DESIGN AND ROBUSTNESS ANALYSIS ON NON-VOLATILE STORAGE AND LOGIC CIRCUIT

by

Peiyuan Wang

B.S., Tsinghua University, P. R. China, 2010

Submitted to the Graduate Faculty of

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science in Electrical Engineering

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

Peiyuan Wang

It was defended on

Oct. 18, 2011

and approved by

Yiran Chen, PhD, Assistant Professor, Electrical and Computer Engineering Department

Steven P. Levitan, PhD, John A. Jurenko Professor, Electrical and Computer Engineering Department

Jun Yang, PhD, Associate Professor, Electrical and Computer Engineering Department

Thesis Advisor: Yiran Chen, PhD, Assistant Professor, Electrical and Computer Engineering Department

DESIGN AND ROBUSTNESS ANALYSIS ON NON-VOLATILE STORAGE AND LOGIC CIRCUIT

Peiyuan Wang, M.S.

University of Pittsburgh, 2011

By combining the flexibility of MOS logic and the non-volatility of spintronic devices, spin-MOS logic and storage circuitry offer a promising approach to implement highly integrated, power-efficient, and nonvolatile computing and storage systems. Besides the persistent errors due to process variations, however, the functional correctness of Spin-MOS circuitry suffers from additional non-persistent errors that are incurred by the randomness of spintronic device operations, i.e., thermal fluctuations. This work quantitatively investigates the impact of thermal fluctuations on the operations of two typical Spin-MOS circuitry: one transistor and one magnetic tunnel junction (1T1J) spin-transfer torque random access memory (STT-RAM) cell and a nonvolatile latch design. A new nonvolatile latch design is proposed based on magnetic tunneling junction (MTJ) devices. In the standby mode, the latched data can be retained in the MTJs without consuming any power. Two types of operation errors can occur, namely, persistent and non-persistent errors. These are quantitatively analyzed by including models for process variations and thermal fluctuations during the read and write operations. A mixture importance sampling methodology is applied to enable yield-driven design and extend its application beyond memories to peripheral circuits and logic blocks. Several possible design techniques to reduce thermal induced non-persistent error rate are also discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Over the past year I have learned a tremendous amount. I am grateful to everyone whom I have learned something from. In particular I thank my supervisor, Professor Yiran Chen, for all of his support and guidance in this work.

I am very thankful to Dr. Steven P. Levitan and Dr. Jun Yang for their kindly consenting to be my committee members, helpful suggestion and useful discussion. I thank my closest collaborators and the rest of my research group here in University of Pittsburgh.

I thank my families who have always been a great support for me. I thank Qian Chen for always believing in me, even when I do not believe in myself. I could not have progressed so far without all of you.

1.0 INTRODUCTION

This thesis presents a design and robustness analysis of both non-volatile store and logic circuits in a 45-nm process. The work analyzes the various aspects of non-volatile memory and logic, focusing on challenges such as process variation, energy-efficiency, and write/read performance.

1.1 BACKGROUND

Current (embedded) computer systems cannot be turned on and off quickly. This inconvenience has limited user behavior; even though systems are not being used, we tend to leave them on anyways. An enormous amount of power is being wasted by keeping these devices on while doing nothing. A solution such as using sleep mode turns off most of the devices, but it leaves the memory on, albeit in low power mode. Hence, it is still wasting energy. Hibernation, which is another such solution, saves the contents of DRAM on non-volatile storage at system turn off and allows all the devices in the system to be powered off. Though hibernation does not waste energy, it still takes a considerable amount of time for the system to turn off and on. Another solution example is “quick boot” that aims to boot the system quickly while limiting the things one can do with the system. In contrast, we would like our system to not only provide instant on/off, but also provide full functionality. Here, the term “instantly” is a subjective term that refers to a short time span, typically less than 1 or 2 seconds, which is where a human feels that an event has occurred without delay. By “full functionality”, we refer to the operating system state that a conventional operating system would provide.

1.2 MOTIVATION

The key idea behind realizing a fully functional instant on/off system is the adoption of new non-volatile component. In conventional systems that employ SRAM and DRAM, when power is turned off, the system loses all its memory state. Hence, recovering the stored state from right before power-down (or returning to system initial state) is the key source of delay when turning systems on. As a non-volatile system, if non-volatile elements retain the system state, then it is not lost with power-down. To build such a system, not only memory should be non-volatile, but also computation should start from where it stops instantly, hence the registers in pipeline or state machine must be non-volatile so that the system can wake up from standby mode to the exact state in a relatively short period.

1.3 PREVIOUS WORK

This section reviews various non-volatile memory and previous work done on non-volatile logic. Robustness analysis including process variation and thermal fluctuation is introduced.

1.3.1 Non-volatile random access memory

One part we can easily think of to add non-volatility is random access memory. Non-volatile random-access memory (NVRAM) is random-access memory that retains its information when power is turned off, which is described technically as being non-volatile. This is in contrast to the most common forms of random access memory today, which require continual power in order to maintain their data.

The best-known form of NVRAM memory today is flash memory. Some drawbacks to flash memory include the requirement to write it in larger blocks than many computers can atomically address, and the relatively limited longevity of flash memory due to its finite number of write-erase cycles (most consumer flash products at the time of writing can only withstand around 100,000 rewrites before memory begins to deteriorate). Another drawback is the performance limitations preventing flash from matching the response times and, in some cases, the random addressability offered by traditional forms of RAM. Flash and EEPROM’s limited write-cycles are a serious problem for any real RAM-like role, however. Additionally, the high power needed to write the cells is a problem in low-power roles, where NVRAM is often used.

To date, the only such system to enter widespread production is ferroelectric RAM, or FeRAM. FeRAM uses a ferroelectric layer in a cell that is otherwise similar to conventional DRAM, this layer holding the charge in a 1 or 0 even with the power removed. To date, FeRAM has been produced on lines with large feature sizes, and even the most advanced research samples are still twice the line width of most flash devices. Although this difference might be addressable under normal circumstances, as flash moves to multi-bit cells the difference in memory density appears to be growing, rather than shrinking.

Another approach to see major development effort is Magnetic Random Access Memory, or MRAM, which uses magnetic elements and generally operates in a fashion similar to core. 1st generation MRAM utilizes cross-point field induced writing. Two 2nd generation techniques are currently in development: Thermal Assisted Switching (TAS) and Spin Torque Transfer (STT). Another solid-state technology to see more than purely experimental development is Phase-change RAM or PRAM. PRAM reads them based on their changes in electrical resistance rather than changes in their optical properties [1].

1.3.2 Non-volatile logic

There is another type of circuit with memory property as part of sequential logic which is called registers (usually implemented as D flip-flops). In electronics, a flip-flop or latch is a circuit that has two stable states and can be used to store state information. The circuit can be made to change state by signals applied to one or more control inputs and will have one or two outputs. It is the basic storage element in sequential logic. Flip-flops and latches are a fundamental building block of digital electronics systems used in computers, communications, and many other types of systems.

Flip-flops can be either simple (transparent or opaque) or clocked (synchronous or edge-triggered); the simple ones are commonly called latches. The flip-flop is one of the most important building blocks for logical circuits since it synchronizes and stores the intermediate computing data. It is now feasible for a nonvolatile memory cell to be connected directly to each flip-flop in a microprocessor with minimal area overhead. The flip-flop with non-volatility allows the logic circuits to be powered off completely in sleep mode and all the data can be retrieved instantly. A nonvolatile synchronous flip-flop circuit that uses a nanoscale memristive device as the nonvolatile memory element was implemented and tested [2]. Several MTJ based non-volatile flip-flop/latches were proposed [3, 4]. The integration of digital logic devices and non-volatile memory cell could open the way for nonvolatile computation with applications in small platforms that rely on intermittent power sources.

1.3.3 Robustness analysis

Process variation has always been a critical aspect of semiconductor fabrication,n. It can severely affect circuit stability and performance. Circuit robustness is extremely important to non-volatile logic. The reason is that whether the right data stored can be read successfully determines the reboot time and correctness. For CMOS, process variations include random dopant fluctuations (RDFs), line-edge roughness (LER), shallow-trench isolation (STI) stress and the geometry variations of transistor channel length/width. Besides the geometry variations, most of the CMOS process variations are reflected as threshold voltage deviations. The random variation of the threshold voltage is prominent in scaled CMOS technology. Besides CMOS process variation, the integration of a non-volatile device has induced its own process variation and the intrinsic thermal fluctuations [5].

To use MTJ as an example, in general, the impact of thermal fluctuations can be modeled by the thermal induced random filed h_{fluc} in stochastic Landau-Lifshitz-Gilbert (LLG) equation [6-8] as,

$$\frac{d\vec{m}}{dt} = \vec{m} \times \left(\vec{h}_{eff} + \vec{h}_{fluc}\right) - \alpha\vec{m} \times \left(\vec{m} \times \left(\vec{h}_{eff} + \vec{h}_{fluc}\right)\right) + \frac{\vec{T}_{norm}}{M_s} \qquad (1.1)$$

Where $\vec{m}$ is the normalized magnetization vector, time t is normalized by γM_s. γ is the gyro-magnetic ratio and M_s is the magnetization saturation. α is the LLG damping parameter. $\vec{h}_{eff} = \vec{H}_{eff}/M_s$is the normalized effective magnetic field and $\vec{h}_{fluc}$is the normalized thermal agitation fluctuating field at finite temperature. $\vec{T}_{norm} = \frac{\vec{T}}{M_s V}$ is the spin torque term with units of magnetic field. The net spin torque $\vec{T}$ can be obtained through microscopic quantum electronic spin transport model. The thermal field's effect on the magnetization vector will influence switching performance of MTJ.

1.4 CONTRIBUTION OF THIS THESIS

This thesis presents the robustness analysis of non-volatile memory and logic design. Two types of operation errors, namely, persistent and non-persistent errors, are quantitatively analyzed by including the process variations and thermal fluctuations during the read and write operations in spin-MOS logic and storage circuitry. A new nonvolatile latch design is proposed based on magnetic tunneling junction (MTJ) devices to offer a promising approach to implement highly integrated, power-efficient, and nonvolatile computing and storage systems. A mixture importance sampling methodology is applied to enable yield-driven design and extend its application beyond memories to peripheral circuits and logic blocks. The possible design techniques to reduce thermal incurred non-persistent error rates are also discussed.

1.5 SUMMARY

In this chapter, the advantage of instant on/off systems was presented. The key ideas to realize a fully functional instant on/off system were outlined, with particular attention paid to non-volatile memory. The two main objects of non-volatility (memory and registers) were introduced. The variation sources in circuit design were analyzed. Finally, the design and robustness analysis of both non-volatile store and logic circuits in a 45-nm process was then introduced as the goal of this thesis.

2.0 NON-PERSISTENT ERRORS OPTIMIZATION IN SPIN-MOS LOGIC AND STORAGE CIRCUITRY

2.1 INTRODUCTION

Spin torque induced magnetization switching in magnetic tunneling junctions (MTJs) is the fundamental of modern spintronic memory, which features nanosecond access time, high programming endurance, nonvolatility and zero standby power [9]. By combining the flexibility of MOS logic and the non-volatility of spintronic devices, Spin-MOS logic and storage circuitry make it possible to implement high-density, low-power, nonvolatile and robust computing and storage systems. Besides the spin-transfer torque random access memory (STT-RAM) [10], spintronic devices have been also used in timing sequential circuitry and simple logics, such as latches [11] and lookup tables [12]. Moreover, prior art has shown that compared to conventional MOS logic whose functionalities are based on the operation of electrical charge, Spin-MOS circuitry are more resilient to soft errors, which are primarily generated by the Alpha particle emissions from chip packaging materials [13].

The functional errors of a circuit can be categorized as either non-persistent or persistent [14]. An error is persistent if it happens deterministically and can be repeated after the chip is fabricated, such as the errors introduced by process variations. The non-persistent errors include those introduced by soft-errors in CMOS circuitry or by thermal fluctuations in Spin-MOS circuitry. In Spin-MOS circuitry, thermal-induced non-persistent errors demand specific optimization design techniques so that even soft-errors are eliminated. In this work, we quantitatively investigate the impact of thermal fluctuations on the operation of two typical Spin-MOS circuitry: a 1T1J spin-transfer torque random access memory (STT-RAM) cell and a nonvolatile flip-flop design. In addition, we exploit the possibility of minimizing the thermally induced non-persistent error rate while taking into account this adverse impact on persistent errors.

The rest of this chapter is organized as follows. Section 2.2 gives a preliminary overview on persistent and non-persistent errors in Spin-MOS circuitry by using a STT-RAM cell as the example. It depicts the quantitative analysis of the impact of non-persistent errors on the operation of Spin-MOS circuitry and its optimization. It also discusses the tradeoff between persistent and non-persistent errors. In Section 2.3, a more complicated case study a nonvolatile flip-flop is presented. Finally the work is concluded in Section 2.4.

