Design and development of low-power and reliable logic circuits based on spin-transfer torque magnetic tunnel junctions

Erya Deng

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **NANO ELECTRONIQUE ET NANO TECHNOLOGIES**

Arrêté ministériel : 25 mai 2016

Présentée par

Erya DENG

Thèse dirigée par **Lorena ANGHEL** et
co-dirigée par **Guillaume PRENAT**

préparée au sein du **Laboratoire Techniques de l'Informatique et de la Microélectronique pour l'Architecture des systèmes intégrés**
dans **l'École Doctorale Electronique, Electrotechnique, Automatique, Traitement du signal (EEATS)**

Conception et développement de circuits logiques de faible consommation et fiables basés sur des jonctions tunnel magnétiques à écriture par transfert de spin

Thèse soutenue publiquement le **10 février 2017**,
devant le jury composé de :

M. Jean-Michel PORTAL
Professeur, Université d'Aix-Marseille, Président et Examinateur
M. Lionel TORRES
Professeur, Université Montpellier 2, Rapporteur
M. Ian O'CONNOR
Professeur, École Centrale de Lyon, Rapporteur
M. Jacques-Olivier KLEIN
Professeur, Université Paris-Sud, Examinateur
M. Weisheng ZHAO
Professeur, Université Beihang, Invité
Mme. Lorena ANGHEL
Professeur, Institut polytechnique de Grenoble, Directrice de thèse
M. Guillaume PRENAT
Chargé de Recherche, CEA, Co-encadrant de thèse

Acknowledgments

I would like to thank all the people who supported and helped me during my years in TIMA laboratory and IEF laboratory.

I would like to thank my supervisor Prof. Lorena Anghel, professor in Grenoble INP, and co-supervisor Guillaume Prenat, researcher in SPINTEC laboratory, for their patience, guidance and encouragement. They helped me to advance my research in the last year and went through the manuscript of my thesis very carefully regardless of their busy schedule.

I sincerely appreciate Prof. Weisheng Zhao, professor in Beihang University, and Prof. Bernard DIENY, chief scientist in SPINTEC, for giving me the opportunity to be a part of the DIPMEM project during the first two years. Prof. Weisheng Zhao taught me how to be an independent researcher and inspired me when I met the research bottleneck or difficulties.

I'm grateful to all the members of the jury for their precious time spent on my thesis. I'm thankful to the rapporteurs, Prof. Lionel Torres and Prof. Ian O'connor, for reviewing and writing reports for the manuscript of my thesis. I would also like to thank the examiner and president, Jacques-Olivier Klein and Jean-Michel Portal, for their evaluation and questions.

I would like to thank my teachers and colleagues: Prof. Jacques-Olivier Klein, professor in University Paris-Sud, Yue Zhang, Wang Kang, Zhaohao Wang, You Wang, Djaafar Chabi and Chenxing Deng for many enlightening discussions and suggestions. It was an excellent experience to work with such a hard-working and talent group.

I would like to thank administrators in TIMA laboratory and IEF laboratory: Mme. Marie-Pierre Caron, Mme. Sylvie Lamour, Mme. Laurence Ben Tito, Mme. Anne-Laure Fourneret and M. Alexandre Chagoya… who helped a lot during my study in both laboratories.

I wish to express my deep gratitude to my parents for supporting me and being my strong backing throughout my study.

Finally, I want to thank China Scholarship Council (CSC) for the financial support.

Abstract

Title: Design and development of low-power and reliable logic circuits based on spin-transfer torque magnetic tunnel junctions

With the shrinking of CMOS (complementary metal oxide semi-conductor) technology, static and dynamic power increase dramatically and indeed has become one of the main challenges due to the increasing leakage current and long transfer distance between memory and logic chips. In the past years, spintronics devices, such as spin transfer torque based magnetic tunnel junction (STT-MTJ), are widely investigated to overcome the static power issue thanks to their non-volatility. Logic-in-memory (LIM) architecture allows spintronics devices to be fabricated over the CMOS circuit plane, thereby reducing the transfer latency and the dynamic power dissipation. This thesis focuses on the design of hybrid MTJ/CMOS logic circuits and memories for low-power computing system.

By using a compact MTJ model and the STMicroelectronics design-kit for regular CMOS design, we investigate the hybrid MTJ/CMOS circuits for single-bit and multi-bit reading and writing. Optimization methods are also introduced to improve the reliability, which is extremely important for logic circuits where error correction blocks cannot be easily embedded without sacrificing their performances or adding extra area to the circuit. We extend the application of multi-context hybrid MTJ/CMOS structure to the memory design. Magnetic random access memory (MRAM) with simple peripheral circuits is designed.

Based on the LIM concept, non-volatile logic/arithmetic circuits are designed to integrate MTJs not only as storage elements but also as logic operands. First, we design and theoretically analyze the non-volatile logic gates (NVLGs) including NOT, AND, OR and XOR. Then, 1-bit and 8-bit non-volatile full-adders (NVFAs), the basic elements for arithmetic operations, are proposed and compared with the traditional CMOS-based full-adder. The effect of CMOS transistor sizing and the MTJ parameters on the performances of NVFA is studied. Furthermore, we optimize the NVFA from two levels. From the structure-level, an ultra-high reliability voltage-mode sensing circuit is used to store the operand of NVFA. From the device-level, we propose 3-terminal MTJ switched by spin-Hall-assisted STT to replace the 2-terminal MTJ because of its smaller writing time and power consumption.

Finally, non-volatile content addressable memory (NVCAM) is proposed. Two magnetic decoders aim at selecting a word line to be read or written and saving the corresponding search location in non-volatile state.

Keywords: spintronics, spin transfer torque, magnetic tunnel junction, hybrid MTJ/CMOS circuits, non-volatile logic/arithmetic circuits

Résumé

Avec la miniaturisation des dispositifs dans les nœuds avancés en technologie CMOS, la consommation statique et dynamique augmente spectaculairement. Cette augmentation des courants de fuite, ainsi que les longues distances d'interconnexions entre les mémoires et les parties logiques des circuits sont les principaux problèmes limitant cette miniaturisation. Au cours des dernières années, les dispositifs de spintronique, tels que la jonction tunnel magnétique (JTM), notamment dans sa version à écriture par transfert de spin, sont largement étudiés pour résoudre le problème de la consommation statique grâce à leur non-volatilité combinée à leur vitesse d'écriture et leur endurance. L'architecture logique-en-mémoire (LEM) hybride consiste à fabriquer les dispositifs de spintronique au-dessus des circuits CMOS, réduisant le temps de transfert et la puissance dynamique. Cette thèse vise à concevoir des circuits logiques et mémoires pour SoC de faible consommation, en combinant les technologies JTM et CMOS.

En utilisant un modèle compact de JTM et le design-kit CMOS de STMicroelectronics, nous étudions des circuits hybrides JTM/CMOS 1-bit et multi-bits, en particulier les opérations de lecture et d'écriture. Des méthodes d'optimisation sont également introduites pour améliorer la fiabilité, ce qui est extrêmement important pour les circuits logiques où les blocs de correction d'erreur ne peuvent pas être facilement intégrés sans sacrifier leurs performances ou augmenter fortement leur surface. Nous étendons la structure JTM/CMOS hybride multi-bit à la conception d'une mémoire MRAM avec des circuits périphériques simples.

Basés sur le concept de LEM, des circuits logiques/arithmétiques non-volatiles sont conçus. Les JTMs sont intégrées non seulement comme des éléments de stockage, mais aussi comme des opérandes logiques. Tout d'abord, nous concevons et analysons théoriquement les portes logiques non-volatiles (PLNVs) NOT, AND, OR et XOR. Ensuite, des additionneurs complets non-volatiles (ACNVs) de 1-bit et 8-bit sont proposés et comparés avec une architecture d'additionneur classique basée sur la technologie CMOS. Nous étudions l'effet de la taille de transistor CMOS et des paramètres de JTM sur les performances d'ACNV. De plus, nous proposons deux solutions pour optimiser l'ACNV. Premièrement, un circuit de détection (mode tension) de très haute fiabilité est proposé. Ensuite, nous proposons de remplacer le JTM à deux électrodes par un JTM à trois électrodes (écrit par transfert de spin assisté par l'effet Hall de spin) en raison du temps d'écriture et de la consommation plus faibles.

