

preferential mapping of the look up table. However, to achieve this, issues of single ended writes and read sneak paths have to be taken care of. In the next chapter, we study how other factors like packaging solutions can modulate the SRAM specifications and how it translates to new design requirements.

# CHAPTER 8

# THERMAL COUPLING IMPACT ON 3D STACKED SRAM CACHE

3D integration allows partitioning different components of a system and stacking them in separate vertical layers connected by through-silicon-via (TSV). Folding a single 2D design into a 3D die-stack, hereafter referred to as die-folding, can reduce system footprint, interconnect delay, and power [91, 92]. Due to its potential advantage, system architecture as well as process technology of 3D integrated SRAM caches are receiving significant attention [60, 93]. When multiple power dissipating dies are stacked vertically with a reduced system footprint and hence, lower cooling efficiency, the system temperature can increase. The analysis and mitigation of thermal behavior of 3D die-stacks have been a key area of research [52, 57, 94].

Fig. 8.1 illustrates the system architecture of a 2D and 3D integrated core and cache system. As illustrated in Fig. 8.1, the heat escape path for 2D and 3D cases are similar, but the heat distribution paths are different in 3D. The heat generated in one tier can flow to the other tier through the die-to-die interface, resulting in the die-to-die thermal coupling [54, 56]. In a many-core processor the processor cores generates much higher and time-varying power compared to the SRAM sub-arrays (much lower power) in cache. Hence, it is important to analyze how die-folding and thermal coupling modulates the cache performance in 3D systems. Such analysis is critical to exploit advantage of 3D integration as performance, leakage power, and stability of SRAM are sensitive to temperature [61]. Further, temperature strongly modulates the time-dependent degradation of devices and SRAM cell stability due to bias temperature instability (BTI) [62].

We analyze the implications of 3D die-to-die thermal coupling on power, performance, and aging of SRAM. We show that in 3D system, the worst case temperature of the cache is tightly coupled to core power and temperature. Hence, we observe that a change in the power spread in cores (due to variations in applications and workload) strongly modulate the temperature variation of the SRAM blocks in 3D integrated cache. Consequently, 3D SRAM blocks have higher spatial and temporal variation. The higher average value and spread in temperature significantly degrade the access time (~30% increase maximum access time), increase array leakage (~2X), and accelerate the time-dependent device aging. Further, the spatial and temporal variations in performance of SRAM blocks become a strong function of the power variations in the cores.

### 8.1. 3D Stacking: The Thermal Issue

Thermal modeling for 3D system on chip has been an active area of research. The finite element methods [52] and the computationally less complex electrical mesh based methods [55] are normally used for analysis. Efforts have also been directed in finding the sensitivity of temperature distribution to various packaging parameters [95, 96]. It has been argued that folding a logic die into multiple stacks can increase the overall chip temperature due to higher power density [58]. This work concludes that thermal coupling is a major factor in 3D without exploring the cache performance impacts for core-cache

stacking. There are works emphasizing the electro-thermal implications of die stacking [53, 59].

However, there is a limited understanding of the electrical impact of the die-to-die thermal coupling between the core and SRAM tiers on the behavior of SRAM. The primary contribution of this paper is the analysis of the above effect considering random variations in the core power symbolizing workloads. To achieve this goal, we adopt a distributed RC based 3D thermal modeling approach suitable to evaluate thermal variations across a plane. Loi et. al [53] have evaluated the effect of thermal cross-talk in a 3D core-cache-memory stack, but they used a lumped model to represent the thermal behavior of each tier. Hence, the effect of within tier power and thermal variations were not captured.

Using the distributed analysis, we have shown that power variations among cores and strong thermal coupling lead to much higher spatial and temporal variations in temperature among SRAM sub-blocks in 3D than in 2D. Further, our analysis connects the effect of temperature to behavior of cache electrical parameters that includes not only performance and leakage but also aging and lifetime reliability. To the best of our knowledge, the interaction of 3D stacking and the aging behavior of SRAM have not been reported in earlier literature. This work concludes that thermal coupling is a major factor in 3D without exploring the cache performance impacts for core-cache stacking [97].

### 8.2. System Models and Simulation Environment

We consider a 64 core system as shown in Fig. 8.1. In 2D system, the cache is placed in between two sets of 32 cores (Fig 8.1(a)) to ensure the shared cache acts as a thermal

buffer between the cores. In the 3D die-folded system cores and cache are distributed in two different layers. We discuss the modeling paradigm in detail in the next subsection.

### 8.2.1. Thermal Modeling with 3D Distributed RC Grid

The thermal framework used in this study is constructed using distributed RC grid where R represents the thermal resistance and C represents the specific heat. The grid was structured for the 2D case to include the impact of heat sink, silicon and insulator layers (Fig. 8.2). Modeling of 3D stacked systems using the electrical equivalent circuit has been extensively used [56]. We use distributed RC models method to analyze the thermal coupling issue. In this work, we restrict ourselves to the steady state study using the R grid. We take a face-to-back bonded die with processor-memory stacking, as shown

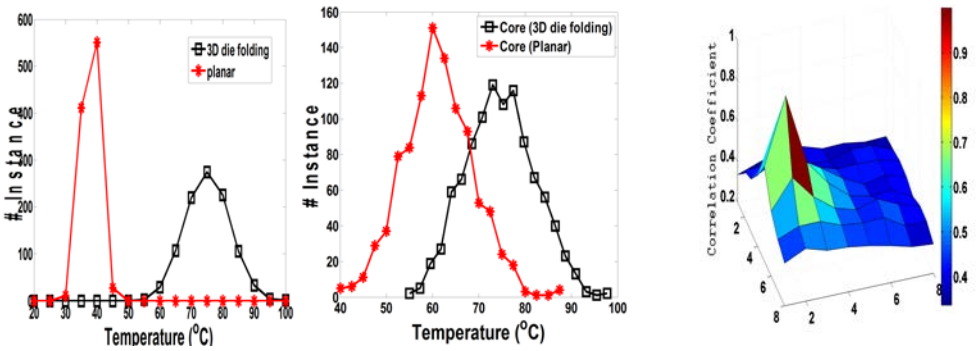
in Fig. 8.2(c). The layers consist of the thermal package i.e. heat sink, spreader, and thermal interface material, the bulk silicon, the active silicon (processor), the back end of the line including metal connections (for the processor layer), the die-to-die interface material, the bulk silicon, the active silicon (memory), the back end of the line (for the memory layer), and the electrical substrate/package (i.e. die-to-package interface). We simulated an 8mmX8mm chip having 64 cores. The width of the die considered was 350µm. The values of the bonding material, heat sink and package conductivity was set in accordance with the values reported in [55]. The thickness of the bonding layer was taken to be 10µm. The back end of the line resistance was determined, using oxide to metal ratio of 1:3. The thermal resistivity of the die-to-die interface layer (core and SRAM) was modified to consider the effect of the heat flow through the TSV. HSPICE was used as the RC circuit simulator. Power consumption wise the cache was assumed to consume 10% of the total power. This choice is mainly motivated by the L2 cache power consumption percentage of current microprocessor [51].