2.2 CASE STUDY ON STT-RAM CELLS

2.2.1 Persistent errors in spin-MOS circuitry

Fig. 2.1 shows the ordinary 1T1J (one-transistor-one-MTJ) STT-RAM cell design, where a MTJ is connected to a NMOS transistor [10]. The MTJ resistance can be changed between the high and the low state under a polarized switching current. It is well-known that the switching time of a MTJ is determined by the switching current: the increases on the switching time leads to the reduction on the switching current [10]. Moreover, when the MTJ switching time is under 10ns, the further scaling of switching time will cause an exponential increase in switching current, as shown in Fig. 2.2. Here, the switching current and time are achieved based on MTJ with a 45×90nm ellipse shape.

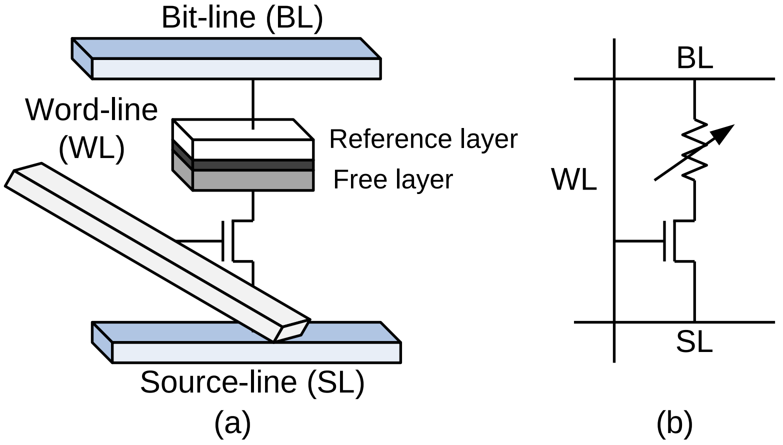


Figure 2.1. 1T1J STT-RAM cell. (a) Cell view. (b) Equivalent schematic.

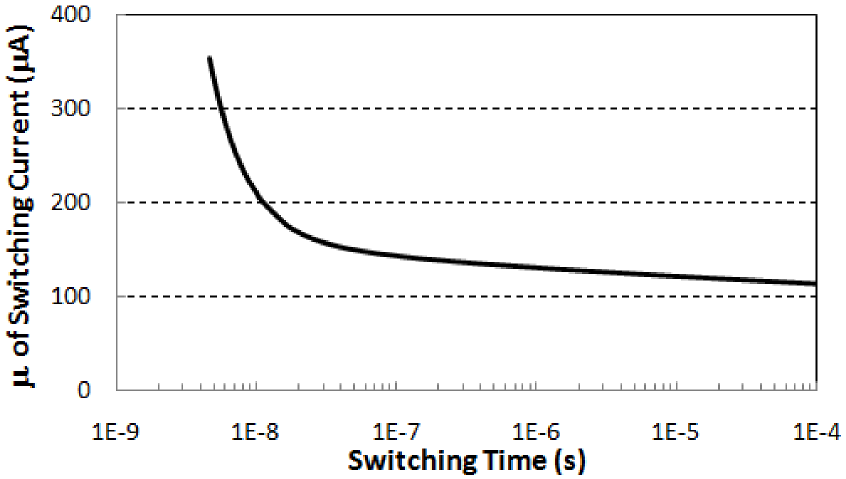


Figure 2.2. Relationship between MTJ switching time and switching current.

Due to the process variation, e.g., the variations of NMOS transistor channel width (W), channel length (L), and threshold voltage (V_t), the current provided by the NMOS transistor to the MTJ varies from memory cell to cell or even from chip to chip associated with the variations of MTJ switching time. Because the parameters of NMOS transistors are fixed after the chip is fabricated, the corresponding errors incurred by the transistor variations, e.g., MTJ fails to switch within the applied write pulse width, are persistent. Similarly, the MTJ geometry and resistance variations, which may cause the MTJ driving current to shift by changing the bias conditions of the NMOS transistor, are also fixed after the chip is fabricated. Therefore, the corresponding errors are also persistent.

Another important persistent error in STT-RAM design is the fault sensing due to the device mismatch in the sense amplifier and/or the small sense margin. During the read operation, a read current, I_r, is injected into STT-RAM cell and generates the corresponding bitline voltage V_{BL}. Then the MTJ resistance state can be obtained by comparing V_{BL} to a reference voltage V_{ref} in the sense amplifier (SenAmp), as shown in Fig. 2.3. However, if the sizes and the threshold voltages of the six MOS transistors (highlighted in RED) deviate from their designed values too much, or the difference between V_{BL} and V_{ref} is too small, SenAmp may give a false result.

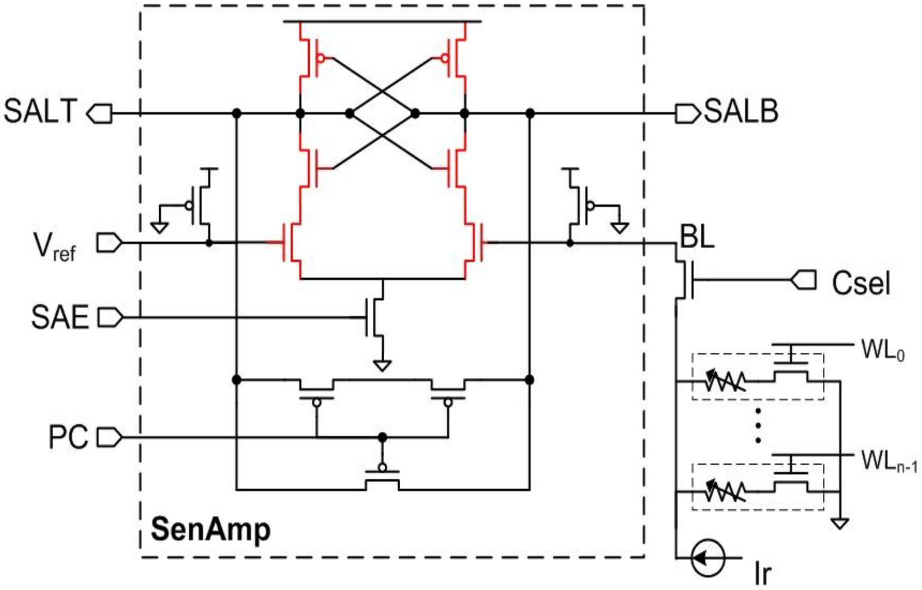


Figure 2.3. Conceptual sense amplifier design.

2.2.2 Non-persistent errors in spin-MOS circuitry

In a Spin-MOS circuit, there are two major non-persistent errors, which occur in write or read operations, respectively. The first one is due to the thermal fluctuation in the write operation of STT-RAM cells. When a MTJ works in a long time region (>10ns), the thermal fluctuation is dominated by the thermal component of internal energy; when MTJ works in sub-10ns region, the thermal fluctuation is dominated by the thermally activated initial angle of procession [15]. The existence of thermal fluctuation causes the deviation of the MTJ switching time from its nominal value. Also, following the increases in MTJ switching current, the ratio between the standard deviation and the mean of MTJ switching time decreases first, mainly due to the increased impact of spin-torque on MTJ switching. Then it increases again after the mean of MTJ switching time enters sub-10ns region and the thermally activated initial angle of procession dominates, as shown in Fig. 2.4. If a MTJ cannot switch by the end of write pulse width, an error will be generated.

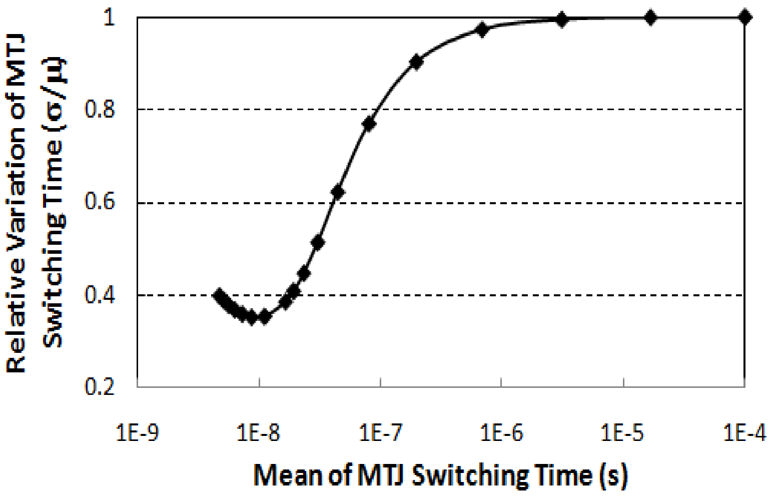


Figure 2.4. Variations of MTJ switching time due to thermal fluctuations.

The second non-persistent error is read disturbance, which denotes the undesired MTJ switching during the read operation. In Ref [16], it is pointed out that the disturbance probability (Pr_{dis}) of a MTJ at a read current of I_R can be expressed as:

$$\mathrm{Pr}_{\mathrm{dis}} = 1 - \exp\left\{-\frac{t}{\tau}\exp\left[-\Delta\left(1-\frac{I_{\mathrm{R}}}{I_{\mathrm{C}}}\right)\right]\right\}, \tag{2.1}$$

which has been proven in [17]. Here t is the read current pulse width. Δ is the magnetic memorizing energy without applying any current or magnetic fields. τ is the inverse of the attempted frequency. I_C is the critical switching current, which is the minimum current amplitude to switch the MTJ resistance with a write pulse width of τ. Usually the read current pulse width is fixed by the timing control circuit. Therefore, Pr_{dis} is mainly determined by the read current amplitude. Fig. 2.5 shows the read disturbance probability of the simulated MTJ with 10ns read pulse width under various read currents. Compared to the non-persistent error resulted by the thermal fluctuation in the write operation, the impact of read disturbance is much smaller.

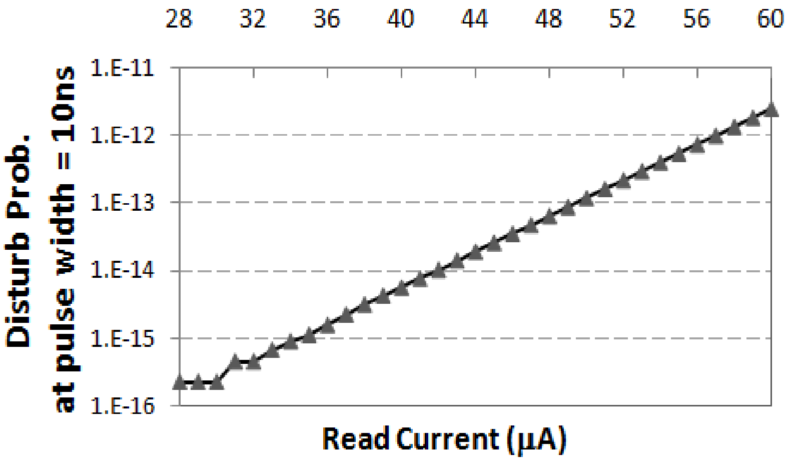


Figure 2.5. The variations of MTJ read disturbance probability when read current amplitude changes.

2.2.3 Minimizing non-persistent errors

The straightforward way to minimize the non-persistent errors in the write operation of STT-RAM cell is increasing the switching current, or sizing up the NMOS transistor. As shown in Fig. 2.2 and Fig. 2.4, increasing switching current (e.g., by increasing the NMOS transistor size) can produce a tighter distribution of the required switching time by reducing both the mean and the standard deviation of switching time.

Fig. 2.6 shows the simulated rate that MTJ fails to switch within the given switching pulse width by varying NMOS transistor size in the range from 270nm to 990nm. At each node, 1000 Monte-Carlo simulations were conducted with the thermal fluctuation in consideration. Here, the switching pulse width increases from 10ns to 16ns with a step of 2ns.

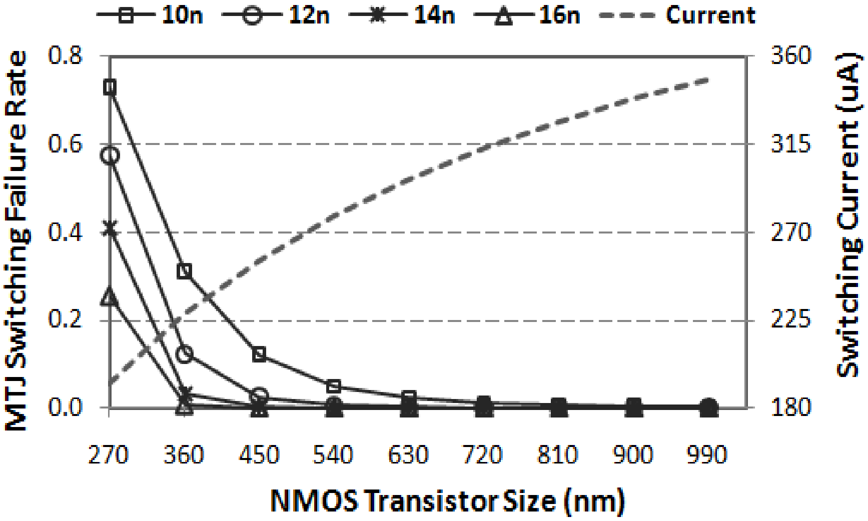


Figure 2.6. The non-persistent failure rate of MTJ as the transistor size varies.