Enfin, une mémoire adressable par contenu non-volatile (MACNV) est proposée. Deux architectures de décodeurs magnétiques sont proposées pour sélectionner des lignes et à enregistrer la position de recherche dans un état non-volatile.

Mots clés: spintronique, transfert de spin, jonction tunnel magnétique, circuits hybrides JTM/CMOS, circuits logiques/arithmétiques non-volatiles

Table of Contents

List of Figures

List of Tables

General introduction

Motivation

Complementary metal oxide semi-conductor (CMOS) has been the dominate technology for integrated circuits (IC) for several decades. It is widely used in both digital and analog circuits such as microprocessors, digital logic circuits, image sensors, etc. The development of CMOS technology enables ICs to follow the Moore's law, that is, the number of transistors doubles every two years [1]. However, this trend of higher density of integration is slowed down with the shrinking feature size of CMOS technology (e.g., < 90 nm), due to the increasing static and dynamic power consumption. First, the static power consumption is significantly increased due to the increasing leakage current, especially in memories that need power supply to maintain data. An unexpected power interruption causes not only data loss but also additional power and time to restart the process. The volatility of CMOS-based memories and logic circuits becomes one of the obstacles for low-power, normally-off and instant-on computing systems. Besides, there is a long distance between the memory units and the logic parts in the computing systems based on the Von Neumann architecture, which results in high dynamic transfer power consumption and long transfer delay.

All of these motivate academic and industrial research to focus on novel technologies that can partly or completely replace the CMOS technology and architecture that can eliminate the communication bottleneck. Since the discovery of giant magnetoresistance effect (GMR) in 1988 [2], emerging spintronics devices are under intensive investigation. They exploit the spin property of electrons (up or down) rather than the charge property. The tunnel magnetoresistance effect (TMR) was observed in the magnetic tunnel junction (MTJ) structure in 1975 [3]. After that, MTJs with different materials have been investigated, especially for increasing the TMR ratio. By using the MgO barrier, the TMR ratio can reach up to 600% at room temperature [4], which makes MTJ one of the most promising spintronics devices for both memory and logic applications. Most effort is devoted to MTJ-based memories such as magnetic random access memory (MRAM). Recently, many MRAM prototypes or chips have been proposed and commercialized [5], [6], [7]. The non-volatility of MTJ allows powering-off completely the system in "idle" state, thus cutting down drastically the static power.

The integration of MTJ directly into the logic circuits would pay a way for high-speed and low-power logic/arithmetic operation. Thanks to the 3-D integration technology of MTJs on

the CMOS-based logic circuits, the communication distance between the memory and logic chips are greatly shortened. Consequently, this significantly reduces the dynamic transfer power and access latency compared to the conventional systems. In the past years, non-volatile logic has been proposed and studied to get the memory closer to the processing unit and bring the non-volatility directly into the logic circuits. But these logic circuits suffer from erroneous switching with the reading current, reliability issue due to process variations and relative long switching time. Therefore, reliable read/write circuits and other writing approach that can overcome the drawbacks of mainstream writing approach (e.g., spin-transfer torque (STT)) need to be explored.

Key contributions of the PhD thesis

The objectives and contributions of this thesis are the following:

1. We study the STT-MTJ model, design and explore ways to improve the performances of the hybrid MTJ/CMOS circuit for non-volatile memory and logic design. We then extend the single-bit to multi-bit hybrid MTJ/CMOS, where several non-volatile memory cells share the same read/write circuits, and optimize its performances in terms of power consumption, density, reliability and read/write speed.

2. Design of a novel structure of 1KB MRAM with simple peripheral circuits, which is part of the project DIPMEM. It takes advantages of the multi-context hybrid MTJ/CMOS structure.

3. Based on the logic-in-memory (LIM) architecture, we propose and theoretically study different non-volatile logic gates (NVLGs) including NOT gate, AND/NAND gate, OR/NOR gate, XOR/NXOR gate.

4. Full-adder is the basic element of many arithmetic operations such as addition, division, etc. Therefore, low-power design of full-adder is becoming more and more important for portable devices including smart phones, tablets and sensors. On this basis, we design single-bit and multi-bit non-volatile full-adders (NVFAs). Optimizations have been proposed to improve the reliability, writing speed and energy from the circuit-level and device-level (Spin-Hal-assisted STT writing approach).

5. A non-volatile content addressable memory (NVCAM) is proposed. In order to store the search location in non-volatile states to avoid data loss in case of unexpected power-off, two magnetic decoders are designed.

All the circuits are designed and simulated on Cadence platform by using the STMicroelectronics CMOS 28 nm and 40 nm design kits. We use the MTJ models based on STT or spin-Hall-assisted STT from the NANOARCHI group in IEF laboratory.

Thesis organization

The thesis is organized as follows:

Chapter 1 gives an overview of the basic technologies and principles related to our work including the development of spintronics, MTJ and its writing approaches, MTJ-based memory and logic circuits and non-volatile logic circuits based on other spintronics devices.

Chapter 2 introduces the compact model of STT-MTJ that will be used in our circuit design and simulation. Single-bit and multi-bit hybrid MTJ/CMOS circuits are studied including read/write circuits and strategies to improve their performances (e.g., speed, power, reliability, etc.). As an example of MTJs used in memory applications, a novel structure of 1KB MRAM is proposed based on the multi-context hybrid MTJ/CMOS circuit.

Chapter 3 focuses on the design and analysis of non-volatile logic circuits, i.e., NVLGs, 1-bit and 8-bit NVFAs, based on the LIM architecture. Comparison between the proposed NVFA and the conventional CMOS-based FA is presented. In order to improve the performances of NVFA for low-power high-speed and reliable operation, we propose optimizations from the circuit- and device- level. Spin-Hall-assisted STT writing mechanism is applied for higher writing speed and lower writing energy.

Chapter 4 details the design of NVCAM by combining the multi-context idea and the LIM-based NVLG. Performance analysis and comparison with other CAM circuits are presented as well. Two magnetic decoders (MD) are designed for word line selection. These MDs store the search location in MTJ cells and save energy if a power-off occurs.

Finally, we will conclude this thesis and provide possible future directions.

Chapter 1 State-of-the-art

This chapter presents the state of the art and development of spintronics. Magnetic tunnel junction (MTJ) technology and different writing approaches are introduced. Finally, the current status of magnetic random access memory (MRAM) and non-volatile logic circuit based on MTJs and other spintronics devices are presented.

1.1 Spintronics

Mass, charge and spin are three intrinsic properties of electron. The mainstream complementary metal oxide semi-conductor (CMOS) technology [8], [9] only considers the charge property (+ or -) of electrons and is used in memories, microprocessors, analog and digital circuits etc. Spintronics (or spin-electronics), however, aims at exploiting the spin property of electrons (up ↑ or down ↓) and create new devices. The concept of spin is firstly proposed by Wolfgang Pauli in 1925 and successfully explained by Paul Dirac in 1928 in the relativistic quantum mechanics [10]. The origin of spintronics can be tracked back to the study of spin-dependent electron transport phenomena in the 1970s [3], [11]. But it wasn't used in practical applications due to the limitations of technology and equipment until the independent discovery of giant magnetoresistance effect (GMR) by Albert Fert in 1988 [2] and by Peter Grünberg in 1989 [12], separately.

GMR is based on spin-dependent scattering in specific FM/NFM/FM structure (ferromagnetic/non-ferromagnetic/ferromagnetic). When the injected electrons pass through the FM material, the scattering probability depends on the magnetization direction of the FM layer. That is, the electrons whose spin direction is identical to the magnetization direction of the FM layer experience less scattering, corresponding to low resistance R_L. Contrarily, the electrons whose spin direction is opposite to the magnetization direction of the FM layer experience significant scattering, corresponding to high resistance R_H [13].