### 8.3.Analysis Methodology - Coupling of Thermal and Circuit Simulations

We first consider that the power of cores (and SRAM) in the 2D and 3D systems are same. To emulate the effect of workload variation on core power, we assume core power follows a Gaussian distribution (mean of 1.5W and a variance of 0.5W) (Fig. 8.2(a)).

We estimate the temperature patterns in cores and caches in 2D and 3D system considering a large number of random power patterns applied to the cores. Each power pattern is applied for constant period of time in simulation (until steady-state is reached). Each such power pattern applied to the cores results in a particular temperature pattern across the cache blocks for 2D and 3D systems. For each power pattern, we estimate the

temperature for each SRAM block and maximum and minimum temperature for the entire cache (i.e. maximum and minimum over all blocks at that power pattern). We first study the temperature behavior of SRAM blocks and maximum/minimum temperature of cache over time (i.e. over power patterns). This study indicates how 2D and 3D integration modulates the temperature behavior of SRAM cache over time (*temporal*). We next study how the *spatial* pattern in temperature in the cache is different in 2D and 3D design. The temporal study is critical to understand what ranges of temperature SRAM blocks will experience in 2D and 3D design. The spatial study also helps analyze the how 2D and 3D integration modulates the location of thermal hotspots in 2D and 3D.




**(a)** **(b)** **(c)**

**Fig 8.4: Temperature distribution across (a) cache and (b) core. 3D integration results in higher mean and spread in temperature distribution of cores and cache blocks considering all random power patterns. (c)Thermal correlation for a cache block (location: 2, 3) with core blocks directly below it. Maximum correlation is observed with the core block directly below and it falls off rapidly with increasing distance. Hence, it shows cache temperature is more strongly correlated to cores vertically below itself.**

Finally, we connect the temperature estimated from thermal simulations to the power, performance, and reliability (stability under PMOS aging) estimated from circuit simulations. The thermal simulations for different random power patterns provide temperature of SRAM blocks. We assume all cells within a block have same temperature.

The spatial and temporal patterns of temperature estimated from thermal simulation are used to compute the corresponding patterns in performance, leakage, and reliability.

## 8.4.Simulation Results and Analyses

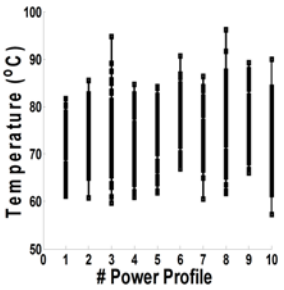
### 8.4.1. Comparison of Thermal Distribution (with and without Stacking)

Fig. 8.3 shows the spatial variation in temperature for cache blocks considering 2D and 3D designs. Fig. 4 shows the possible temperature variation for a given core and SRAM cache block in the 2D and 3D case considering different power pattern (i.e. at different time point). In Fig. 8.5(a), given a power pattern (in x-axis), the points in the y-axis represent the temperature of different cache blocks in the 3D design. We obtained similar plot for 2D as well (not shown). Fig. 8.5(b) summarizes the spatial observation and plots the histogram (over all power patterns) of identification number of the cache block that becomes the hottest for a power pattern. We summarize the observation as follows:

- ***The core and cache temperature in 3D is higher than 2D***. We first observe that the cores and cache experience a much lower temperature in 2D case due to the larger spatial area (i.e. lower average power density and larger heat spreader/heat sink contact area) (Fig. 8.3, 8.4). In the 3D scenario, the temperature of the cores is higher due to the reduced footprint (higher average power density and smaller heat sink/heat spreader contact area). The elevated core temperature is coupled through the die-to-die interface to the caches on next tier resulting in an increased temperature for the SRAM blocks (~ 30-40$^0$C higher than 2D SRAM blocks) (Fig. 8.4).
- ***The temporal variation in temperature of cache blocks is much higher for the 3D system compared to the 2D system.*** Fig. 8.4(a, b) shows that for the 2D scenario, at any

particular location of the cache block, the variance in temperature over time is small. For the 3D system, the individual cache blocks may experience significantly higher average temperature as well as larger temperature spread over time (Fig. 8.4(a), (b)).

- ***The spatial thermal pattern in 3D is more non-deterministic and more strongly correlated to power patterns of cores compared to 2D.*** Fig. 8.3(a)-(b) shows that for 2D case, although the power patterns across cores are very different, the cache blocks towards the central region remain the coldest and those towards the periphery are the hottest. For the 3D scenario we observe that the temperature pattern of the cache follow the temperature pattern across the core distribution very closely. This suggests that as the power pattern across the cores can vary significantly depending on the workload pattern so does the thermal pattern across the cache. This is observed in Fig. 8.5(a) which shows that in 3D one can have significant spatial variation in temperature across cache blocks. This is further evident from Fig. 8.4(c) which shows correlation between




**(a)** **(b)**

**Fig 8.5: (a) Distribution of temperature with power profile in 3D and (b) Histogram of the identification number of the hottest cache block. In 3D the spatial difference in temperature is much higher. Moreover, all the blocks have equal probability of being the hottest block in 3D while in 2D location of hottest blocks is deterministic.**

the temperature of a cache block and all the cores. We observe that the cache temperature is very closely correlated to the temperature of the core directly below it and falls off very sharply (within the order of 1 core distance) to a relatively low value. When the cores have spatiotemporally non-uniform and random power profile (as is expected in most applications), temperature can be considered as a source of spatiotemporal random variation for the cache blocks in 3D system. Therefore, for the 2D case, even under statistical variation in the power profile of cores, both the temperature of cache blocks and locations of hottest cache blocks are more predictable. But in 3D each block has much higher temperature spread (over time) and all blocks are equally likely to be hottest block. This is illustrated in Fig. 8.5(b) which shows the location of the hottest cache block for 2D and 3D case.