The simulation results show that increasing the size of the NMOS transistor, and hence, increasing switching current in a STT-RAM cell can effectively reduce the MTJ switching failure rate when the NMOS is small. However, further increasing the NMOS transistor size (e.g., when NMOS is wider than 360nm at 16ns switching pulse width for 45mm process) does not improve MTJ switching failure rate much. Although the mean of MTJ switching time goes below 10ns when the NMOS channel width is larger than 360nm, significant timing error rate can still be observed due to the variations of MTJ switching performance. In the practice of STT-RAM cell design, the target write error rate is usually predetermined by the memory specification.

The minimization of read disturbance probability is usually achieved by controlling the read current amplitude through a clamping magnetic field which may be applied to enhance the MTJ stability during the read operation [18].

2.2.4 Tradeoff between persistent and non-persistent errors

In STT-RAM designs, the amplitude of the read current is usually controlled by a global read driver. We note that the sense margin of a STT-RAM cell ΔV is proportional to $I_R \cdot \Delta R$, where ΔR is the difference between the high- and the low-resistance states of the MTJ. Thus, reducing read current, I_R, minimizes the read disturbance probability while simultaneously increasing the sensing error rate due to the degraded sense margin.

The Monte-Carlo simulation results are shown in Fig. 2.7. Here, we assume the standard deviation of the NMOS transistor channel width and length are 5% of their nominal values and the standard deviation of V_t for minimum is 33mV for 45nm process. The variations of MTJ read disturbance probability with various read current amplitudes can be seen in Fig. 2.5. The simulation of a conventional sense amplifier generated by applying process variation on CMOS transistors shows that the sense margin degradation dominates within the given read current range (<60 μA). For the given MTJ device, the read disturbance increases sharply after the read current exceed this value. Therefore, the non-persistent errors due to the read disturbance start dominating, which is hard to control in design.

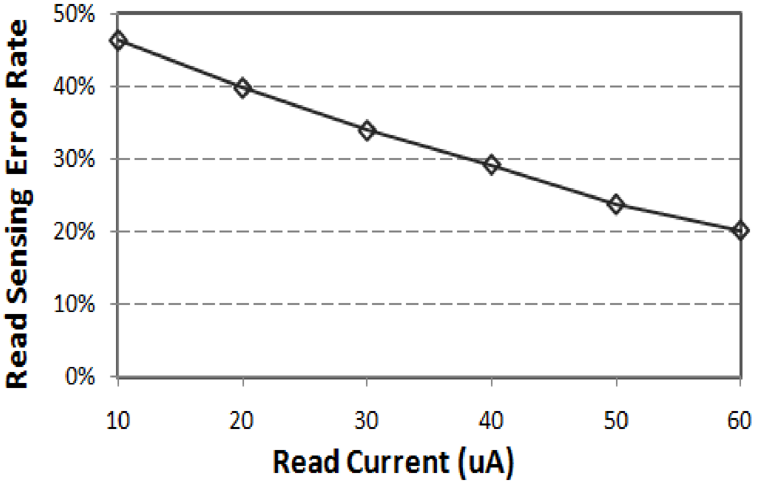


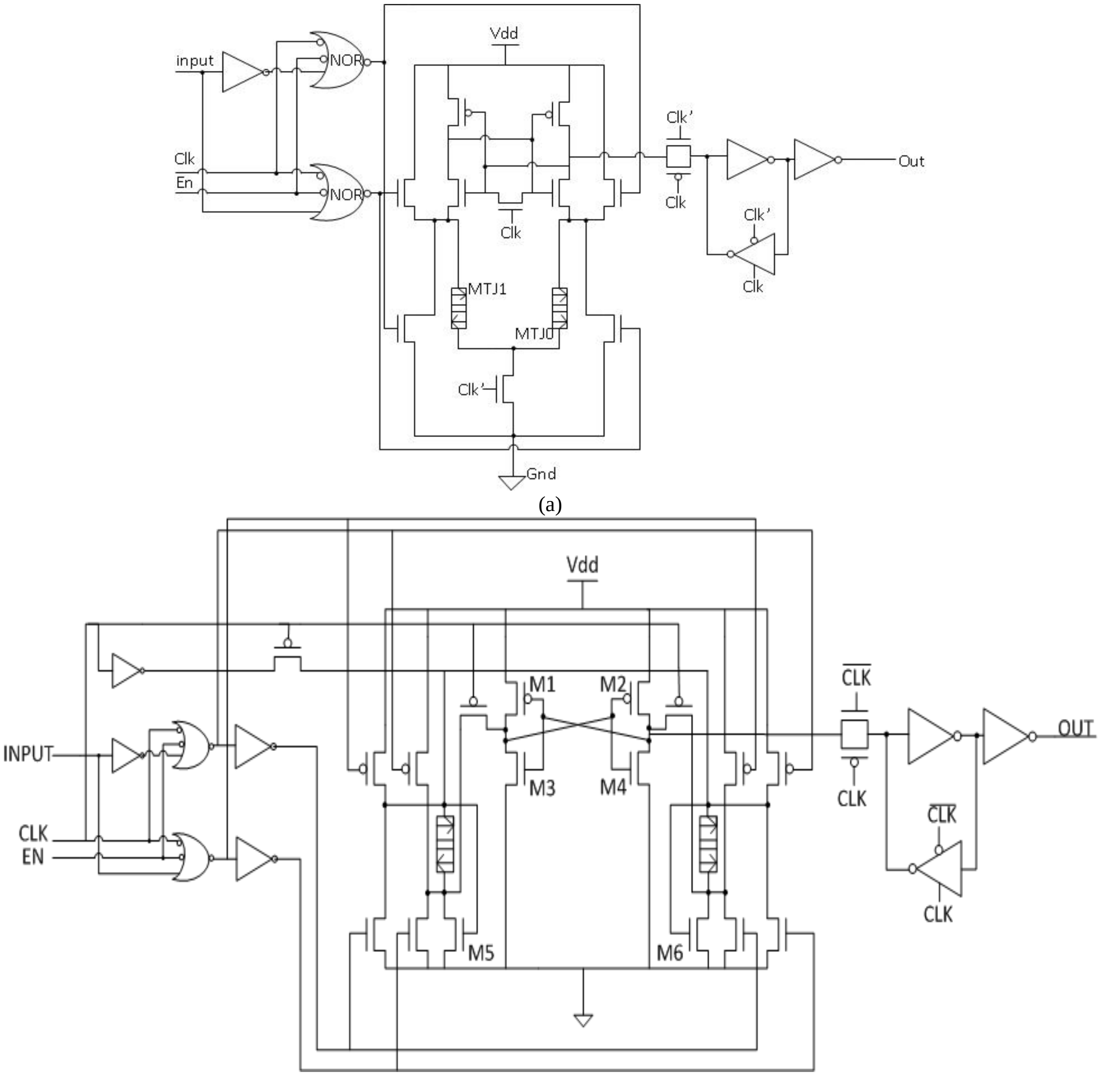
Figure 2.7. Tradeoff between read disturbance probability and sensing errors.

2.3 CASE STUDY ON NONVOLATILE FLIP-FLOP

The non-volatility of MTJ devices are also utilized in other circuit component designs, e.g., Flip-Flops. Fig. 2.8(a) shows a recently proposed nonvolatile flip-flop design where two MTJs are embedded into the traditional flip-flop design with opposite stack structures [11]. In the normal operation, the whole flip-flop works as the conventional flip-flop. When the circuit is entering standby or power down mode, the ‘EN’ signal is raised and the stored value is written into the two MTJs by a current whose direction is controlled by the stored value.

One disadvantage of this design is that the write path always includes two NMOS and two MTJs. The large voltage drops across the MTJs degrade the driving ability of MOS transistors by reducing the voltage difference between their gate and source (V_{gs}).

In this work, we proposed a new flip-flop design with separated write paths of the two MTJs to overcome the above disadvantage, as shown in Fig. 2.8(b). Each MTJ has its own PMOS-NMOS transistor pair to supply the switching current during the write operation. Obviously, the size of PMOS-NMOS transistor pair must be sufficiently large to minimize the non-persistent timing errors during the write operations.



(a)

(b)

Figure 2.8. Nonvolatile Flip-flop designs. (a) Original design in [11]. (b) Our modified design.

When the nonvolatile flip-flop wakes up from the standby mode, the difference between the resistances of two MTJs is sensed. Similar to the SenAmp in STT-RAM designs, the device mismatch among the cross coupled inverters (M1-M4) may cause false sensing when the generated voltage difference on the two MTJs is too small. The new design connects MTJ and NMOS transistor (M5 or M6) in series to provide credible inputs for M1-M4. Depending on the data stored in two MTJs, one of M5 and M6 works in saturation region and another works in linear region. By properly sizing M5 and M6, the currents through MTJs can be adjusted, and consequently, the design can be more process-variation tolerant. However, similar to STT-RAM cell designs, increasing the read current may result in the increase in read disturbance probability.

Fig. 2.9 shows the non-persistent write failure rate of our nonvolatile flip-flop when increasing the size of PMOS-NMOS write-driver pair. Here, we assume the PMOS transistor size is always twice of the NMOS transistor size. The failure rate follows a similar trend to the one of the STT-RAM.

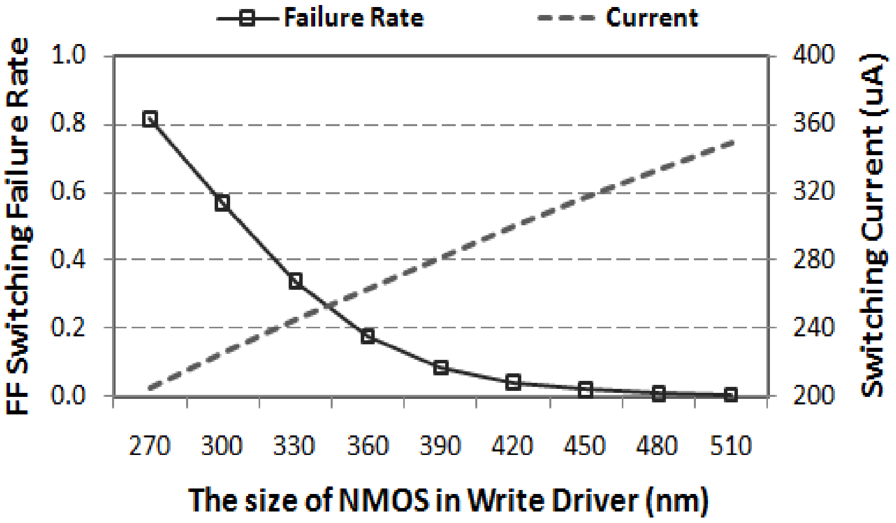


Figure 2.9. The non-persistent write failure rate of nonvolatile flip-flop when the transistor size varies.

Fig. 2.10 shows the tradeoffs between the persistent and non-persistent read errors when sizing up M5-M6. We use the same simulation setup as Section 2.2.4 for process variations. As expected, increasing the size of these transistors can produce higher reading current through the MTJ pair. Hence, the persistent errors due to process variation can be reduced significantly.

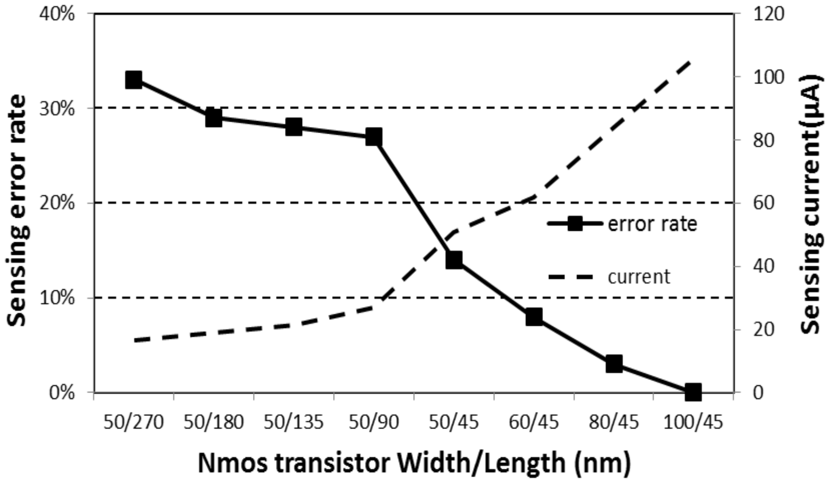


Figure 2.10. Tradeoff between read disturbance probability and sensing errors of nonvolatile flip-flop when the transistor size varies.

2.4 CONCLUSION

In this work, we thoroughly analyzed the persistent and non-persistent errors in Spin-MOS circuitry: the former mainly comes from process variations, and the later one results from thermal fluctuations and read disturbance. Additionally, we quantitatively investigated the impact of these variations and fluctuations on the operations of two typical Spin-MOS circuitry: 1T1J spin-transfer torque random access memory (STT-RAM) cell and a nonvolatile flip-flop design. The possible design techniques to reduce thermal incurred non-persistent error rate were also discussed. Our experimental results show that the optimization of non-persistent and persistent errors are closely entangled with each other and should be conducted from both circuit design and magnetic device engineering perspectives simultaneously.

3.0 A 1.0V 45NM NONVOLATILE MAGNETIC LATCH DESIGN AND ITS ROBUSTNESS ANALYSIS

3.1 INTRODUCTION

Technology scaling rapidly increases the power density and clock frequency of systems, which makes low-power design essential to modern VLSI systems. As a popular technique, standby mode can inactivate the unnecessary circuit module for dynamic and leakage power reduction. When more hardware resources are required, these circuit modules can be activated again.