The basic mechanism of GMR effect was explained by Mott's two spin-channel model (or two-current model) [14]. The electrons can be divided in two channels, spin-up electron channel and spin-down electron channel.

- When the relative magnetization direction of two FM layers is parallel, the spin-up electrons are barely scattered and pass through the three-layer structure easily, whereas the spin-down electrons suffer from significant scattering in both FM layers. The resistor network in illustrated in Figure 1.1(a) and this stack performs low resistance

$R_P = 2R_H R_L (R_H + R_L)$ in this case.

- When the FM layers are anti-parallel, both spin-up and spin-down electrons suffer from scattering in either FM_1 or FM_2. As shown in Figure 1.1(b), each channel can be regarded as a low resistance R_L and a high resistance R_H connected in series. The total anti-parallel resistance is $R_{AP} = (R_H + R_L)/2$. The magneto-resistance ratio is defined as:

$$GMR = \frac{\Delta R}{R_P} = \frac{(R_H - R_L)^2}{4R_H R_L} \qquad \text{Eq. 1.1}$$

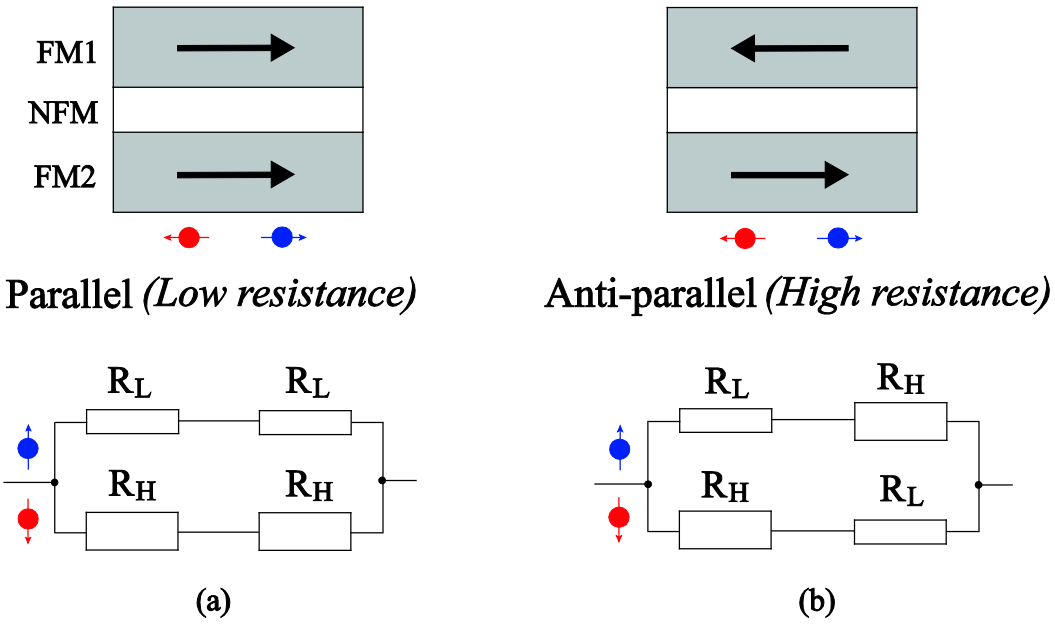


Figure 1.1 Two spin-channel model of GMR effect induced by spin-dependent scattering

GMR is shown as one of the most important milestones in physics research. It is not only the "birth" of a new discipline, but also a model that pushes the fundamental research towards industrial products. The first commercial GMR sensor was announced in 1994 [15]. Nowadays, GMR sensor is used in data storage, biological applications, space applications etc. [16]. The first hard disk drive (HDD) with GMR read head produced by IBM in 1994, increasing more than ten times the storage density [17]. Currently, the storage density of HDD based on GMR effect is more than 500 Gb/in^2. GMR effect was also exploited in the development of MRAM until the discovery of tunneling magnetoresistance effect (TMR) (discussed in the next paragraph), which shows more advantages in MRAM applications.

1.2 Magnetic tunnel junction (MTJ)

1.2.1 Tunneling magnetoresistance (TMR) effect and MTJ structure

Tunneling magnetoresistance effect (TMR) was first observed by Jullière in Fe/Ge/Co junction in 1975 [3]. Conductance measurement depends on the spin polarizations. As can be seen in Figure 1.2(a), a magnetic tunnel junction (MTJ) is principally composed of a thin insulating barrier (e.g., Al_2O_3) sandwiched between two FM layers. The spin direction in one FM layer (reference layer or pin layer) is fixed, whereas in the other FM layer (storage layer or free layer) is free to change. Due to the TMR effect, MTJ is able to present two states, i.e., parallel (P) and anti-parallel (AP), corresponding to low- and high- resistance by changing the relative magnetization orientation of two FM layers (see Figure 1.2(c)). MTJ can be switched between two states (P state and AP state) by external magnetic fields or an injecting current flowing through the nanopillar. Therefore, an MTJ can be essentially considered as a two-value resistor.

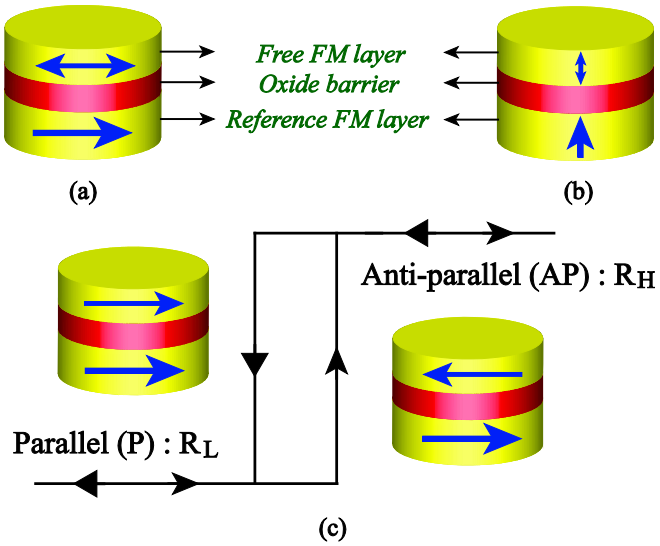


Figure 1.2 (a) In-plane magnetic tunnel junction (MTJ) (b) Perpendicular MTJ (c) Tunneling magnetoresistance (TMR) effect in an MTJ nanopillar

TMR ratio, which characterizes the amplitude of resistance change, is defined as:

$$TMR = \frac{\Delta R}{R_P} = \frac{R_{AP} - R_P}{R_P} = \frac{\Delta G}{G_{AP}} = \frac{G_P - G_{AP}}{G_{AP}} \qquad \text{Eq. 1.2}$$

where R_P and R_{AP} are the resistances of MTJ in P and AP state, G_P and G_{AP} are the conductances of MTJ in P and AP state, respectively.

The electrons travel from a FM layer, through an insulating layer (1~2 nm), and then into another FM layer by spin-dependent tunneling, as illustrated in Figure 1.3. In an FM material, the number of spin-up and spin-down carriers at Fermi level is unequal, resulting in an imbalance of spin populations [18].

- In the electric transport of P state, all the spin-up and spin-down electrons at the Fermi level of one FM layer can easily tunnel into another FM layer because the states available for spin-up and spin-down electrons of two FM layers are equal in number. Therefore, MTJ presents low resistance characteristic.

- In the electric transport of AP state, on the contrary, only partial electrons of FM_1 can reach FM_2 due to imbalance of state density of two FM layers. It can cause a higher resistance than that of P state (see Figure 1.4) [19].