## 8.5. Implications for Power, Performance and Reliability of SRAM

In this section we evaluate effect of changing temperature distribution on the read access time, leakage power, and reliability of SRAM blocks and cache considering process variation using 32nm predictive technology [98].

- Leakage: With increase in temperature the nominal value of the leakage increases exponentially. Moreover, due to the exponential dependence, the rate of change of leakage due to threshold voltage variation depends exponentially on temperature. For any instance of power distribution across the cores (i.e. at a given time point), the spatial spread in temperature is much higher in 3D [Fig. 8.6(a)]. This translates to large spatial spread in leakage of different sub-arrays for 3D integrated SRAM than the 2D counterpart. For a given power pattern, we next add the estimated leakage of all blocks to compute the *cache leakage* for that power pattern. The statistical distribution

of the total cache leakage for all power patterns (i.e. over time) is shown in Fig. 8.6(b). Due to the higher average temperature and spread, both the mean leakage power (~2X) and the spread in the cache leakage over time (~4X) increase significantly as we move to 3D.

- Read Access Time of an SRAM cell is defined as the time required to develop a pre-defined bit-differential (~100mV) during read operation. During reading, the bit line BL discharges due to the read current ($I_{read}$) flowing through the selected cell while bit

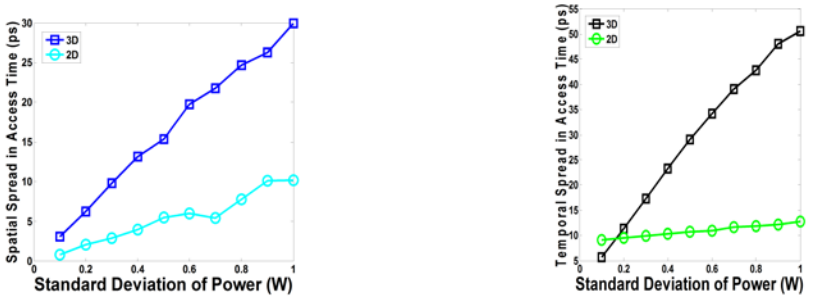
line complement BR discharges due to bit-line leakage ($I_{bitline\_leakage}$) current. At higher temperature the cell read current reduces due to lower on current for access and pull-down NMOS devices of the selected SRAM cell. Moreover, higher temperature also increases the bit line leakage current through the other access transistors. The overall performance of the cache (*cache access time*) for a power pattern (i.e. at a given time point) is determined by the maximum of the block access time over all SRAM blocks. Fig. 8.6(c) shows the distribution of *block access time* of SRAM sub-arrays for a given power pattern i.e. the spatial variation in access time across the entire cache. We observe that the access times of the cells in 2D are much predictable (tighter distribution) while in a 3D scenario, they are much widespread. The statistical distribution of the *cache access time* (i.e. maximum of the block access time for a power pattern) for all power patterns (i.e. over time) is shown in Fig. 8.6(d). We observe that the average *cache access time* increases by nearly 50ps leading to ~28% performance degradation.

- Device Degradation due to NBTI caused with prolonged usage is known to have significant dependence on the voltage and thermal stress. It is known that increase in the Vth of the PMOS devices due to NBTI degrades the cell stability – i.e. reduces read margin and increases minimum data retention voltage (DRV). Using the NBTI models presented in [62], we estimate the change in PMOS Vth for each block considering its *average operating temperature.* The estimated PMOS Vth shift is for an SRAM block used to compute the 3σ worst-case values of read margin and $V_{min}$ (both under process variation) of that block. As the cache temperature is much higher for a 3D SRAM cache block, the rate of degradation of threshold voltage for PMOS

will also be faster in comparison to the 2D case (Fig. 8.7(a)). But for 2D the blocks near the cores degrade at a faster rate. The faster degradation of PMOS threshold voltage also results in a faster reduction of read margin and increase of minimum DRV for 3D caches (Fig. 8.7(b) and 8.7(c)).

## 8.6. Effect of Non-uniformity of the Power Pattern

We next study the correlation between spatiotemporal variation in the performance of SRAM and the non-uniformity in the power profile of cores. We maintain a constant mean (1.5W) but repeat the Monte-Carlo simulations for different standard deviation (0.1W to 1W). This study is performed to understand the impact of running different




**(a)** **(b)**

**Fig 8.8: Effect of varying spread in the power variation of cores on the performance of SRAM cache in 2D processor and 3D die stack. (a) Spatial Spread (b) Temporal Spread. Increase in the standard deviation of core power variation results in an increase in spatial and temporal spread in access time**

workload with varying power variability among the cores (as different threads running on different cores will have varying computational power). For each standard deviation of power profile, we compute the spatial and temporal spread in SRAM properties. For brevity, here we present the analysis of the access time for 2D and 3D SRAM for different non-uniformity in the power pattern – the leakage and device aging follows the same trend. For a given standard deviation of core power, we consider the different instances of power pattern and estimate the spatial variation in block access time for each pattern. We next compute the statistical variation in the *cache access time* considering those 1000 Monte-Carlo instances. The computed spread of the cache access time distribution for a given standard deviation of core power variation is defined as the 'temporal spread' in access time for that standard deviation. The results are shown in Fig 8.8. We clearly observe that an increase in the standard deviation of core power variation results in an increase in spatial and temporal spread in access time. However, the increase in much higher in 3D integrated SRAM. Therefore, the workload that introduces a

significant non-uniformity in power patterns (higher standard deviation) across cores increases the access time spread compared to workload creating uniform power profile.

### 8.7.Conclusion

3D die folding leads to significant deviations of the cache thermal profile in a core-cache die-stack. Temperature acts an additional source of variation leading to significant degradation for leakage and performance and accelerated aging. 3D integration reduces interconnect latency but increases the access delay and leakage power of the cache. Our analysis points to the fact that the power pattern and workloads are important factors deciding the extent of thermal coupling enabled performance degradation. This suggests a strong need for a cohesive analysis considering the trade-off between the gains in interconnect delay and temperature induced degradation in SRAM power and performance.