In synchronous circuits, the sequential devices (such as latches and flip-flops) are placed at the outputs of pipeline stages to maintain the timing. To implement the "instant-on" concept, some data retention techniques must be applied during the standby mode: for example, data can be stored in the shadow latches that are powered by a separate source [19]. When switching back to the active mode, the system states can be restored from the shadow latches. However, such a design may introduce large leakage power and power routing overhead when the number of the shadow latches is large.

Recent research on the emerging memories inspired the use of nonvolatile sequential circuit designs. By storing the data in nonvolatile devices, the power supply can be safely removed during the standby mode while still maintaining the "instant-on" capability. Among all the emerging nonvolatile devices, magnetic tunneling junction (MTJ) is a promising candidate in high-speed sequential circuit design for its nanosecond programming time, high endurance, and good CMOS process compatibility [20]. Many MTJ-based nonvolatile flip-flop designs have been reported in past years [20-23]. However, all these designs include the MTJs in the latch loop and generally suffer from slow data backup/recovery time and poor process variation tolerance.

In this chapter, we present a new MTJ-based nonvolatile latch design for standby mode usage. The performance and robustness of our latch design are improved by separating the paths of recovery signal generation, sensing, and MTJ writing. We discuss persistent and non-persistent errors in our MTJ-based latches designs, which are caused by the process variations and the thermal fluctuations, respectively. The design tradeoffs for error reduction are also investigated.

The rest of this chapter is organized as follows: Section 3.2 gives the fundamentals of MTJ devices and MTJ-based nonvolatile latch designs; Section 3.3 presents our latch designs; Section 3.4 discusses the operation errors and the design tradeoff; Section 3.5 shows our experimental results; and Section 3.6 concludes our work.

3.2 PRILIMINARY

3.2.1 MTJ basics

MTJ has been widely used as the data storage device in spin-transfer torque random access memory (STT-RAM). As shown in Fig. 3.1, an MTJ is composed of two ferromagnetic layers (FLs) and one oxide barrier layer, e.g., MgO. When the magnetization directions (MDs) of the two FLs are parallel (anti-parallel), MTJ is in low (high) resistance state. The MD of one FL (reference layer) is pinned while the MD of the other FL (free layer) is switchable: when a current passes through the MTJ from B (A) to A (B), the MD of free layer flips to be parallel (anti-parallel) to that of reference layer [24].

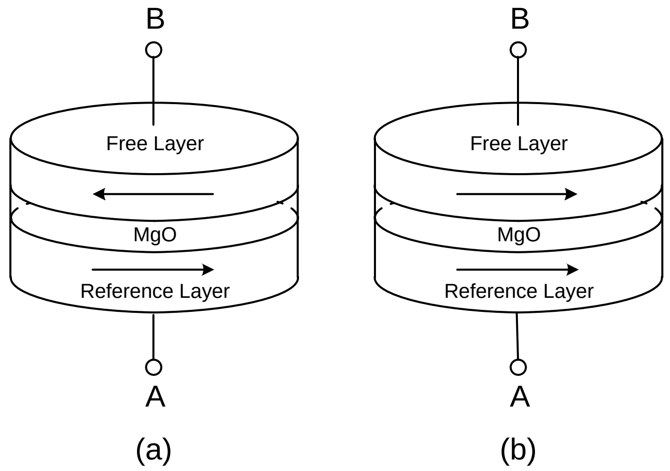


Figure 3.1. MTJ structure. (a) Anti-parallel state “1”. (b) Parallel state “0”.

The switching time of MTJ resistance is determined by the amplitude of the applied MTJ switching current, as shown in Fig. 3.2. The required MTJ switching current increases when the MTJ switching time decreases.

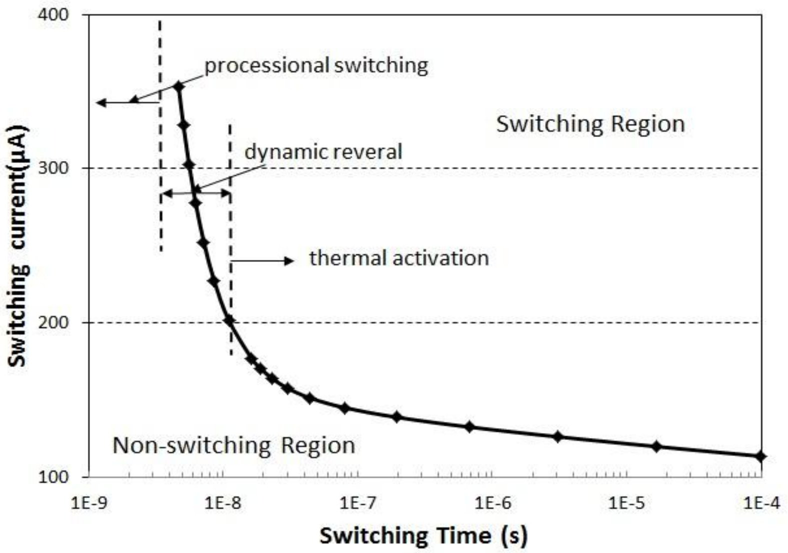


Figure 3.2. Switching time vs. switching current for a MTJ with 45×90nm ellipse shape.

3.2.2 Conventional MTJ-based latch design

Fig. 3.3 shows the schematic of an SRAM-cell based non-volatile latch design by using MTJ as data storage element [20]. A pair of MTJs is embedded below the two back-to-back connected inverters. As pointed out by the authors, this design has two major issues: 1) the MOS transistors connected to the MTJ must be sufficiently large to supply the required write current. Therefore, the normal latch operation speed is degraded due to the large parasitic capacitances at the output/ input nodes; 2) the two MTJs reduce the actual voltage applied to the SRAM structure. Consequently a higher operation voltage is required to ensure the correct circuit functions [20]: 2.6V supply voltage is needed by the latch implemented with 1.5V process, which incurs severe reliability concerns.

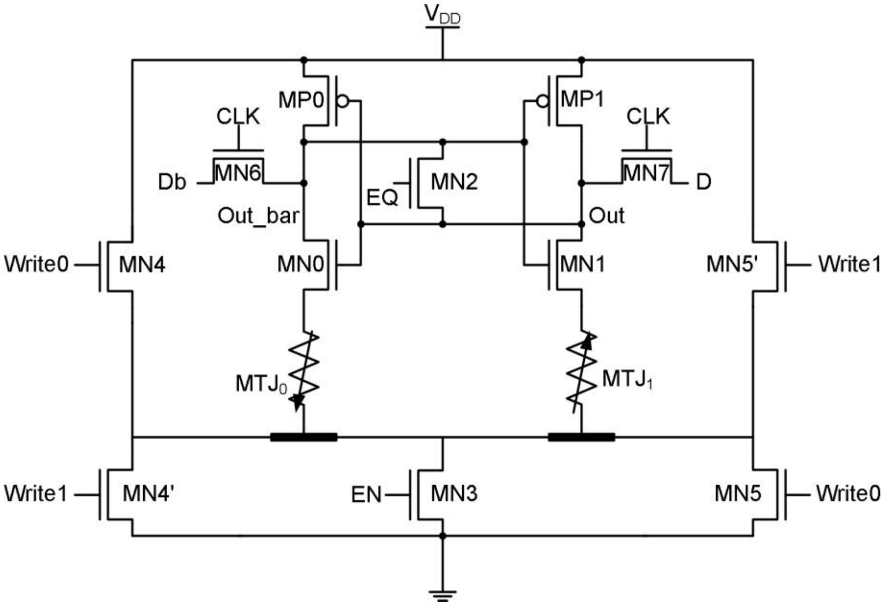


Figure 3.3. Schematic of the existing MTJ-based nonvolatile latch [20].

3.3 PROPOSED MAGNETIC LATCH

Our proposed latch design is shown in Fig. 3.4. Although a SRAM cell structure is still adopted, the MTJ pair is moved out of the inverter loop and controlled separately. The functionalities of our latch design can be summarized as:

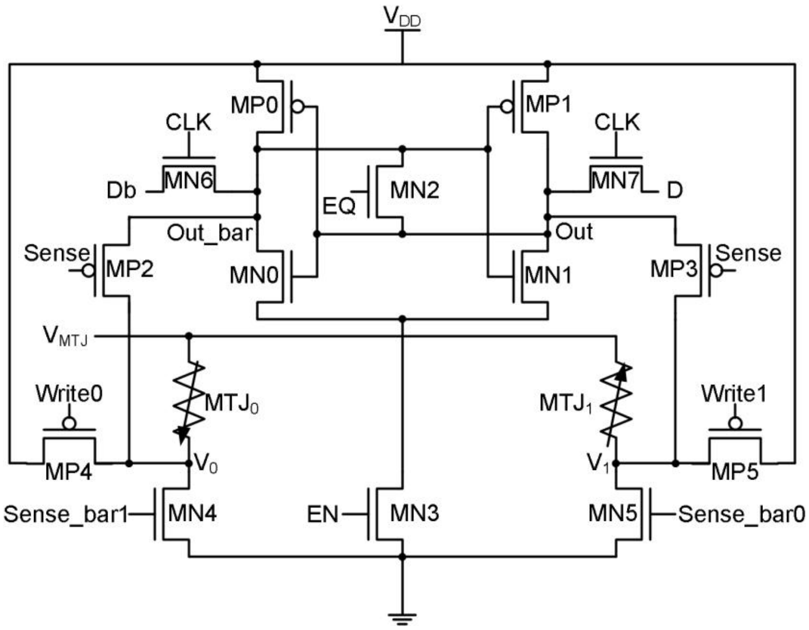


Figure 3.4. Our proposed nonvolatile latch design.

3.3.1 Normal latch mode

During the normal latch mode, the enable signal EN raises high to turn on MN3. The signal EQ is pulled down to turn off the equalizing transistor MN2. The transistors on the data backup and recovery paths, including MP2-MP5, MN4-5 are all turned off while the reference voltage V_{MTJ} is grounded. The design works as a conventional latch: data is written into the latch through MN6 and MN7, and stored at the outputs of the two inverters.

3.3.2 Data backup mode

Before the system enters the standby mode, our nonvolatile latch enters the data backup mode. The data stored in the latch will be differentially written into the MTJ pair. In our design, the reference layers of MTJ_0 and MTJ_1 are connected to MN4 and MN5, respectively. If the stored data in the latch is ‘1’, transistors MP4 and MN5 are turned on to allow a write current to pass through MTJ_0 and MTJ_1 in sequence. The MTJ_0 and MTJ_1 are then programmed to ‘1’ and ‘0’, respectively. If the stored data is ‘0’, MP5 and MN4 are turned on to program the MTJ_0 and MTJ_1 to ‘0’ and ‘1’, respectively. During the data backup mode, the input of V_{MTJ} keeps a high-impedance. After MTJ programming completes, the power supply can be safely removed. The cross section of the MTJ integration scheme and the write paths are shown in Fig. 3.5.

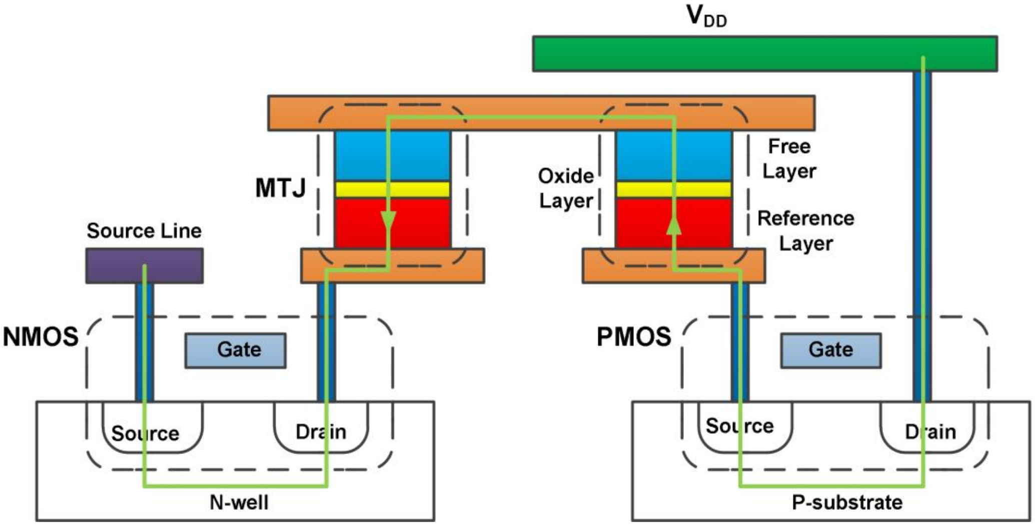


Figure 3.5. Cross section of the MTJ integration and write path.

Our latch design separates the MTJ write path from the normal latch operation circuit. The parasitic capacitances at the outputs are significantly reduced. The data backup time can be improved by simply sizing up MN4-5 and MP4-5 without degrading the normal latch operation performance.