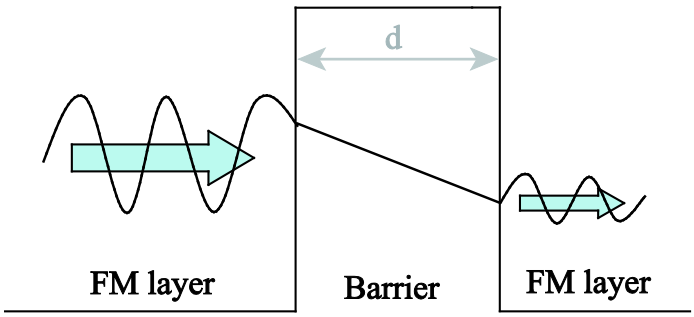


Figure 1.3 Tunneling through an insulating barrier

In the Jullière model, the conductance (G_P for P state or G_{AP} for AP state) is proportional to the density of spin-up and spin-down states, as expressed by equations Eq. 1.3 and Eq. 1.4.

$$G_P = G_{\uparrow\uparrow} + G_{\downarrow\downarrow} \propto N_{1\uparrow}N_{2\uparrow} + N_{1\downarrow}N_{2\downarrow} \qquad \text{Eq. 1.3}$$

$$G_{AP} = G_{\uparrow\downarrow} + G_{\downarrow\uparrow} \propto N_{1\uparrow}N_{2\downarrow} + N_{1\downarrow}N_{2\uparrow} \qquad \text{Eq. 1.4}$$

where $N_{1\uparrow}$ and $N_{2\uparrow}$ are spin-up electrons in FM_1 and FM_2, $N_{1\downarrow}$ and $N_{2\downarrow}$ are spin-down electrons in FM_1 and FM_2.

The spin polarization at Fermi level is defined in terms of spin-up and spin-down carriers' number.

$$P_1 = \frac{N_{1\uparrow} - N_{1\downarrow}}{N_{1\uparrow} + N_{1\downarrow}} \quad and \quad P_2 = \frac{N_{2\uparrow} - N_{2\downarrow}}{N_{2\uparrow} + N_{2\downarrow}} \qquad \text{Eq. 1.5}$$

Based on Eq. 1.2-Eq. 1.5, TMR ratio can be defined by the following equation, which shows that TMR effect strongly depends on spin polarizations of two FM layers.

$$TMR = \frac{2P_1P_2}{1-P_1P_2} \qquad \text{Eq. 1.6}$$

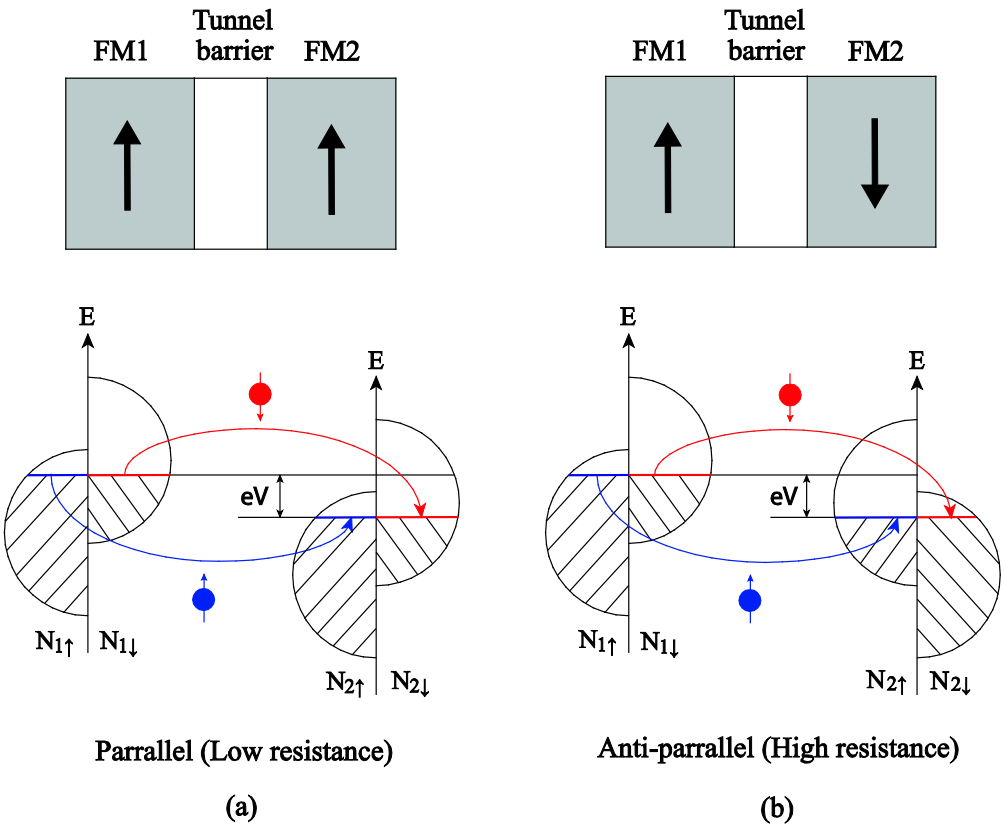


Figure 1.4 Spin-dependent tunneling in MTJ nanopillar which is in (a) parallel state (b) anti-parallel state

When compared to the GMR device, a TMR device replaces the non-ferromagnetic layer (e.g., Cr) with an insulator (e.g., Al_2O_3). TMR effect didn't attract much attention at that time due to technical limitations until the reports of repeatable TMR effect by using Al_2O_3 as the insulating layer (TMR ratio up to 18% at room temperature) in 1995 [20], [21].

1.2.2 TMR ratio enhancement

As mentioned above, an MTJ can store stable data '0' (AP state) or '1' (P state). It has compatible resistance values (~ kΩ) with the CMOS transistor technology. Besides, MTJs can be fabricated above CMOS-based circuits by back-end process [22] to reduce die area. All these advantageous features make MTJ possible to develop hybrid MTJ/CMOS circuits with high performances. TMR ratio is an important factor for detecting the MTJ state by

CMOS-based circuit such as pre-charge sense amplifier (PCSA) [23]. In order to ensure reliable sensing face to process variations, high TMR ratio is strongly required, especially for hybrid logic circuits. TMR ratio can only reach 70%-80% if the amorphous Al_xO_y is used as the insulating layer [24], [25]. During the last years, researches have focused on MTJ using MgO barrier for higher TMR ratio [26], [27].

W. H. Butler [28] and J. Mathon [29] predicted in 2001 that TMR ratio of Fe/MgO/Fe junction may exceed 1000%. Later in 2004, Parkin from IBM and Yuasa from AIST separately obtained TMR ratio of 220% [30] and 180% [31] at room temperature. In the following years, higher TMR ratios were successively observed, for instance, 230% [32] and 260% [33] in 2005, 410% in 2006 [34], 500% in 2007 [35] etc. The latest report in [4] showed that 604% TMR ratio was observed with CoFe/MgO/CoFe junction.

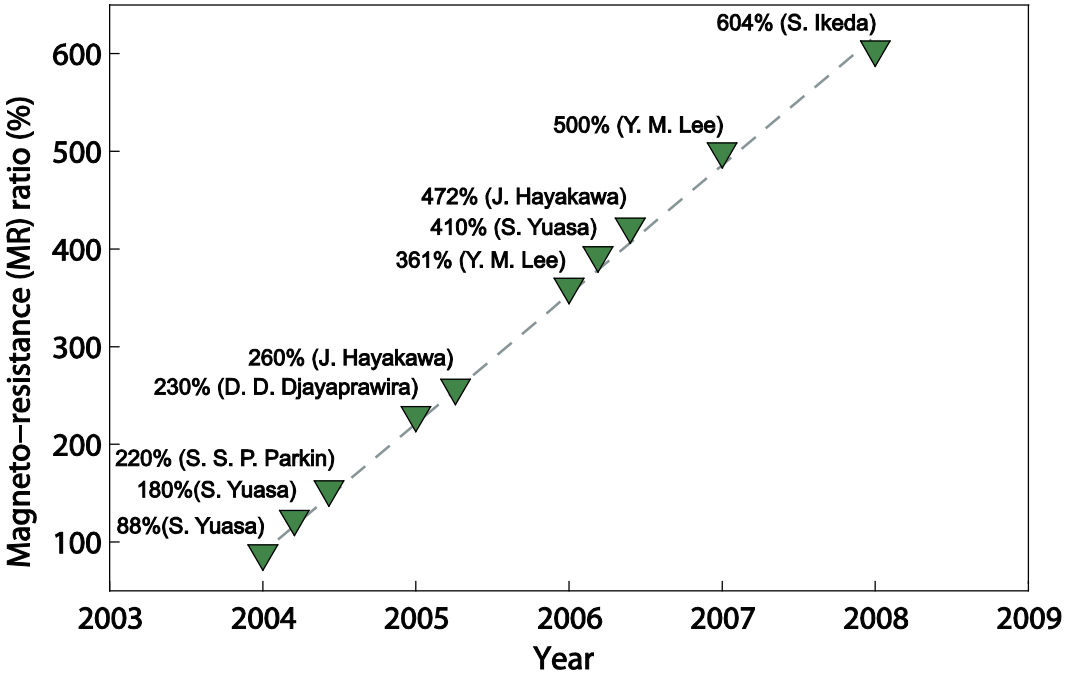


Figure 1.5 Research progress of TMR ratio (MgO based MTJ)