# CHAPTER 9

# CONCLUSION AND FUTURE WORK

The dissertation showed how introduction of application-aware design methodologies and design techniques could result in power-performance benefits in embedded memory domain. The design methodologies and design techniques span across the domain of volatile and non-volatile memory technology. In this chapter, we summarize the contributions made in this dissertation, evaluate their impact on the embedded memory application and suggest areas where there is possibility of extension of this work.

## 9.1. Summary of Contributions

The first half of the dissertation deals with STTRAM as a prototype for emerging non-volatile memory. Being a maturing technology, there are areas in design methodology development which need to be addressed. The dissertation points out the area of energy-efficiency as one of the areas where there is a requirement of significant design methodology development. An energy model for the STTRAM array structure is developed. The developed energy model is coupled with cell-specification-constrained design space exploration to arrive at the optimal energy design point. The sensitivity of the optimal design point with application characteristics are studied extensively. The design point is found to be particularly sensitive to the percentage of read and write across applications.

Adoption of the STTRAM technology requires a detailed study of the reliability aspect of it. STTRAM is particularly known for its high programming current which gives rise to a high temperature. The evaluation of the thermal distribution in STTRAM considering the MTJ and transistor behavior together was performed. The results obtained thereby were used to find the impact on circuit and system level properties. Sensing results were particularly sensitive to the thermal distribution. Also, as different vectors and operations require different current to flow across the RAM, pattern history dependence was suggested. The results corresponding to the pattern history dependence were presented. The impact of the source biasing circuit technique was assessed to this end to reduce the leakage in unselected cells of selected column. The choice of the voltage level is a trade-off between the sensing accuracy and dynamic power overhead.

STTRAM design methodology was applied across cache and MBC to find design optimizations possible. While the read-write ratio may vary widely for cache applications, for look up based applications like MBC reads are high majority operations. Hence, the cell design could be optimized to get additional energy benefits as proposed in the methodology. Beyond that, a further reduction was suggested to be possible by preferential mapping of “0” versus “1”.

In the volatile domain, SRAM based modifications were suggested to improve the EDP of MBC application. The read dominant nature was exploited to introduce an alternate topology with pulsed, single ended detection mechanism. The use of the cell has the potential to reduce power and improve scalability but is bounded by its problems with single-ended write and a pathological read scenario. We suggested techniques to overcome these limitations- by adopting word line overdrive and pulsed read approaches.

Lastly, 3D offers a unique packaging platform for die stacking and die folding to keep up with Moore's law in future. One of the possibilities 3D opens up is that of stacking SRAM on top of core, which would lead to operations under conditions of temperature variability. The dissertation evaluates the impact of SRAM integration in 3D and uses these results towards a detailed failure mechanism study of SRAM. Design constraints in 3D are found to be tighter than that imposed by the usual process corner based analysis.

## 9.2. Future Work

The studies for STTRAM have been conducted based on the available in-plane MTJ data. Perpendicular-plane MTJs are slowly becoming popular with efforts being made in the direction of CMOS integration. Hence, the thermal modeling work can be extended to such perpendicular-plane MTJs to register the difference in the thermal impact. The interplay between temperature for these cells and their properties is an interesting and new topic for future work. Also, the pattern history dependence suggested in our studies has been estimated at the worst case corners. Our focus has been to understand temperature rise in STTRAM and its coupled impact on the STTRAM parameters. Separately we have also analyzed how reads and writes will affect the temperature distribution. Beyond that we have analyzed the worst case scenarios that could affect STTRAM. However, the worst case is an unlikely scenario. Reads to writes will follow in a certain ratio (e.g. depending on the hardware and application there will be a fixed number of read operations per write operations). This opens up a whole new dimension of analysis ( E.g.: dependence of the failure probability depending on data access patterns for different sizes of cache and benchmarks etc.). A more optimistic measure of this dependence may be found by coupling the device level findings with

detailed architecture benchmark simulation characteristics. This might be an interesting and necessary topic for enablement of STTRAM technology in future.

3D stacking of STTRAM has been suggested. However, the studies have been restricted to a circuit-microarchitecture level understanding of the benefits of stacking. However, the dynamics between temperature, performance and failure mechanisms need to be considered to get a realistic estimate of the 3d stacking performance for STTRAM. This is an important area that needs exploration from NVRAM and VLSI perspectives. Also, as 3D applications are particularly power sensitive (leads to high temperature), the optimal design point must be considered in any such design to improve power efficiency.

Also, the dissertation shows that volatile memories can be altered topologically and architecturally to better suit the needs of a particular application. This realization can be carried forward to other forms like eDRAM especially in the domain of reconfigurable embedded applications.

### 9.3. Conclusion

The dissertation deals with the analysis and implementation of energy-efficient design methodologies and design styles which translate to power-performance benefits in embedded application domain. We proceed with an emerging non-volatile memory technology (STTRAM) and a volatile memory technology (SRAM), and show that the understanding of the application can be incorporated into the design of the memory hierarchy for power-performance leverage. The learning from this dissertation is applicable across embedded memory applications, which are becoming increasingly vast. To keep up with Moore’s law, innovative embedded and packaging solutions are being

exploited. This requires logic and memory exploration in detail. The memory design exploration requires detailed design space exploration framework and energy and reliability modeling. The hierarchical approach followed in this thesis can be extended towards any generic memory technology. A co-design which takes into account the memory technology details and the application characteristics can be particularly helpful in reducing the power and increasing the performance in such future systems. The application-aware techniques suggested for emerging and existing technologies can be exploited to design complex computing systems of the future.

# REFERENCES


[1] A. Driskill-Smith, S. Watts, D. Apalkov, D. Druist, X. Tang, Z. Diao, X. Luo, A. Ong, V. Nikitin, and E. Chen, "Non-volatile Spin-Transfer Torque RAM (STT-RAM): An analysis of chip data, thermal stability and scalability," in *Memory Workshop (IMW), 2010 IEEE International*, 2010, pp. 1-3.

[2] M.Hosomi, H.Yamagishi, and T.Yamamoto, "A Novel Nonvolatile Memory with Spin Torque Transfer Magnetization Switching: Spin-RAM," presented at the IEEE International Electron Devices Meeting (IEDM), 2005.