3.3.3 Data recovery mode

When the system wakes up from the standby mode, the data in the latch will be recovered from resistance states of the MTJs. The data recovery mode includes two phases, namely, recovery signal generation and sensing:

In the recovery signal generation phase, signal EQ is raised high to turn on MN2 and equalize latch outputs Out and Out_bar. This operation minimize the impact of the driving competition at Out and Out_bar from the connected MOS transistors and the device mismatch of MP0, MP1, MN0 and MN1 in the cross-bar structure. Also, signal Sense and Sense_barX, (X = 0 or 1) are pulled down to ground and raised to VDD, respectively, to turn on transistors MN4, MN5, MP2, and MP3. After an appropriate reference voltage V_{MTJ} is applied, the conductive path formed by MN4 (MN5) and the MTJ_0 (MTJ_1) generates the signal V_0 (V_1), which is determined by the MTJ resistance state. Here MN4 and MN5 are designed to be identical and working in linear region. V_0 and V_1 are then transmitted into the cross-bar structure as the inputs.

In the sensing phase, MN2 is turned off. A voltage difference appears at Out and Out_bar due to the different resistance states of the MTJs. MN4, MN5, MP2 and MP3 are all turned off. The positive feedback of the cross-bar structure amplifies the initial voltage difference at Out and Out_bar to a full swing outputs, or logic one and zero.

There are two important design metrics to meet in our latch design: First, MN4 and MN5 must work in the linear region to generate the correct voltage levels of V_0 and V_1, say $V_0 < V_1$ when data = ‘1’ or $V_0 > V_1$ when data = ‘0’; Second, the voltage difference at Out and Out_bar at the beginning of the sensing phase must be large enough to conquer the device mismatch in the cross-bar structure. Here reusing MN4 and MN5 in both the reference signal generation path and the MTJ write path helps to reduce the layout area.

3.4 ERROR MECHANISMS AND ANALYSIS

The operation errors of a magnetic latch can be categorized into two types: persistent error and non-persistent error. Their differences were introduced in chapter 2.

3.4.1 Persistent errors in nonvolatile latch

The persistent errors are mainly incurred by the process variations of both CMOS transistors and MTJ. Transistor and MTJ device deviations causes the imbalance of the cross-bar structure, variations of the MTJ write current and the data recovery signals. Fig. 3.6 shows the error rates of the data recovery operations by considering the process variations of MOS transistors and MTJ under different sizes of MN4 and MN5. A larger transistor supplies high sensing current and a bigger voltage difference (ΔV_{out}) between Out and Out_bar. As a result, the error rate reduces. The simulation setup, such as the mean and the standard deviations of the MOS device and MTJ parameters, are summarized in Table 3.1. The error rates are significantly reduced when the ΔV_{out} increases or the device deviations decreases.

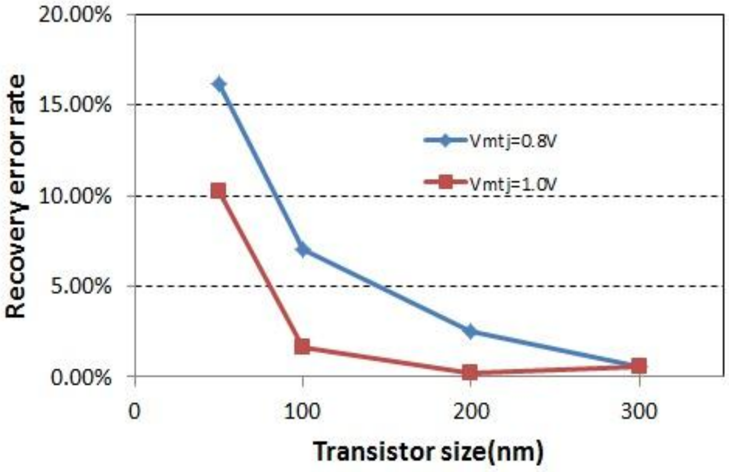


Figure 3.6. Data recovery error rate under various sizes of MN4 and MN5.

Table 3.1. Device parameters in our simulations.

Device Parameters		Mean	Standard Deviation
Transistor	Channel Length	45 nm	2.25 nm
	Channel Width	design dependent	2.25 nm
	Threshold Voltage	0.466 V [26]	σ_{Vth0}=30 mV
MTJ	MgO Thickness	2.2 nm	
	Shape Area	45x90 nm^2	
	Low Resistance	1000 Ω	
	High Resistance	2000 Ω	

3.4.2 Non-persistent errors in nonvolatile latch

The non-persistent errors are mainly introduced by the thermal fluctuation (TF) process during the data backup and recovery modes, i.e., the thermal component of internal energy (when MTJ switching >10ns) or the thermally activated initial angle of procession (when MTJ switching <10ns) [25]. In the data backup mode, TF induces the variation of the MTJ switching time. Write failure happens if the write current is removed before the MTJ completely switches. In the data recovery mode, the applied read current may flip the MTJ resistance accidently.

As the switching current increases, the MTJ switching time variation changes from Poisson distribution to Gaussian distribution. The ratio the standard deviation (σ) and the mean (μ) of MTJ switching time reduces first due to the increased impact of spin-torque on MTJ switching. Then the ratio ramps up at sub-10ns region because the thermally activated initial angle of procession starts to dominate, as shown in Fig. 3.7.

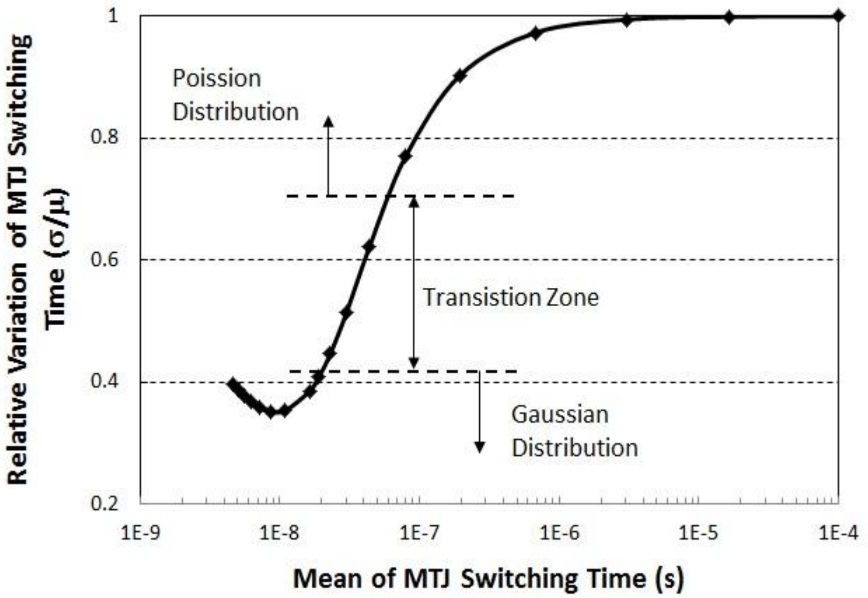


Figure 3.7. Variations of MTJ switching time due to thermal fluctuations.

3.5 EXPERIMENTAL RESULT AND DISCUSSION

3.5.1 Data recovery function

We designed the proposed nonvolatile latch with PTM 45nm technology [26] and conducted simulations with Spetre under Cadence design environment. Fig. 3.8 shows the Out (solid line) and Out_bar (dash line) waveforms during the data recovery mode. First, EQ is raised high to equalize Out and Out_bar. Then, Sense_bar0/1 switches to high to generate the data recovery signal V_0/V_1. After ΔV_{out} becomes stable, Sense_bar switches to low and the latch enters the sensing phase. If ΔV_{out} is large enough, cross-bar structure will drive Out and Out_bar to V_{DD} or ground.

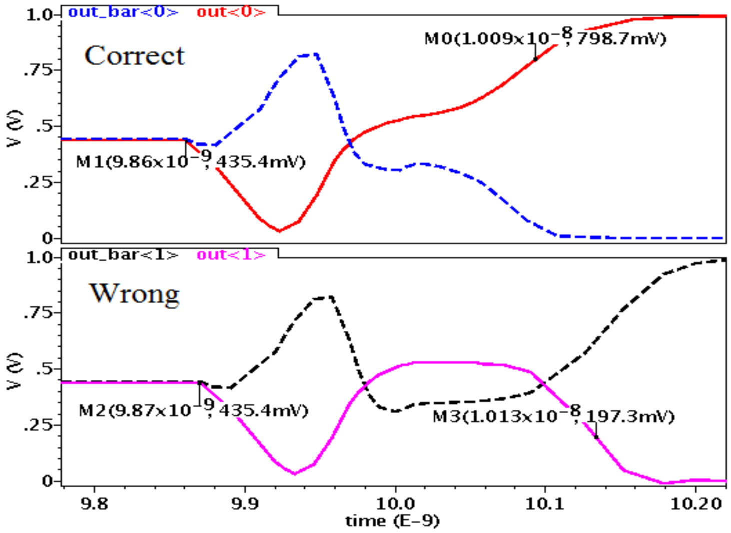


Figure 3.8. Timing waveforms of the data recovery operations.

Fig. 3.9 shows the data recovery latency at different device mismatch conditions of the cross-bar structure. The longest delay 230ps occurs at corner SFFS. The setup of Fast (F) and Slow (S) corners of MOS transistor is shown in Table 3.2. Fig. 3.10 shows the layout of our design by using FreePDK design rule [27]. The total area is 1.944 μm^2.

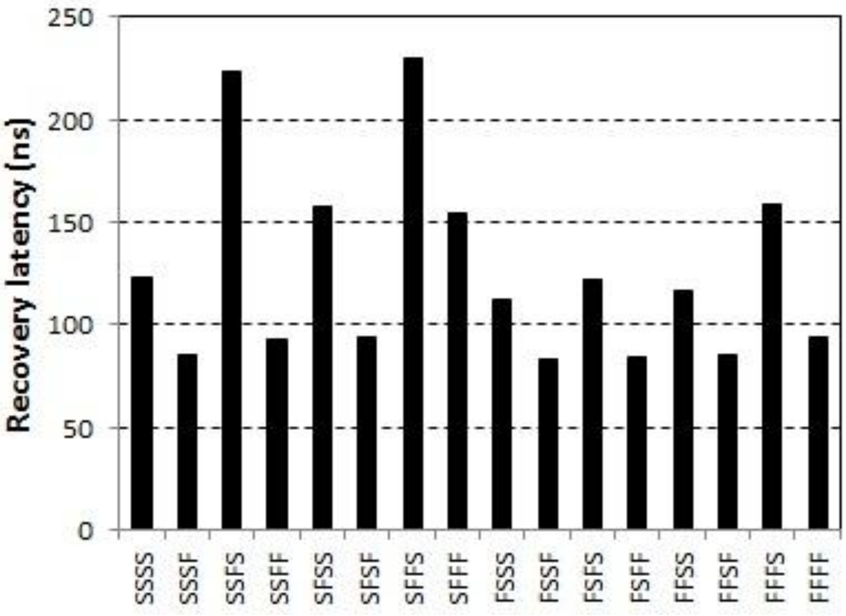


Figure 3.9. Data recovery latency under different device mismatch conditions. F: fast corner, S: slow corner. The device sequence in the corner representation is MN0-MN1-MP0-MP1.

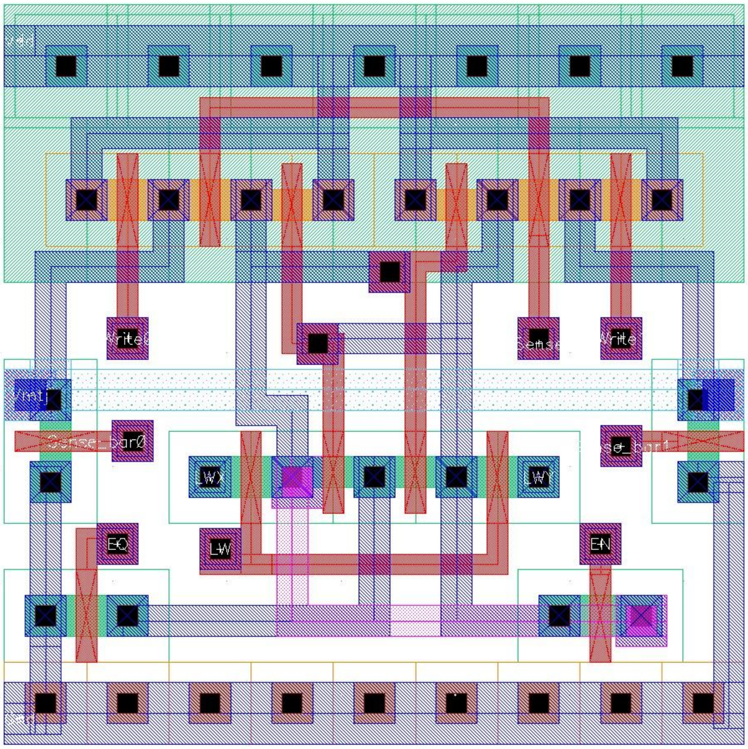


Figure 3.10. Layout of our proposed nonvolatile magnetic latch cell

Table 3.2. Corner cases of the device mismatch models for the transistors in cross-bar structure

Transistor Size	MN0/MN1		MP0/MP1	
	F	S	F	S
ΔVth(V)	-0.03	0.03	0.03	-0.03
Width(nm)	62.25	57.25	62.25	57.25
Length(nm)	42.75	47.25	42.75	47.25

3.5.2 Operation errors and design optimizations

When the data backup time requirement is fixed, the write failure can be minimized by increasing the sizes of MTJ write transistors due to the improved MTJ write current and the reduced influence of process variations. This conclusion is supported by Fig. 3.11, which shows that the write error rate decreases when raising the sizes of MN4-5 and MP4-5.