Thanks to the progress of TMR ratio in the past decade, two states of MTJ (P and AP) can be easier to be distinguished. This makes MTJ become one of the most promising candidates for both non-volatile memory and logic circuits.

1.2.3 Writing approaches

1.2.3.1 Field-induced magnetic switching (FIMS)

Field-induced magnetic switching mechanism is a writing approach of MTJ for the first

generation MRAM [36]. The magnetization of the free layer is switched by easy-axis and hard-axis external magnetic fields generated by two orthogonal current lines, i.e., bit line and digit line as shown in Figure 1.6 [37]. The switching from P to AP (P→AP) or from AP to P (AP→P) depends on the direction of the current flowing though the bit line.

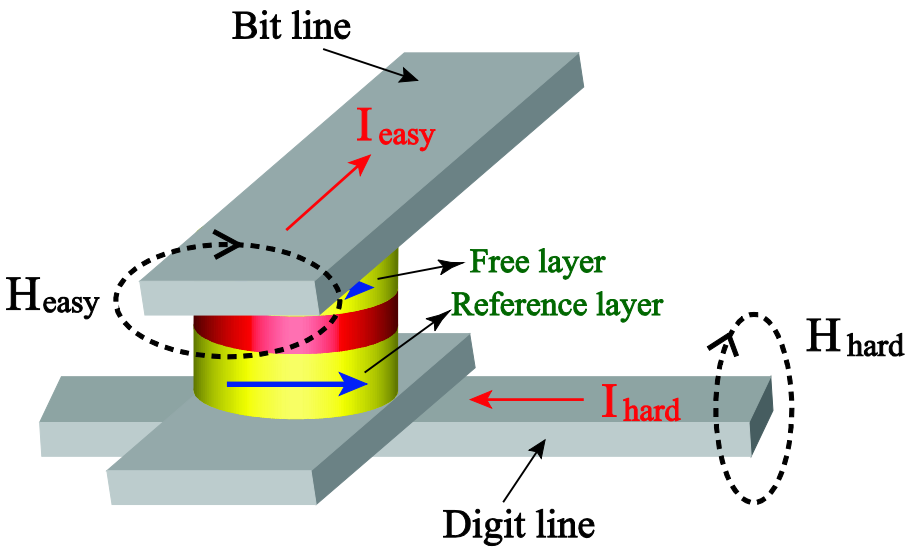


Figure 1.6 Field-induced magnetic switching (FIMS) writing approach

FIMS writing approach suffers from two main issues.

- First, the currents needed for MTJ switching are too high (~ mA), leading to high power consumption, low density and low scalability.
- Besides, half-selectivity disturbance is another disadvantage, especially for FIMS based MRAM (FIMS-MRAM) [5]. As shown in Figure 1.7, the selected MTJ is situated at the cross point of the word line (WL) and the bit line (BL). The MTJs near the selected MTJ (half-selected MTJs), however, are also influenced by the external magnetic fields generated by the WL and BL. The fields at the selected MTJ must be large enough to switch its configuration. Conversely, the fields at the half-selected MTJs generated by WL or BL should be small for not switching their states.

Engel et al. from Freescale proposed the toggling switching mechanism to solve the half-selectivity problem [38]. The free layer of MTJ is replaced by a synthetic antiferromagnet (SAF) layer, two FM layers sandwiched by a non-magnetic coupling layer. MTJ is placed 45° to the current lines. The pulse sequence of two write lines and the switching principle is illustrated in Figure 1.8.

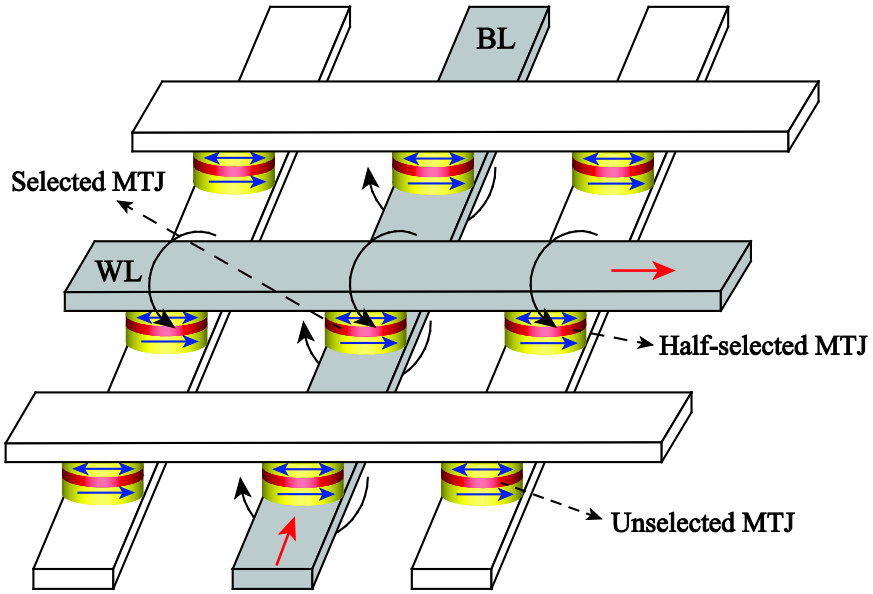


Figure 1.7 Half-selectivity issue of FIMS based MRAM (FIMS-MRAM)

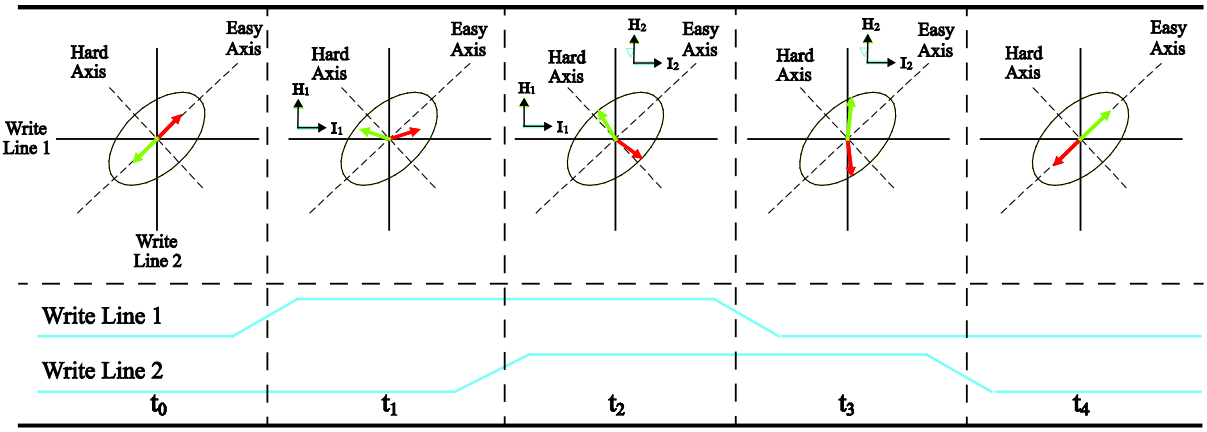


Figure 1.8 Schematic of the toggling operation [38]

Because the toggling switching method still uses magnetic field to change the state of MTJ, it cannot avoid the disadvantages of high currents and power consumption, large area and low density. Therefore, other switching approaches need to be exploited, especially for MTJs to be embedded in low-power processors.