[3] H. Sunami. (2008, The Role of the Trench Capacitor in DRAM Innovation. *IEEE Solid State Circuits*. Available: http://www.ieee.org/portal/site/sscs/menuitem.f07ee9e3b2a01d06bb9305765bac26c8/index.jsp?&pName=sscs_level1_article&TheCat=6010&path=sscs/08Winter&file=Sunami.xml

[4] S. Chatterjee, M. Rasquinha, S. Yalamanchili, and S. Mukhopadhyay, "A methodology for robust, energy efficient design of Spin-Torque-Transfer RAM arrays at scaled technologies," in *Computer-Aided Design - Digest of Technical Papers, 2009. ICCAD 2009. IEEE/ACM International Conference on*, 2009, pp. 474-477.

[5] T. Shimizu, J. Korematu, M. Satou, H. Kondo, S. Iwata, K. Sawai, N. Okumura, K. Ishimi, Y. Nakamoto, M. Kumanoya, K. Dosaka, A. Yamazaki, Y. Ajioka, H. Tsubota, Y. Nunomura, T. Urabe, J. Hinata, and K. Saitoh, "A multimedia 32 b RISC microprocessor with 16 Mb DRAM," in *Solid-State Circuits Conference, 1996. Digest of Technical Papers. 42nd ISSCC., 1996 IEEE International*, 1996, pp. 216-217, 448.

[6] G. Wang, K. Cheng, H. Ho, J. Faltermeier, W. Kong, H. Kim, J. Cai, C. Tanner, K. McStay, K. Balasubramanyam, C. Pei, L. Ninomiya, X. Li, K. Winstel, D. Dobuzinsky, M. Naeem, R. Zhang, R. Deschner, M. J. Brodsky, S. Allen, J. Yates, Y. Feng, P. Marchetti, C. Norris, D. Casarotto, J. Benedict, A. Kniffm, D. Parise, B. Khan, J. Barth, P. Parries, T. Kirihata, J. Norum, and S. S. Iyer, "A 0.127 $\mu m^2$ High Performance 65nm SOI Based embedded DRAM for on-Processor Applications," in *Electron Devices Meeting, 2006. IEDM '06. International*, 2006, pp. 1-4.

[7] S. Paul and S. Bhunia, "MBARC: a scalable memory based reconfigurable computing framework for nanoscale devices," presented at the Proceedings of the 2008 Asia and South Pacific Design Automation Conference, Seoul, Korea, 2008.

[8] S. Paul and S. Bhunia, "Memory based computing: reshaping the fine-grained logic in a reconfigurable framework (abstract only)," presented at the Proceedings of the 19th ACM/SIGDA international symposium on Field programmable gate arrays, Monterey, CA, USA, 2011.

[9] J. C. Slonczewski, "Currents, torques, and polarization factors in magnetic tunnel junctions," *Physical Review B*, vol. 71, p. 024411, 2005.

[10] Y. H. Driskill-Smith Alexander, "STT-RAM – A New Spin on Universal Memory," *Future Fab Intl.*, 2007.

[11] M. H. Kryder and K. Chang Soo, "After Hard Drives;What Comes Next?," *Magnetics, IEEE Transactions on*, vol. 45, pp. 3406-3413, 2009.

[12] Y. Huai, D. Apalkov, Z. Diao, Y. Ding, A. Panchula, M. Pakala, L.-C. Wang, and E. Chen, "Structure, Materials and Shape Optimization of Magnetic Tunnel Junction Devices: Spin-Transfer Switching Current Reduction for Future Magnetoresistive Random Access Memory Application," *Japanese Journal of Applied Physics*, vol. 45, p. 6, 2006.

[13] Z. Diao, D. Apalkov, M. Pakala, Y. Ding, A. Panchula, and Y. Huai, "Spin transfer switching and spin polarization in magnetic tunnel junctions with MgO and AlOx barriers," *Applied Physics Letters*, vol. 87, pp. 232502-232502-3, 2005.

[14] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano, "A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram," in *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, 2005, pp. 459-462.

[15] C. Suock, K. M. Rho, S. D. Kim, H. J. Suh, D. J. Kim, H. J. Kim, S. H. Lee, J. H. Park, H. M. Hwang, S. M. Hwang, J. Y. Lee, Y. B. An, J. U. Yi, Y. H. Seo, D. H. Jung, M. S. Lee, S. H. Cho, J. N. Kim, G. J. Park, J. Gyuan, A. Driskill-Smith, V. Nikitin, A. Ong, X. Tang, K. Yongki, J. S. Rho, S. K. Park, S. W. Chung, J. G. Jeong, and S. J. Hong, "Fully integrated 54nm STT-RAM with the smallest bit cell dimension for high density memory application," in *Electron Devices Meeting (IEDM), 2010 IEEE International*, 2010, pp. 12.7.1-12.7.4.

[16] A. Driskill-Smith, S. Watts, V. Nikitin, D. Apalkov, D. Druist, R. Kawakami, X. Tang, X. Luo, A. Ong, and E. Chen, "Non-volatile spin-transfer torque RAM (STT-RAM): Data, analysis and design requirements for thermal stability," in *VLSI Technology (VLSIT), 2010 Symposium on*, 2010, pp. 51-52.

[17] A. Raychowdhury, "Model study of 1T-1STT MTJ memory arrays for embedded applications," in *Circuits and Systems (MWSCAS), 2010 53rd IEEE International Midwest Symposium on*, 2010, pp. 5-8.

[18] L. Hai and C. Yiran, "An overview of non-volatile memory technology and the implication for tools and architectures," in *Design, Automation & Test in Europe Conference & Exhibition, 2009. DATE '09.*, 2009, pp. 731-736.

[19] W. Xiaoxia, L. Jian, Z. Lixin, E. Speight, and X. Yuan, "Power and performance of read-write aware Hybrid Caches with non-volatile memories," in *Design, Automation & Test in Europe Conference & Exhibition, 2009. DATE '09.*, 2009, pp. 737-742.

[20] C. W. Smullen, V. Mohan, A. Nigam, S. Gurumurthi, and M. R. Stan, "Relaxing non-volatility for fast and energy-efficient STT-RAM caches," in *High Performance Computer Architecture (HPCA), 2011 IEEE 17th International Symposium on*, 2011, pp. 50-61.

[21] A. Nigam, C. W. Smullen, V. Mohan, E. Chen, S. Gurumurthi, and M. R. Stan, "Delivering on the promise of universal memory for spin-transfer torque RAM (STT-RAM)," in *Low Power Electronics and Design (ISLPED) 2011 International Symposium on*, 2011, pp. 121-126.