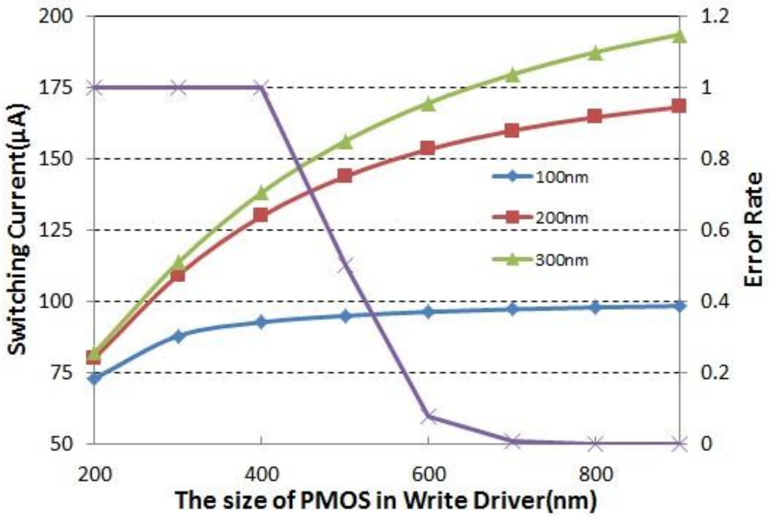


Figure 3.11. Data backup error rate under various write transistor size and write pulse width.

When the sensing currents through the MTJs increase, the voltage difference between V_0 and V_1 rises. The error rate of the data recovery mode decreases accordingly because of the increased ΔV_{out}. However, large sensing current also raises the probability to overwrite the MTJ to the undesired value. In our nonvolatile latch, the sensing current can be adjusted by sizing up MN4 and MN5 or increasing V_{MTJ}. In our design, the amplitude of the sensing current is about 250 µA at $W_{MN4,5}$ = 300nm, which is higher than that in the previous design [20]. However, the error rate still maintains at a low level close to zero because the current only flow through MTJ in a short period (95ps).

3.5.3 Design scalability

Compared to the previous designs, our design can work at low supply voltage (1.0V) while maintaining low error rate. For comparison, we simulate the data recovery error rate of the previous design [20] with the same technology. Even working at 2V, the error rates are still as high as 45%. If the design is working at the normal supply voltage of 1V, the error rates are 46.3%.

As technology keeps scaling down, the process variations will become more prominent and hence increase the minimum required ΔV_{out}. It can be achieved by applying a higher V_{MTJ} and/or continuing to increase MN4-5. In such cases, thick gate oxide devices may be used in the V_{MTJ} driver design for lifetime consideration. Our latch structure does not need to change.

3.6 CONCLUSION

We proposed a novel 1.0V 45nm nonvolatile magnetic latch for SoC power management technique. Our simulation shows that the data recovery time of the latch can be as low as 230ps. Both persistent and non-persistent error mechanisms are analyzed based on Monte-Carlo simulations. High robustness to the process variations is achieved in our design while maintaining low supply voltage, low power consumption and high operation speed.

4.0 STATISTICAL ANALYSIS FOR PREDICTING MEMORY LOSS

4.1 INTRODUCTION

Following scaling described by Moore's Law, the conventional memory technologies, i.e. SRAM, DRAM, and Flash memory, have achieved remarkable success in the applications of modern computing systems and portable electronics in the last several decades. However, when the technology scaling of the conventional memories enters 22nm process node and below [28-30] process variations significantly increase the fabrication and design costs.

Recently, a new concept called “Universal Memory” rises above the horizon. The expected characteristics of a universal memory include high-density (low-cost), high-speed (for both read and write operations), low-power (both access and standby powers), random-accessibility, non-volatility and unlimited endurance. These characteristics allow universal memory to meet the requirements of various applications: from a large, expensive supercomputer to a low-cost, ubiquitous, consumer handheld device. These memories can be excellent candidates for making “More than Moore” come true. Some promising candidates of universal memory include Phase-Change RAM (PCRAM) [31], Spin-Torque Transfer RAM (STT-RAM) [32, 33], and Resistive RAM (R-RAM) [34]. Table 4.1 lists some important qualitative features of these memory technologies predicted in ITRS 2009 [28], compared to the conventional SRAM.

Table 4.1. Comparison of Different Memory Technologies

Features	**SRAM**	**PCRAM**	**STT-RAM**	**RRAM**
Nonvolatility	No	Yes	Yes	Yes
Memory Cell Factor (F^2)	50-120	6-12	4-20	<1
Read Time (ns)	1	20-50	2-20	<50
Write/Erase Time (ns)	1	50-120	2-20	<100
Number of Rewrites	10^{16}	10^{10}	10^{15}	10^{15}
Power Consumption – Read/Write	Low	Low	Low	Low
Power Consumption – Other than R/W	Leakage Current	None	None	None

Although the emerging memories have demonstrated many promising characteristics overwhelming their technology ancestors, process variations continue to be the biggest challenge in the fabrication of these nanoscale devices. Moreover, these emerging memory technologies utilize new materials as storage devices, which bring in some new failure modes. Some of them have been summarized in Table 4.2. Therefore, statistical approaches for yield estimation and robust design are becoming more and more important.

Table 4.2. Sources of Variabilities

Sources	SRAM	PCRAM	STT-RAM	RRAM
Temperature	Y	Y	Y	Y
Geometry variations (LER, TF, etc)	Y	Y	Y	Y
Random dopant fluctuations	Big	Small	Small	Small
Radiation effects	Y	N	N	N
Aging effect	NBTI, HCI, etc	Resistance shifting	Resistance shifting	Resistance shifting

Moreover, the statistical analysis for memory-interacting logic has received little attention in the conventional SRAM design, especially analysis involving rare-failure estimation [35]. This is true from the functional behavior perspective as well as the performance perspective. However, the process variations of the peripheral logic can affect or even result in the design failure of the emerging memory technology, as we shall show in the paper. Thus, we need to capture not only average logic delay distributions but also possible design failures, especially when we want to guarantee the yield for millions of chips. Also the methodology is well suited for optimizing logic and memory elements. Furthermore, we must analyze the yield of the memory design in situ with the peripheral logic, raising the need for simultaneous statistical analysis of the memory/logic unit.

In this chapter, we will mainly focus on STT-RAM. The basic component of magnetic random access memory (MRAM) is magnetic tunneling junction (MTJ). Data storage is realized by switching the resistance of MTJ between high- and low-resistance states [36]. MRAM features non-volatility, fast writing/reading speed (<10ns), almost unlimited programming endurance (>10^{15} cycles) and zero standby power.

In conventional MRAM design (known as “toggle-mode”), MTJ resistance is changed by using the current induced magnetic field to switch the magnetization of MTJ. When the size of MTJ scales, the amplitude of the required magnetic field is increased correspondingly. The high write power consumption severely limits the scaling of conventional MRAM. Recently, a new write mechanism based on spin polarization current induced magnetization switching, is introduced to MRAM design. This new STTRAM design is believed to have a better scalability than conventional MRAM. Various designs of STT-RAM were proposed by both industry and academia in the past several years [33].

Previously Li et al. [37] had discussed the variation sources of MTJ and proposed 2T1J STT-RAM design for yield enhancement. And Chen and Sun et al. [38, 39] proposed self-reference schemes to overcome the read failures in STT-RAM design. However, these work mainly focused on STT-RAM cells. There is lack of a statistical analysis flow for overall STT-RAM system design including both memory cells and peripheral circuitry.

In this chapter, we use STT-RAM as example to discuss the implication of statistical analysis to the emerging nonvolatile memory design. We extended a mixture importance sampling methodology, a fast Monte Carlo technique [35], to STT-RAM yield analysis. The methodology not only targets the memory elements, but also builds a holistic yield analysis methodology, which goes beyond memory to peripheral logic.

The rest of this chapter is organized as follows: Section 4.2 provides the preliminaries of MTJ and STT-RAM; Section 4.3 discuss the impact of process variation in STT-RAM read operation; Section 4.4 explains the mixture importance sampling methodology used for STT-RAM read failure analysis; Section 4.5 concludes the paper.

4.2 PRELIMINARIES OF STT-RAM

4.2.1 The basic of MTJ

MTJ – the data storage element of STT-RAM - includes two ferromagnetic layers and one oxide barrier layer, e.g., MgO. MTJ resistance is determined by the relative magnetization directions of the two ferromagnetic layers: when the magnetization directions are anti-parallel (parallel), MTJ is in high- (low-) resistance state, as shown in Figure 3.1. In STT-RAM, the magnetization direction of one ferromagnetic layer (called "reference layer") is fixed by coupling to a pinned magnetization layer; the magnetization direction of the other ferromagnetic layer (called "free layer") is changed by passing a driving current polarized by reference layer [33].

A typical R-I sweep curve of an MgO-based MTJ is shown in Figure 4.1 [38]. Application of a positive voltage on point B in Figure 3.1, the magnetization direction of free layer rotates to the opposite direction of reference layer. MTJ resistance switches from low to high. On the other hand with a positive voltage on point A, the magnetization direction of free layer rotates to the same direction of reference layer. MTJ resistance switches from high to low.

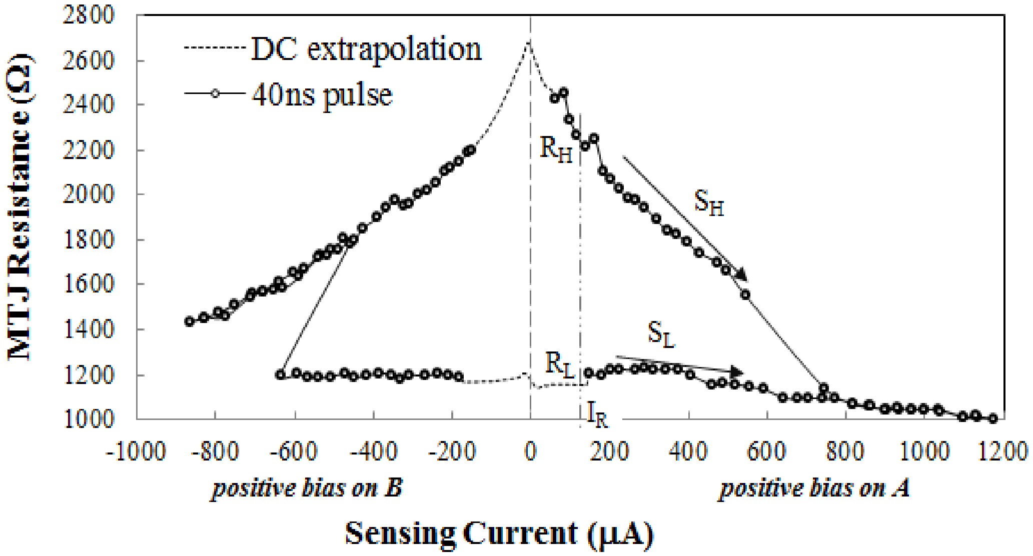


Figure 4.1. The measure static R-I curve of an MgO-based MTJ.

An important parameter of MTJ is Tunneling Magneto Resistance Ratio (TMR), which is defined as

$$TMR = \frac{R_H - R_L}{R_L}. \quad (4.1)$$

Here R_L and R_H denote the low and the high MTJ resistances, respectively. As shown in Figure 4.1, R_H and R_L (and hence TMR) actually depends on the magnitude of read current.

4.2.2 STT-RAM cell

Because of its simplicity, one-transistor-one-MTJ (or 1T1J) structure [33], where one MTJ is connected to one NMOS transistor in series, becomes the most popular design of STT-RAM. As shown in Figure 2.1(a), we usually call interconnects connected to MTJ, to the source/drain and to the gate of NMOS transistor as bit-line (BL), source-line (SL) and word-line (WL), respectively. MTJ is modeled as a current-dependent resistor in the equivalent circuit schematic [40], as shown in Figure 2.1(b). The direction of the switching current of MTJ is polarized by the different biasing on BL and SL.

Fig. 4.2 depicts a conventional voltage sensing scheme for STT-RAM design [38]. Read current I_R is sent to the STT-RAM cell and generates the BL voltage as:

$$\begin{aligned}&\text{if MTJis in low resistance state}: V_{\mathrm{BL,L}} = I_{\mathrm{R}} \cdot (R_{\mathrm{L}} + R_{\mathrm{TR}}) \quad \text{or}\\ &\text{if MTJis in high resistance state}: V_{\mathrm{BL,H}} = I_{\mathrm{R}} \cdot (R_{\mathrm{H}} + R_{\mathrm{TR}}).\end{aligned} \tag{4.2}$$

Here R_L and R_H are the low and the high MTJ resistance at read current I_R, respectively. R_{TR} is the resistance of NMOS transistor. $V_{BL,L}$ and $V_{BL,H}$ are the BL voltage when the MTJ is at the low and the high resistance state, respectively. By comparing the BL voltage to a reference voltage V_{REF} between $V_{BL,L}$ and $V_{BL,H}$, the MTJ resistance state can be readout. If a V_{REF} is shared by multiple STT-RAM bits, it needs to satisfy:

$$Max(V_{\mathrm{BL,L}}) < V_{\mathrm{REF}} < Min(V_{\mathrm{BL,H}}). \tag{4.3}$$

Here Max($V_{BL,L}$) and Min($V_{BL,H}$) denote the maximal $V_{BL,L}$ and the minimal $V_{BL,H}$ generated by all involved STT-RAM bits, respectively. Unfortunately, Max($V_{BL,L}$) < Min($V_{BL,H}$) may not be always true when the bit-to-bit variation of MTJ resistance is large.