1.2.3.2 Thermal assisted switching (TAS)

Thermal assisted switching (TAS) was proposed to overcome the aforementioned issues caused by field magnetic switching [39]. Different antiferromagnetic layers are added below the pinned FM layer (AF_1) and above the free FM layer (AF_2) [40]. The blocking temperature of AF1 is higher than that of AF_2. As shown in Figure 1.9, a current I_h passing through the MTJ stack heats the MTJ. When the temperature is above the blocking temperature of the free

layer, its magnetization is free to be changed by an external magnetic field induced by another current I_m.

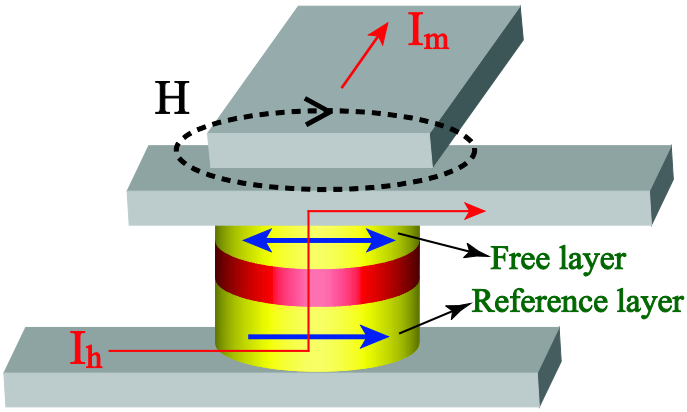


Figure 1.9 Thermal assisted switching (TAS) writing approach for MTJ

When compared to the FIMS writing approach, TAS can minimize the half-select switching since only the free FM layer of one selected MTJ is unpinned by current I_h. Besides, only one magnetic field (H in Figure 1.9) is needed to switch the state of MTJ, thus greatly reducing the writing energy and circuit area. Thanks to its lower power, higher density and higher thermal stability, TAS has been used to build MRAM [7], [41], look-up-table (LUT) [42]. Nevertheless, MTJ has to be cooled after switching operation with relative long cooling duration (tens of nanoseconds). For this reason, TAS cannot meet the requirement of high speed for applications of logic or register.

1.2.3.3 Spin-transfer torque (STT)

Spin transfer torque (STT) is another breakthrough since the discovery of GMR effect. In 1996, Slonczewski [43] and Berger [44] theoretically predicted that the magnetization of free layer (FL) could be influenced by the injected current larger than a critical current, denoted as I_{C0}. When the injected electrons flow perpendicular from the reference layer (RL) to the FL, they are spin-polarized and aligned to the magnetization direction of the RL. When the electrons reach the FL, their spin angular momentum is transferred to the magnetization of FL following the total angular momentum conservation law. A large torque called spin-transfer torque (STT) (or spin-current-induced torque) is applied to align the magnetization of FL towards that of RL. When the electrons flow in the opposite direction, the magnetization of FL is forced to be anti-parallel to RL by the reflected electrons. The switching of the magnetization of FL in two cases is shown in Figure 1.10, assuming that the current injected

from FL (or RL) is positive (or negative).

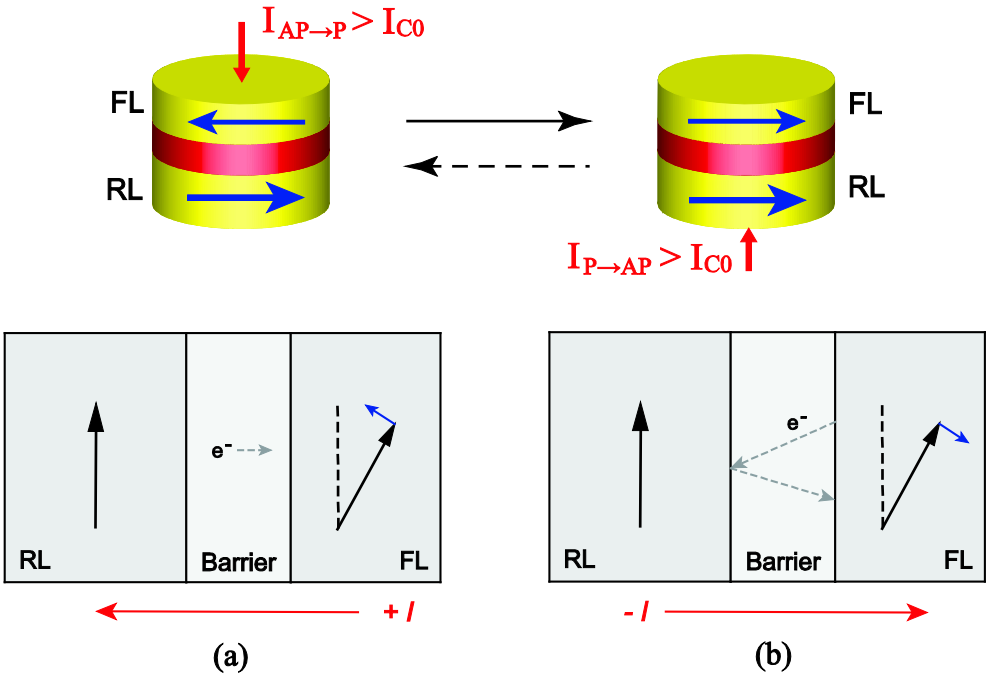


Figure 1.10 Spin-transfer switching (a) to parallel state (b) to anti-parallel state

The dynamic behavior of the magnetization of FL can be explained by the modified Laudau-Lifshitz-Gilbert (LLG) equation [45], [46]. The precession of moment $\boldsymbol{m}$, unit vector along the magnetization direction of FL, is influenced by three torques, shown in Figure 1.11 [47], [48]. Field torque makes $\boldsymbol{m}$ precess in circle. Gilbert damping torque decreases the precession angle θ and pushes $\boldsymbol{m}$ back towards the magnetic field H_{eff} ($\hat{\boldsymbol{z}}$ direction). The direction of STT is either parallel or anti-parallel to the Gilbert damping torque, depending on the sign of current shown in Figure 1.10. In the former case, STT increases the speed back to $\hat{\boldsymbol{z}}$. In the latter case, the STT weakens the damping. And if the current is above I_{C0}, $\boldsymbol{m}$ processes away from H_{eff} by increasing the θ value. In this case, the magnetization of FL is switched after a certain delay.

When compared to the field-based switching approaches FIMS or TAS, STT only needs a bi-directional current and the current density is lower (10^6~10^7 A/cm^2). STT writing mechanism greatly simplifies the writing circuit in hybrid circuit design, while keeping lower power and higher density. Moreover, half-selectivity can be avoided because the writing current only passes through the selected MTJ. Currently, MTJ based on STT switching (STT-MTJ) is widely investigated and applied to both memory (STT-MRAM) [49] and logic design, which are also the main topic of this thesis.