[22] C. W. Smullen, A. Nigam, S. Gurumurthi, and M. R. Stan, "The STeTSiMS STT-RAM simulation and modeling system," in *Computer-Aided Design (ICCAD), 2011 IEEE/ACM International Conference on*, 2011, pp. 318-325.

[23] A. Nigam, K. Munira, A. Ghosh, S. Wolf, E. Chen, and M. R. Stan, "Self consistent parameterized physical MTJ compact model for STT-RAM," in *Semiconductor Conference (CAS), 2010 International*, 2010, pp. 423-426.

[24] W. Zhao, E. Belhaire, Q. Mistral, C. Chapped, V. Javerliac, B. Dieny, and E. Nicolle, "Macro-model of Spin-Transfer Torque based Magnetic Tunnel Junction device for hybrid Magnetic-CMOS design," in *Behavioral Modeling and Simulation Workshop, Proceedings of the 2006 IEEE International*, 2006, pp. 40-43.

[25] J. D. Harms, F. Ebrahimi, Y. Xiaofeng, and W. Jian-Ping, "SPICE Macromodel of Spin-Torque-Transfer-Operated Magnetic Tunnel Junctions," *Electron Devices, IEEE Transactions on*, vol. 57, pp. 1425-1430, 2010.

[26] Y. Chen, X. Wang, H. Li, H. Liu, and D. V. Dimitrov, "Design Margin Exploration of Spin-Torque Transfer RAM (SPRAM)," in *Quality Electronic Design, 2008. ISQED 2008. 9th International Symposium on*, 2008, pp. 684-690.

[27] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. De, "Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, 2009, pp. 1-4.

[28] L. Jing, C. Augustine, S. Salahuddin, and K. Roy, "Modeling of failure probability and statistical design of Spin-Torque Transfer Magnetic Random Access Memory (STT MRAM) array for yield enhancement," in *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, 2008, pp. 278-283.

[29] L. Jing, L. Haixin, S. Salahuddin, and K. Roy, "Variation-tolerant Spin-Torque Transfer (STT) MRAM array for yield enhancement," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, 2008, pp. 193-196.

[30] D. Apalkov, D. Zhitao, A. Panchula, W. Shengyuan, H. Yiming, and K. Kawabata, "Temperature Dependence of Spin Transfer Switching in Nanosecond Regime," *Magnetics, IEEE Transactions on*, vol. 42, pp. 2685-2687, 2006.

[31] S. Salahuddin and S. Datta, "Self-consistent simulation of quantum transport and magnetization dynamics in spin-torque based devices," *Applied Physics Letters*, vol. 89, pp. 153504-153504-3, 2006.

[32] S. Salahuddin and S. Datta, "Electrical detection of spin excitations," *Physical Review B*, vol. 73, p. 081301, 2006.

[33] B. Murmann, P. Nikaeen, D. J. Connelly, and R. W. Dutton, "Impact of Scaling on Analog Performance and Associated Modeling Needs," *Electron Devices, IEEE Transactions on*, vol. 53, pp. 2160-2167, 2006.

[34] R. Scheuerlein, W. Gallagher, S. Parkin, A. Lee, S. Ray, R. Robertazzi, and W. Reohr, "A 10 ns read and write non-volatile memory array using a magnetic tunnel junction and FET switch in each cell," in *Solid-State Circuits Conference, 2000. Digest of Technical Papers. ISSCC. 2000 IEEE International*, 2000, pp. 128-129.

[35] Y. Iwata, K. Tsuchida, T. Inaba, Y. Shimizu, R. Takizawa, Y. Ueda, T. Sugibayashi, Y. Asao, T. Kajiyama, K. Hosotani, S. Ikegawa, T. Kai, M.

Nakayama, S. Tahara, and H. Yoda, "A 16Mb MRAM with FORK Wiring Scheme and Burst Modes," in *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*, 2006, pp. 477-486.

[36] T. Kawahara, R. Takemura, K. Miura, J. Hayakawa, S. Ikeda, Y. Lee, R. Sasaki, Y. Goto, K. Ito, I. Meguro, F. Matsukura, H. Takahashi, H. Matsuoka, and H. Ohno, "2Mb Spin-Transfer Torque RAM (SPRAM) with Bit-by-Bit Bidirectional Current Write and Parallelizing-Direction Current Read," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, 2007, pp. 480-617.

[37] N. Sakimura, T. Sugibayashi, T. Honda, S. Miura, H. Numata, H. Hada, and S. Tahara, "A 512kb cross-point cell MRAM," in *Solid-State Circuits Conference, 2003. Digest of Technical Papers. ISSCC. 2003 IEEE International*, 2003, pp. 278-279 vol.1.

[38] N. Sakimura, T. Sugibayashi, R. Nebashi, H. Honjo, S. Saito, Y. Kato, and N. Kasai, "A 250-MHz 1-Mbit embedded MRAM macro using 2T1MTJ cell with bitline separation and half-pitch shift architecture," in *Solid-State Circuits Conference, 2007. ASSCC '07. IEEE Asian*, 2007, pp. 216-219.

[39] Z. Ping, Z. Bo, Y. Jun, and Z. Youtao, "Energy reduction for STT-RAM using early write termination," in *Computer-Aided Design - Digest of Technical Papers, 2009. ICCAD 2009. IEEE/ACM International Conference on*, 2009, pp. 264-268.

[40] Y. Chen, H. Li, X. Wang, W. Zhu, W. Xu, and T. Zhang, "Combined magnetic- and circuit-level enhancements for the nondestructive self-reference scheme of STT-RAM," in *Low-Power Electronics and Design (ISLPED), 2010 ACM/IEEE International Symposium on*, 2010, pp. 1-6.

[41] C. Yiran, L. Hai, W. Xiaobin, Z. Wenzhong, X. Wei, and Z. Tong, "A nondestructive self-reference scheme for Spin-Transfer Torque Random Access Memory (STT-RAM)," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010*, 2010, pp. 148-153.

[42] C. Yiran, L. Hai, W. Xiaobin, Z. Wenzhong, X. Wei, and Z. Tong, "A 130 nm 1.2 V/3.3 V 16 Kb Spin-Transfer Torque Random Access Memory With Nondestructive Self-Reference Sensing Scheme," *Solid-State Circuits, IEEE Journal of*, vol. 47, pp. 560-573, 2012.