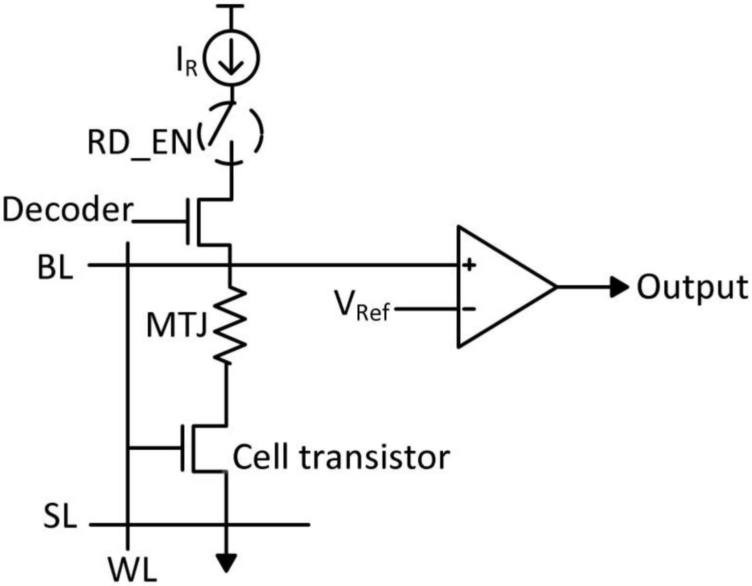


Figure 4.2. Read-out scheme of STT-RAM.

4.3 PROCESS VARIATIONS IN STT-RAM READ OPERATION

4.3.1 Process variations in STT-RAM design

The main electrical properties of an MTJ that affect STT-RAM read operations are: R_{L0}-low resistance at a close-to-zero read current, R_{H0}-high resistance at a close-to-zero read current, S_L-low state roll-off slope, and S_H-high state roll-off slope, as shown in Figure 4.1. The MTJ resistances $R_L(I_R)$ and $R_H(I_R)$ at read current I_R can be expressed as

$$R_L(I_R) = R_{L0} - S_L \cdot I_R \text{, and } R_H(I_R) = R_{H0} - S_H \cdot I_R \text{.} \qquad (4.4)$$

The variations of MTJ resistances R_{L0} and R_{H0} are mainly determined by the thickness of MgO layer and the geometrical size of MTJ. Usually the uniformity of MTJ stack is evaluated by resistance-area product (RA), and cross-sectional area (A). For a given MTJ stack, R_{L0} and R_{H0} can be calculated by:

$$R_{L0} = \frac{(RA)_{L0}}{A}\text{, and } R_{H0} = \frac{(RA)_{H0}}{A}\,. \qquad (4.5)$$

Here, $(RA)_{H0}$ and $(RA)_{L0}$ represent the RA's at high and low resistance states and measured at a close-to-zero read current, respectively.

A high TMR (>100%) can be achieved by an MgO-based MTJ because the quantum tunneling selection rule prohibits the transition of minority spins. However, the actual TMR and RA are also affected by some quantum effects caused by material interfacial state, as well as some lattice defects, lattice dislocation and discrenation because of the difficulty to control thin film growth process like annealing time and sputtering uniformity. Usually the variation of $(RA)_H$ is larger than that of $(RA)_L$ due to the interaction of the intrinsic quantum tunneling process and the extrinsic scattering process. Similarly, the asymmetric S_H and S_L of an MTJ depends upon the micro-structure of the interfaces separating the two electrodes. For example, a ballistic electronic and spin transport model was proposed to explain the state roll-off asymmetry of MTJs in [41]. The variations of the state roll-off slope, however, can be addressed by taking into account both elastic and inelastic tunneling processes.

Table 4.3 summarizes the statistical data of the MTJ process and electrical parameters adopted in our work [38]. We assume that $(RA)_H$, $(RA)_L$, A, S_L and S_H all follow Gaussian distribution. The means (μ) and the standard deviations (std. dev. σ) of $(RA)_L$ and $(RA)_H$ are estimated based on the measured data presented in [37]. The std. dev. of A is set to 5%, which is the same as the one reported in [42]. Since there is no any public data on roll-off variations available, we assume that the std. dev. of S_H is 10%, which is pessimistic enough to cover the normal range.

Table 4.3. Electrical parameters of MTJ

Parameters	**Mean μ**	**Std. Dev. σ**	**σ/μ**
τ	10 Å	0.5 Å	5%
A	$90\times180\text{nm}^2$	810nm^2	5%
$(RA)_{L0}$	$20\ \Omega\cdot\mu\text{m}^2$	$1.6\ \Omega\cdot\mu\text{m}^2$	7.8%
$R_{L,0}$	1230 Ω	114.4 Ω	9.3%
$R_{H,0}$	2650 Ω	273.0 Ω	10.3%
S_L	5×10^4 Ω/A	5×10^3 Ω/A	10%
S_H	3×10^6 Ω/A	3×10^5 Ω/A	10%

The variations of the electrical and geometry parameters of transistor, i.e., threshold voltage V_{th} and transistor dimension ratio W/L are also considered. According to [37] V_{th} follow Gaussian distributions with a std. dev. σ = 8.2%. For simplicity we assume both V_{th} and W/L are lumped into the Vth variability, which follows Gaussian distribution with a std. dev. σ = 10% for the cell CMOS device.

4.3.2 Impact of peripheral circuit

Figure 4.3 shows a basic cross section of the read path in STT-RAM design. It consists of a cell and a fully loaded bitline segment consisting of two smaller local bitline segments. Each local bitline segment has 32 cells above and below the local bit-select circuit.

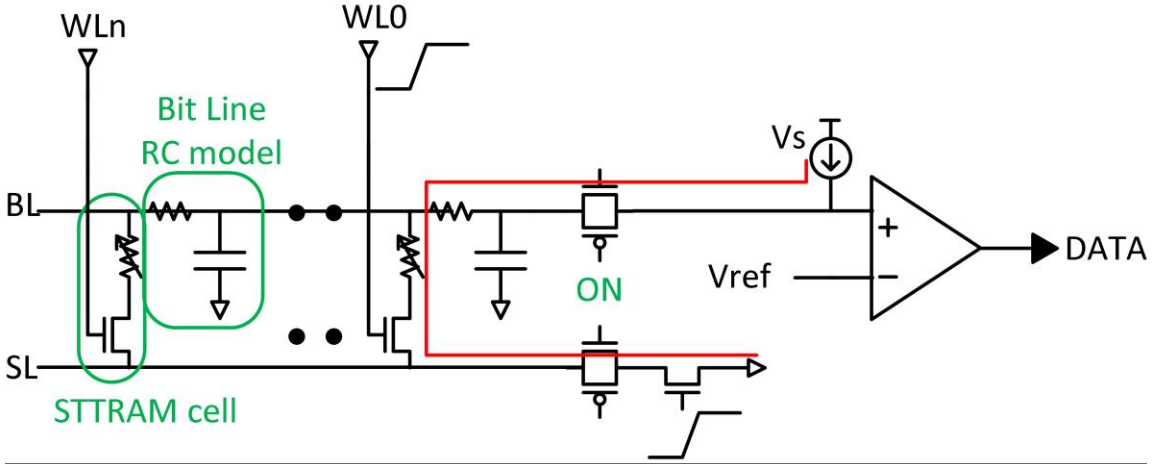


Figure 4.3. Cross section of a local bit-select and evaluate circuit of STT-RAM.

In a read path, the read current indeed goes through at least two multiplexers, one STT-RAM cell, and an NMOS transistor. The CMOS process variations can result in the deviation of read current, and hence, affect the read MTJ resistance as shown in Figure 4.1. This means, on top of the variations of MTJ characteristics, the peripheral circuit introduces one more local variation, which also affect the BL voltage V_{BL}.

Moreover, the changing of read current I_R and the MTJ resistance has an opposite impact on the BL voltage V_{BL}: when I_R increases, the MTJ resistance decreases. From this perspective, increasing I_R may not always help improve sense margin and reduce read failure.

4.4 STATISTICAL ANALYSIS IN MEMORY YIELD DESIGN

4.4.1 Traditional Monte Carlo simulations

Monte Carlo method has been the most popular method to estimate design yield and fail probabilities. The increasing demand of density and chip-yield requirements, however, raises stringent requirements on the fail probability of less than 1-per-million parts. This in turn requires an extremely large number of Monte Carlo simulations to achieve good confidence and accuracy. Equation (4.6) represents the number of Monte Carlo simulations needed to accurately estimate P_f= Prob(x > z_0) with the 95% confidence interval and error of estimate criteria α=10%; x follows a standard normal distribution. The number of samples needed for estimating low failure probabilities z_0>4 exceeds $1e^6$ samples and is not practical.

$$N \approx \frac{4}{\alpha^2} \bullet \frac{(1-P_f)}{P_f} \tag{4.6}$$

4.4.2 Mixture importance sampling

Variance reduction methods are intended to reduce the error in the estimate and hence improve the efficiency of the statistical simulations for a given number of samples. Importance sampling [43] is one form of variance reduction that enables estimating excessively low fail probabilities and is suitable to address the memory design yield problem. The method relies on distorting the (natural) Monte Carlo sampling function, to produce more samples in the important region(s). It is based on the following fact.

$$E_{p(x)}[\theta] = E_{g(x)}[\theta \bullet \frac{p(x)}{g(x)}] \tag{4.7}$$

where $E_p[\Theta]$ is the expected value of Θ with respect to the sampling function p, g(x) is the distorted sampling function, and p(x) is the natural distribution. The method is theoretically sound, and with the proper choice of g(x), we are able to obtain accurate results with relatively small number of simulations.

4.4.3 MixIS in STT-RAM design

We apply the methodology [36] to the STT_RAM cell read yield analysis. Both R_L and R_H conditions are considered as illustrated in Figures 4.4 and 4.5 respectively. The experiments represent different combinations of the sources of variability and are labeled as follows.

a. Cell (RES): variability considered only in the magnetic resistor of the cell to be read.

b. Cell (device): variability considered only in the cmos device of the cell to be read.

c. Cell (RES+device): variability considered in both the magnetic resistor and the cmos device of the cell to be read.

d. Cell (RES+device) +Ref Cell (device): variability considered in both the magnetic resistor and the cmos device of the cell to be read. Also the reference cell cmos device is subject to variability.

e. Cell (RES+device) +Ref Cell (RES+device): variability is similar to (d). Also the reference cell magnetic resistor is subject to variability.

f. Cell (RES+device) +Ref Cell (RES+device) +SA: variability is similar to (e). Also we assume variability in very small sense amplifier devices.

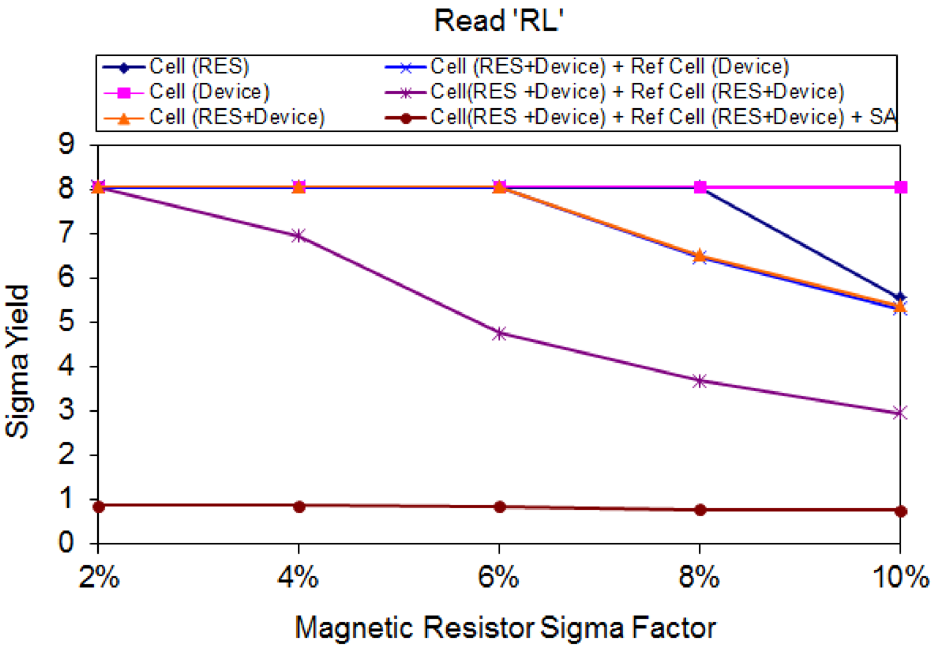


Figure 4.4. Read Yield analysis for RL case.