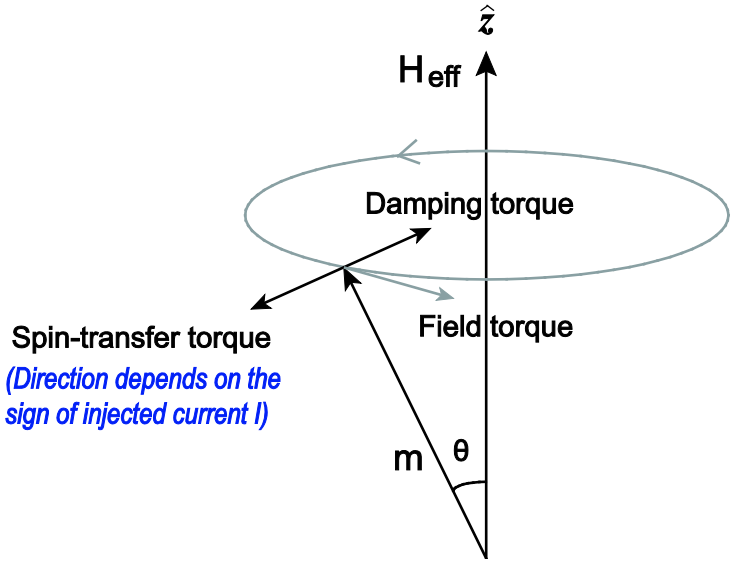


Figure 1.11 Schematic of the Laudau-Lifshitz-Gilbert (LLG) dynamic model

In order to fully take advantages of MTJ in hybrid circuit design, low critical current and high thermal stability are two criteria that should be ensured. With the shrinking of MTJ size (< 40 nm), in-plane anisotropy MTJ (i-MTJ) can no longer provide high energy barrier and high thermal stability performances. The energy barrier and thermal stability factor are given by Eq. 1.7-Eq. 1.8 [50], [51]. If the size of MTJ decreases, thickness t and AR need to be increased to maintain the thermal stability without changing the free layer material. The value of AR is nearly 3 when the width W shrinks down to 40 nm. Therefore, i-MTJ is usually ellipse. Perpendicular magnetic anisotropy (PMA) based MTJ (p-MTJ) is illustrated in Figure 1.2(b). The energy barrier and the thermal stability factor of p-MTJ are given by Eq. 1.9-Eq. 1.10. By comparing Eq. 1.8 and Eq. 1.10, p-MTJ has higher thermal stability when maintaining the same size because H_K is much larger than H_C.

$$E(\text{in}-\text{plane}) = \frac{\mu_0 M_S \times V \times H_C}{2} \qquad \text{Eq. 1.7}$$

where μ_0 is the permeability in the free space, M_S is the saturation magnetization, V is the volume of the free layer, H_C is the in-plane anisotropy field.

$$\Delta(\text{in}-\text{plane}) = \frac{E}{k_B T} = \frac{\mu_0 M_S \times V \times H_C}{2k_B T} \propto t^2 W(\text{AR}-1) \qquad \text{Eq. 1.8}$$

Where k_B is the Boltzmann's constant, T is the temperature, t, W and AR are thickness,

width and aspect ratio of the free layer.

$$E(\text{perpendicular}) = \frac{\mu_0 M_S \times V \times H_K}{2} \qquad \text{Eq. 1.9}$$

where H_K is the perpendicular anisotropy field.

$$\Delta(\text{perpendicular}) = \frac{E}{k_B T} = \frac{\mu_0 M_S \times V \times H_K}{2 k_B T} \qquad \text{Eq. 1.10}$$

In addition, i-MTJ has to overcome additional out-of-plane demagnetizing field for STT switching. Therefore, p-MTJ has lower critical current and higher switching speed than i-MTJ when keeping the same thermal stability [52]. p-MTJ has attracted much interest in recent research [53], [54], [55], [56], [57]. An exciting result was reported by using the Ta/CoFeB/MgO structure in 2010 [58]. High tunnel magnetoresistance ratio (> 120%), low critical current (49 μA), high thermal stability at small MTJ dimension (40 nm) were obtained. It is the experimental basis of the compact PMA STT-MTJ model that we use for the hybrid logic and memory design.

1.2.3.4 Spin Hall effect (SHE)

Even though current-induced STT writing mechanism exhibits many attractive features, it still has some disadvantages for MTJs to be embedded in logic circuits where speed is critical, STT needs long incubation delay (several nanoseconds) at the initial switching stage, due to random thermal fluctuations [59]. The low switching speed greatly limits its development for faster computing system. Besides, large bi-directional current passing though the MTJ nanopillar leads to larger writing circuit and higher risk of barrier breakdown. Since the read and write of the two-terminal MTJ device share the same current path, read and write operations should be separated, and the read current should be small enough to avoid erroneous writing.

Spin Hall effect (SHE) is another way to switch the magnetization of the free layer by an in-plane injecting current [60], [61]. Three-terminal magnetic device based on SHE has been proposed, where a heavy metal strip (e.g., Ta, Pt) with a large spin-orbit coupling parameter is placed below the free layer. When a current passes through the heavy metal, electrons with different spin directions are scattered in opposite directions (see Figure 1.12). The spin-orbit coupling converts the charge current into perpendicular spin current, generating a torque

called spin-orbit torque (SOT, or spin Hall torque) to assist magnetization reversal [62].

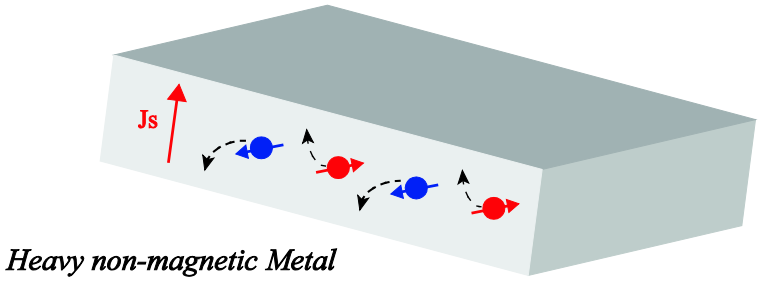


Figure 1.12 Spin Hall effect

The orientation of the free layer is controlled by the direction of the injecting current. For i-MTJ shown in Figure 1.13(a) [63], [64], a current large enough can switch the state of MTJ, similar to the STT switching. But for p-MTJ shown in Figure 1.13(b) [65], an external field is required because the direction of electron spin and the anisotropy axis are not collinear. The additional field makes the hybrid MTJ/CMOS circuit design more complex. Another solution proposed in [66] is to use STT writing current in the place of external field. As it can be seen in Figure 1.14, a short (0.5 ns) SHE current pulse is sufficient to eliminate the incubation delay and increase the writing speed. This SHE-assisted STT writing approach (STT+SHE) will be used in our design for higher writing performances.

High writing speed and low writing energy can be achieved thanks to the low resistance and strong spin-orbit interaction of the heavy metal strip. The three-terminal device can also solve the endurance and read disturbance problems of the two-terminal device by separating the read and write current paths. Recently, many circuits have been designed based on SHE assisted switching such as SHE flip-flop [67], [68] and SHE MRAM [66].

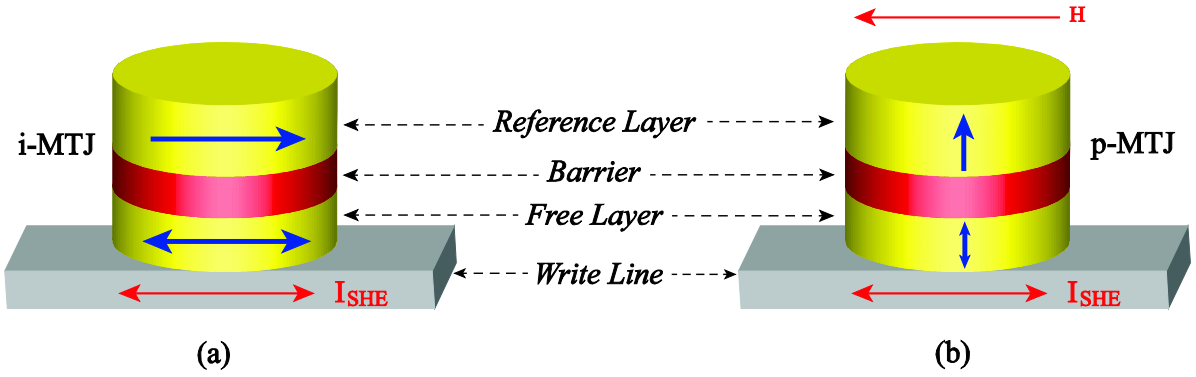


Figure 1.13 Schematic of the three-terminal device based on spin Hall effect (SHE) using (a) i-MTJ (b) p-MTJ