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

[44] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," *Solid-State Circuits, IEEE Journal of*, vol. 39, pp. 1148-1158, 2004.

[45] S. Paul, S. Chatterjee, S. Mukhopadhyay, and S. Bhunia, "Energy-Efficient Reconfigurable Computing Using a Circuit-Architecture-Software Co-Design Approach," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 1, pp. 369-380, 2011.

[46] S. Paul and S. Bhunia, "Reconfigurable computing using content addressable memory for improved performance and resource usage," in *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, 2008, pp. 786-791.

[47] S. Paul, S. Mukhopadhyay, and S. Bhunia, "Hybrid CMOS-STTRAM non-volatile FPGA: Design challenges and optimization approaches," in *Computer-Aided Design, 2008. ICCAD 2008. IEEE/ACM International Conference on*, 2008, pp. 589-592.

[48] S. Paul, S. Mukhopadhyay, and S. Bhunia, "A variation-aware preferential design approach for memory based reconfigurable computing," in *Computer-Aided Design - Digest of Technical Papers, 2009. ICCAD 2009. IEEE/ACM International Conference on*, 2009, pp. 180-183.

[49] J. Singh, J. Mathew, S. P. Mohanty, and D. K. Pradhan, "Single Ended Static Random Access Memory for Low-Vdd, High-Speed Embedded Systems," in *VLSI Design, 2009 22nd International Conference on*, 2009, pp. 307-312.

[50] J. Singh, D. K. Pradhan, S. Hollis, S. P. Mohanty, and J. Mathew, "Single ended 6T SRAM with isolated read-port for low-power embedded systems," in *Design, Automation & Test in Europe Conference & Exhibition, 2009. DATE '09.*, 2009, pp. 917-922.

[51] G. Varghese, J. Sanjeev, T. Chao, K. Smits, D. Satish, S. Siers, N. Ves, K. Tanveer, S. Sanjib, and S. Puneet, "Penryn: 45-nm next generation Intel core™ 2 processor," in *Solid-State Circuits Conference, 2007. ASSCC '07. IEEE Asian*, 2007, pp. 14-17.

[52] I. Sungjun and K. Banerjee, "Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs," in *Electron Devices Meeting, 2000. IEDM Technical Digest. International*, 2000, pp. 727-730.

[53] G. L. Loi, B. Agrawal, N. Srivastava, L. Sheng-Chih, T. Sherwood, and K. Banerjee, "A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy," in *Design Automation Conference, 2006 43rd ACM/IEEE*, 2006, pp. 991-996.

[54] A. Jain, R. E. Jones, R. Chatterjee, and S. Pozder, "Analytical and Numerical Modeling of the Thermal Performance of Three-Dimensional Integrated Circuits," *Components and Packaging Technologies, IEEE Transactions on*, vol. 33, pp. 56-63, 2010.

[55] A. Jain, S. M. Alam, S. Pozder, and R. E. Jones, "Thermal-electrical co-optimisation of floorplanning of three-dimensional integrated circuits under manufacturing and physical design constraints," *Computers & Digital Techniques, IET*, vol. 5, pp. 169-178, 2011.

[56] C. Ting-Yen, S. J. Souri, C. Chi On, and K. C. Saraswat, "Thermal analysis of heterogeneous 3D ICs with various integration scenarios," in *Electron Devices Meeting, 2001. IEDM Technical Digest. International*, 2001, pp. 31.2.1-31.2.4.

[57] M. S. Bakir, C. King, D. Sekar, H. Thacker, D. Bing, H. Gang, A. Naeemi, and J. D. Meindl, "3D heterogeneous integrated systems: Liquid cooling, power delivery, and implementation," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, 2008, pp. 663-670.

[58] D. Sekar, C. King, B. Dang, T. Spencer, H. Thacker, P. Joseph, M. Bakir, and J. Meindl, "A 3D-IC Technology with Integrated Microchannel Cooling," in *Interconnect Technology Conference, 2008. IITC 2008. International*, 2008, pp. 13-15.

[59] H. Hao, C. Mineo, K. Schoenfliess, A. Sule, S. Melamed, R. Jenkal, and W. R. Davis, "Exploring compromises among timing, power and temperature in three-dimensional integrated circuits," in *Design Automation Conference, 2006 43rd ACM/IEEE*, 2006, pp. 997-1002.

[60] S. TEZZARON. (2005, Tezzaron unveils 3d SRAM. .

[61] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 24, pp. 1859-1880, 2005.

[62] R. Vattikonda, W. Wenping, and C. Yu, "Modeling and minimization of PMOS NBTI effect for robust nanometer design," in *Design Automation Conference, 2006 43rd ACM/IEEE*, 2006, pp. 1047-1052.

[63] MOSIS. [Online]. Available: www.mosis.org

[64] J.Li, C.Augustine, S.Salahuddin, and K.Roy, "Modeling of failure probability and statistical design of spin-torque transfer magnetic random access memory (STT MRAM) array for yield enhancement," presented at the IEEE/ACM Design Automation Conference (DAC), 2008.

[65] S. Chatterjee, M. Rasquinha, S. Yalamanchili, and S. Mukhopadhyay, "A Scalable Design Methodology for Energy Minimization of STTRAM: A Circuit and Architecture Perspective," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 19, pp. 809-817, 2011.

[66] M. Rasquinha, D. Choudhary, S. Chatterjee, S. Mukhopadhyay, and S. Yalamanchili, "An energy efficient cache design using spin torque transfer (STT) RAM," presented at the Proceedings of the 16th ACM/IEEE international symposium on Low power electronics and design, Austin, Texas, USA, 2010.

[67] G. H. Loh, S. Subramaniam, and X. Yuejian, "Zesto: A cycle-level simulator for highly detailed microarchitecture exploration," in *Performance Analysis of Systems and Software, 2009. ISPASS 2009. IEEE International Symposium on*, 2009, pp. 53-64.

[68] M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge, and R. B. Brown, "MiBench: A free, commercially representative embedded benchmark suite," in *Workload Characterization, 2001. WWC-4. 2001 IEEE International Workshop on*, 2001, pp. 3-14.

[69] T. L. Gilbert, "A phenomenological theory of damping in ferromagnetic materials," *Magnetics, IEEE Transactions on,* vol. 40, pp. 3443-3449, 2004.