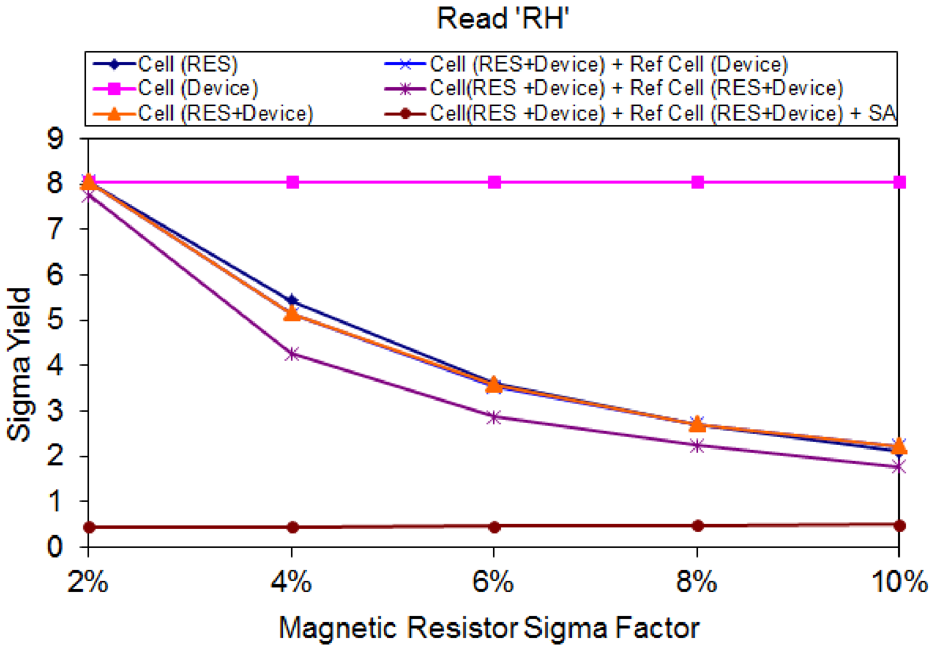


Figure 4.5. Read Yield analysis for RH case.

The yield is then studied in sensitization to the magnetic resistor standard deviation. This helps identify the rate of yield improvement versus process improvement. Note that magnetic resistor variability accounts for both the resistor intrinsic variability and the sensitivity to device current variability factors. Yield is reported based on sigma numbers. For example a 5-sigma yield number is equivalent to fail probability of (P(x>5)), where x is distributed according to a standard normal distribution. We do not compute yields beyond 8-sigma, hence the saturation trend in the estimated yield).

We note that for the ‘RL’ the design is sensitive to both variability in the magnetic resistance of the accessed device and the reference device. CMOS device variability in the accessed and STTRAM cells magnifies the variability in the resistor but is not sufficient alone to impact design yield. For the ‘RH’ case the magnetic resistor variability of the accessed cell is dominant (ignoring sense amp). The cell CMOS device variability affects yield a bit in the high yield region. The yield is tolerable above the 5-sigma range if the magnetic resistor variability factor drops below 4%. Finally, the design is very sensitive to variability in the sense amp.

4.5 CONCLUSION

In this chapter, we used STT-RAM as examples to illustrate the implication of resistance-based nonvolatile memory for yield analysis and the corresponding design consideration. Not only the memory elements itself, but also the peripheral logic, should be comprehensively considered in the yield analysis. A universal statistical methodology presented here to predict memory loss and enable robust design practices is highly desired by emerging nonvolatile memory design.

5.0 CONCLUSIONS AND FUTURE WORK

This inconvenience of unable to turn on/off current (embedded) computer system has limited user behavior and wasted an enormous amount of power to keep these devices on while doing nothing. Non-volatile computing is an attractive solution, as non-volatile storage and logic circuit can memorize the initial state before system sleeps and restore system instantly in relatively low power consumption. By combining the flexibility of MOS logic and the non-volatility of spintronic devices, spin-MOS logic and storage circuitry offer a promising approach to implement a highly integrated, power-efficient, and nonvolatile computing and storage systems.

This thesis presented non-volatile logic and storage circuitry design in the perspective of robustness. The persistent and non-persistent errors in Spin-MOS circuitry are defined and analyzed: the former mainly comes from process variations, and the later one is resulted by thermal fluctuations and read disturbance. This work quantitatively investigates the impacts of these variations and fluctuations on the operations of spin-MOS circuitry. A mixture importance sampling methodology is applied to enable yield-driven design and extended beyond memories to peripheral circuits and logic blocks.

On top of it, a novel 1.0V 45nm nonvolatile magnetic latch for SoC power management technique is proposed. By running Monte-Carlo simulation, high robustness to both persistent and non-persistent error mechanisms is proved in our design while maintaining low supply voltage, low power consumption and high operation speed.

The field of non-volatile storage and logic is still an emerging one, and as such there are several areas in which more work can be done, from variation modeling, to the design of circuits and the peripheral supporting blocks to post-processing techniques.

BIBLIOGRAPHY

1. Non-volatile memory. http://en.wikipedia.org/wiki/Non-volatile_memory

2. W. Robinett, M. Pickett, J. Borghetti, Q. F. Xia, G. S. Snider, G. Medeiros-Ribeiro, and R. S. Williams, “A memristor-based nonvolatile latch circuit,” Nanotechnology, 21, 235203, 2010.

3. Y. Lakys, W. Zhao, J.-O. Klein, and C. Chappert, “Low power, high reliability magnetic flip-flop,” Electron. Lett., vol. 46, pp.1493-1494, 2010.

4. W. Zhao, E. Belhaire, and C. Chappert, “Spin-MTJ based Non-volatile Flip-Flop,” Proceedings of the 7th IEEE International Conference on Nanotechnology, pp. 399-402, Aug. 2007.

5. K. Kuhn, et al., “Managing process variation in Intel’s 45 nm CMOS technology,” Intel Technology Journal, Jun. 2008.

6. Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L. Wang and Y. Huai, “Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory,” J Phys : Condensed Matter, 19, 2007, 165209.

7. L. Berger, “Emission of spin waves by a magnetic multilayer traversed by a current,” Phys. Rev. B., 54, 1996, pp. 9353-9358.

8. T. L. Gilbert, “A Lagrangian Formulation of the Gyromagnetic Equation of the Magnetization Field,” Phys. Rev., 100, 1955, 1243.

9. X. Wang, Y. Chen and T. Zhang, “Magnetization Switching in Spin Torque Random Access Memory: Challenges and Opportunities”, CMOS Processors and Memories, Springer, 2010.

10. Y. Chen, X. Wang, H. Li, H. Liu, D. V. Dimitrov, “Design Margin Exploration of Spin-Torque Transfer RAM (SPRAM),” International Symposium on Quality Electronic Design, pp. 684-690, 2008.

11. N. Sakimura, et al, “Nonvolatile Magnetic Flip-Flop for Standby-power-free SoCs,” IEEE Custom Integrated Circuits Conf., Sep. 2008, pp.355-358.

12. X. Guo, E. Ipek, and T. Soyata, “Resistive Computation: Avoiding the Power Wall with Low-leakage, STT-AM based Computing,” 37th Annual Int’l Symp. On Computer Architecture, Jun. 2010, pp. 371-382.

13. T. C. Chen, "Overcoming Research Challenges for CMOS Scaling: Industry Directions", 8th Int’l. Conf. on Solid-State and Integrated Circuit Technology, 2006, pp. 4-7.

14. S. Lu, “Microprocessor Memory Circuits,” Presentation in Workshop on Technology-Architecture Interaction: Emerging Technologies and their Impact on Computer Architecture, 2010 (Held in conjunction with 43rd Annual IEEE/ACM Int’l Symp. on Microarchitecture).

15. 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,” IEEE Int’l Electron Devices Meeting, pp. 1 –4, Dec. 2009.

16. Y. Higo, et al., “Thermal Activation Effect on Spin Transfer Switching in Magnetic Tunnel Junctions,” Appl. Phys. Lett., 87, 082502 (2005).

17. M. Hosomi, et al., “A Novel Nonvolatile Memory with Spin Torque Transfer Magnetization Switching: Spin-RAM,” IEEE Int’l Electron Device Meeting, Dec. 2005, pp. 459–462.

18. 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,” ACM/IEEE Int’l Symp. on Low Power Electronics and Design, 2010, pp. 1-6.

19. D. Lammers, TI Moves Ahead with 65-nm Chips by Next Year, EE Times, Mar. 22, 2004.

20. N. Sakimura, T. Sugibayashi, R. Nebashi, and N. Kasai, “Nonvolatile magnetic flip-flop for standby-power-free SoCs,” IEEE J. Solid- State Circuits, vol. 44, no. 8, pp. 861–869, Aug. 2009.

21. W. C. Black, Jr., and B. Das, “Programmable logic using giantmagnetoresistance and spin-dependent tunneling devices,” J. Appl. Phys., vol. 87, No. 9, pp. 6674-6679, May. 2000.

22. W. Zhao, E. Belhaire, C. Chappert, F. Jacquet, and P. Mazoyer, “New non-volatile logic based on spin-MTJ,” Phys. Stat. Sol. (a) 205, No. 6, pp. 1373-1377, May. 2008.

23. W. Zhao, E. Belhaire, C. Chappert, and P. Mazoyer, “Power and area optimization for run-time reconfiguration SOPC based on MRAM,” IEEE Trans. Magn., vol. 45, pp. 776–780, 2009.

24. M. Hosomi, et al., “A Novel Nonvolatile Memory with Spin Torque Transfer Magnetization Switching: Spin-RAM,” International Electron Device Meeting Tech. Dig., 2005, pp. 473-476.

25. Z. Diao, et al., “Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory,” J. Phys: Condensed Matter, 19, 2007, 165209.

26. Predictive Technology Model (PTM). http://www.eas.asu.edu/~ptm/.

27. FreePDK45. http://www.eda.ncsu.edu/wiki/FreePDK45:Contents.

28. International Technology Roadmap for Semiconductor, 2007. http://www.itrs.net/.

29. K. Kinam and J. Gitae, “Memory technologies for sub-40nm node,” International Electron Devices Meeting (IEDM), pages 27-30, 2007.

30. William J. Gallagher, “Emerging Nonvolatile Magnetic Memory Technologies,” IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT), Nov.2010, pp. 1073-1076.

31. T. Bedeschi, et al, “A bipolar-selected phase change memory featuring multi-level cell storage, JSSC, vol. 44, no. 1, pp. 217-227, Feb. 2008.

32. R. Beach, et al, “A Statistical Study of Magnetic Tunnel Junctions for High-Density Spin Torque Transfer-MRAM (STT-MRAM),” IEEE International Electron Devices Meeting (IEDM), Dec. 2008, pp. 1-4.

33. M. Hosomi, et al, “A Novel Nonvolatile Memory with Spin Torque Transfer Magnetization Switching: Spin-RAM,” International Electron Devices Meeting (IEDM), pp. 459-462, 2005.

34. G. W. Burr, et al, “Overview of candidate device technologies for storage-class memory,” IBM Journal Research and Device, vol. 52, no. 4/5, pp. 449-464, 2008.

35. S. Tehrani, et al, “Magnetoresistive Random Access Memory Using Magnetic Tunnel Junctions,” Proc. IEEE, pp. 703–714, May 2003.

36. R. Joshi, R. Kanj A. R. Pelella, A. Tuminaro, and Y. Chan, “The Dawn of Predictive Chip Yield Design: Along and Beyond the Memory Lane,” IEEE Design & Test of Computers, vol. 27, no. 6, pp. 36-45, 2010.

37. Z. Sun, et al, “Variation Tolerant Sensing Scheme of Spin-Transfer Torque Memory for Yield Improvement,” International Conference on Computer-Aided Design (ICCAD), Nov. 2010, pp. 432-437.

38. Y. Chen, et al, “Design Margin Exploration of Spin-Torque Transfer RAM (SPRAM)”, Proc. Intl. Symp. On Quality Electronic Design (ISQED), 2008, pp. 684-690.

39. X. Wang, et al, “Quantum Transport and Magnetization Dynamics Simulation on Intrinsic Spin Torque Switching Asymmetry”, Phys. Rev. B., vol. 79, 104408, 2009.

40. J. Li, H. Liu, S. Salahuddin, and K. Roy, "Variation-tolerant Spin-Torque Transfer (STT) MRAM array for yield enhancement," in Custom Integrated Circuits Conference, 2008, pp. 193-196

41. H.Raymond and Ping.Wang, “Variability in Sub-100nm SRAM Designs,” International Conference on Computer-Aided Design (ICCAD), pp. 347-352, 2004.

42. Y. Chen, et al, “Combined Magnetic- and Circuit-level Enhancements for the Nondestructive Self-Reference Scheme of STT-RAM,” ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), pp. 1-6, 2010.

43. R. Kanj, R. Joshi, and S. Nassif, “Mixture Importance Sampling and Its Application to the Analysis of SRAM Designs in the Presence of Rare Failure Events”, Design Automation Conference (DAC), pp. 69-72, 2006.