[70] S. Salahuddin, D. Datta, P. Srivastava, and S. Datta, "Quantum Transport Simulation of Tunneling Based Spin Torque Transfer (STT) Devices: Design Trade offs and Torque Efficiency," in *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, 2007, pp. 121-124.

[71] S. Chatterje, S. Salahuddin, S. Kumar, and S. Mukhopadhyay, "Electro-Thermal Analysis of Spin-Torque-Transfer Random Access Memory Arrays," *ACM Journal of Emerging Technolgy and Circuits (JETC) (in communication),* 2012.

[72] D. Mazumdar, X. Liu, B. D. Schrag, W. Shen, M. Carter, and G. Xiao, "Thermal stability, sensitivity, and noise characteristics of MgO-based magnetic tunnel junctions (invited)," *Journal of Applied Physics,* vol. 101, pp. 09B502-09B502-6, 2007.

[73] X. Liu, D. Mazumdar, W. Shen, B. D. Schrag, and G. Xiao, "Thermal stability of magnetic tunneling junctions with MgO barriers for high temperature spintronics," *Applied Physics Letters*, vol. 89, p. 023504, 2006.
[74] "Taurus Medici."
[75] S. Chatterjee, S. Salahuddin, S. Kumar, and S. Mukhopadhyay, "Modeling of the self-heating in STTRAM and analysis of its impact on reliable memory operations," in *Non-Volatile Memory Technology Symposium (NVMTS), 2009 10th Annual*, 2009, pp. 86-89.
[76] S. Chatterjee, S. Salahuddin, S. Kumar, and S. Mukhopadhyay, "Analysis of thermal behaviors of Spin-Torque-Transfer RAM: A simulation study," in *Low-Power Electronics and Design (ISLPED), 2010 ACM/IEEE International Symposium on*, 2010, pp. 13-18.
[77] S. Chatterjee, S. Salahuddin, S. Kumar, and S. Mukhopadhyay, "Impact of Self-Heating on Reliability of a Spin-Torque-Transfer RAM Cell," *Electron Devices, IEEE Transactions on*, vol. 59, pp. 791-799, 2012.
[78] www.itrs.net [Online].
[79] S. Chatterjee, S. Salahuddin, and S. Mukhopadhyay, "Dual-Source-Line-Bias Scheme to Improve the Read Margin and Sensing Accuracy of STTRAM in Sub-90-nm Nodes," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 57, pp. 208-212, 2010.
[80] J.Li, H.Liu, S.Salahuddin, and K.Roy, "Variation-tolerant Spin-Torque Transfer (STT) MRAM array for yield enhancement," presented at the IEEE Custom Integrated Circuits Conference (CICC), 2008.
[81] S.Yoon, S.Kang, and M.Sani, "Word Line Transistor Strength Control for Read and Write in Spin Transfer Torque Magnetoresistive Random Access Memory," 2008.
[82] "BPTM 65nm: Berkeley Predictive Technology Model.."
[83] B.Wicht, T.Nirschl, and D.Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 39, pp. 1148-1158, July 2004.
[84] S.Salahuddin and S.Datta, "Self-Consistent Simulation of Hybrid Spintronic Devices," presented at the IEEE Electron Devices Meeting (IEDM), Dec 2006.
[85] S. Paul, S. Chatterjee, S. Mukhopadhyay, and S. Bhunia, "Nanoscale reconfigurable computing using non-volatile 2-D STTRAM array," in *Nanotechnology, 2009. IEEE-NANO 2009. 9th IEEE Conference on*, 2009, pp. 880-883.
[86] "Predictive Technology Models: Available online at http://www.eas.asu.edu/~ptm/latest.html."
[87] "Full CAD Flow for Heterogeneous FPGAs: Version 5 available online at: http://www.eecg.utoronto.ca/vpr/."
[88] V. Naveen and A. P. Chandrakasan, "A 65nm 8T Sub-Vt SRAM Employing Sense-Amplifier Redundancy," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, 2007, pp. 328-606.
[89] S. Paul, S. Chatterjee, S. Mukhopadhyay, and S. Bhunia, "A circuit-software co-design approach for improving EDP in reconfigurable frameworks," in *Computer-*

*Aided Design - Digest of Technical Papers, 2009. ICCAD 2009. IEEE/ACM International Conference on*, 2009, pp. 109-112.

[90] M. Khellah, Y. Yibin, K. Nam Sung, D. Somasekhar, G. Pandya, A. Farhang, K. Zhang, C. Webb, and V. De, "Wordline & Bitline Pulsing Schemes for Improving SRAM Cell Stability in Low-Vcc 65nm CMOS Designs," in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, 2006, pp. 9-10.

[91] C. C. Liu, I. Ganusov, M. Burtscher, and T. Sandip, "Bridging the processor-memory performance gap with 3D IC technology," *Design & Test of Computers, IEEE*, vol. 22, pp. 556-564, 2005.

[92] Banerjee Kaustav, Souri Shukri J., K. Pawan, and S. K. C., "3-D Heterogeneous ICs: A Technology for the Next Decade and Beyond," presented at the 5th IEEE Workshop on SIGNAL PROPAGATION ON INTERCONNECTS, Venice, Italy, 2001.

[93] K. Puttaswamy and G. H. Loh, "Implementing caches in a 3D technology for high performance processors," in *Computer Design: VLSI in Computers and Processors, 2005. ICCD 2005. Proceedings. 2005 IEEE International Conference on*, 2005, pp. 525-532.

[94] K. Puttaswamy and G. H. Loh, "Thermal Herding: Microarchitecture Techniques for Controlling Hotspots in High-Performance 3D-Integrated Processors," in *High Performance Computer Architecture, 2007. HPCA 2007. IEEE 13th International Symposium on*, 2007, pp. 193-204.

[95] Sun G, Wu X, and X. Y, "Exploration of 3D Stacked L2 Cache Design for High Performance and Efficient Thermal Control," in *ISLPED*, San Francisco, 2009, pp. 295-298.

[96] Y. Woojin, K. Kyungsu, and K. Chong-Min, "Thermal-aware energy minimization of 3D-stacked L3 cache with error rate limitation," in *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, 2011, pp. 1672-1675.

[97] S. Chatterjee, C. Minki, R. Rao, and S. Mukhopadhyay, "Impact of die-to-die thermal coupling on the electrical characteristics of 3D stacked SRAM cache," in *Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2012 28th Annual IEEE*, 2012, pp. 14-19.

[98] "asu.ptm.edu."