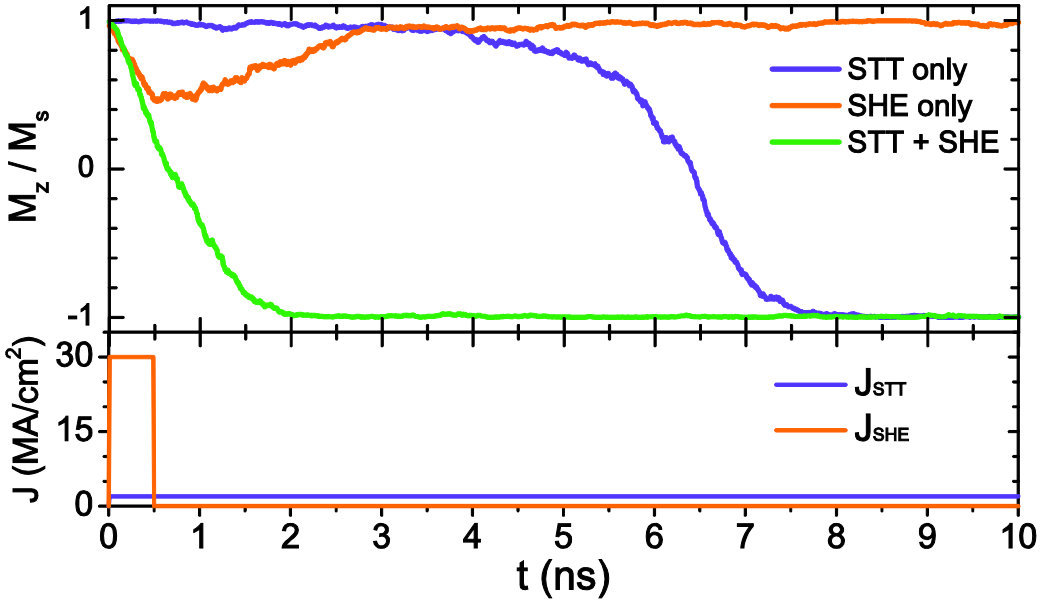


Figure 1.14 Magnetization trajectories along with the applied current pulses. 0.5 ns in-plane polarized current pulse is applied for the STT+SHE case [66].

Table 1.1 concludes the comparison of the four writing approaches including FIMS, TAS, STT and SHE. It should be noted that the writing speed of TAS approach is relative low due to the cooling phase after switching. In the following of the thesis, we use STT as the writing mechanism for MTJ and SHE as a solution to the low writing speed issue of STT.

Table 1.1 Comparison of different writing approaches

	Switching mechanism	**Power**	**Speed**	**Area**	**Half-selectivity**	**Read/Write paths**
FIMS	Field	High	Low	Large	Yes	Same
TAS	Field	Medium	Low	Medium	No	Same
STT	Current	Low	Medium	Small	No	Same
SHE	Current[a)]	Low	High	Small	No	Independent

a) An applied current can switch the state of i-MTJ. But an additional field or STT writing current is necessary for switching the state of p-MTJ.

1.3 MTJ-based hybrid memory and logic circuits towards low-power computing system

1.3.1 Magnetic random access memory (MRAM)

Magnetic random access memory (MRAM) is a promising non-volatile memory. The concept of MRAM was proposed in 1972, but the discovery of GMR and TMR phenomenon has pushed forward the development of MRAM. Many MRAM prototypes or chips based on integration MTJ and CMOS technology have been proposed and commercialized. The first generation of MRAM was based on FIMS switching and the first commercial Toggle-MRAM was commercialized by IBM in 2006 [6]. TAS-MRAM and STT-MRAM are the following generation of high-speed low-power consumption MRAMs. STT-MRAM has good scalability (e.g., 22 nm [69]) and becomes a promising candidate for the universal memory [70].

Table 1.2 summarizes the features of MRAM compared with other semiconductor memory technologies [71]. Static random access memory (SRAM) has fast read/write speed, but it needs large cell area and suffers from increasing static power consumption due to leakage current. Dynamic random access memory (DRAM) has simpler cell structure (one pass transistor and one capacitor), but it has to be refreshed to preserve data. Flash memory has limited endurance, high write power and slow write speed. When compared to the other memory technologies, MRAM combines the features of non-volatility, unlimited read/write endurance ($> 10^{15}$ cycles), fast read/write time (< 10 ns), large capacity and nearly zero standby power consumption.

Table 1.2 Comparison of MRAM with other memory technologies

	SRAM	**DRAM**	**FLASH**	**STT-MRAM**
Cell size	Large	Small	Small	Small-Medium
Read time	Fast	Slow	Medium	Medium-Fast
Write time	Fast	Medium	Slow	Medium-Fast
Write power	Low	Low	High	Low
Endurance	High	High	Slow	High
Non-volatility	No	No	Yes	Yes
Refresh	No	Yes	No	No
Low voltage	Yes	Limited	No	Yes

The conventional memory hierarchy is illustrated in Figure 1.15. Each level is distinguished by speed and capacity. The higher-level memories are faster but of smaller size and are used for the active data. The lower-level memories are slower but of larger size to store large data. Since the working memories (SRAM for cache memory and DRAM for main memory) are volatile, they consume large energy to store data. During the booting process, data is transmitted all the way from the storage memory (Flash or hard drive) to the main and active memories, wasting large amount of energy and time [72]. Before the system is shut down, data has to be backuped in storage memory, which requires additional energy and time.

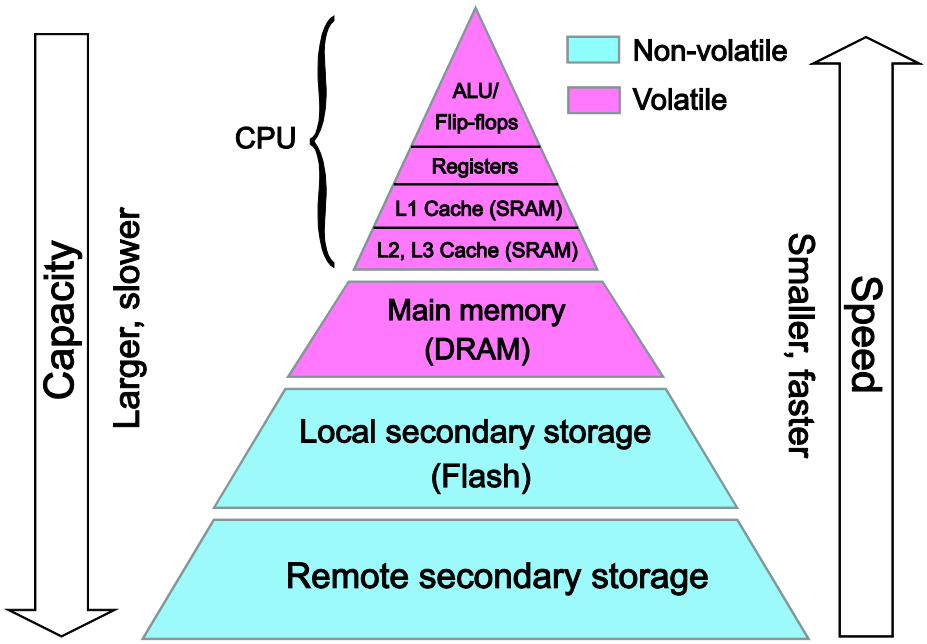


Figure 1.15 Structure of the current computer memory hierarchy

MRAM is a potential technology to eliminate most of the issues by bringing non-volatility into the main and active memory while keeping high-speed and sufficient density. But MRAM is not used to directly replace all the volatile memories. The increment of writing power of MRAM compared with that of SRAM should be smaller than the standby power. The power consumption due to the leakage current in SRAM cache memory is dominated by level-2 (L2) cache because it has a larger capacity than level-1 (L1) cache. Besides, most of the power dissipation of L2 is standby power while that of L1 is active power. Therefore, MRAM is a possible solution for the main memory and low-level cache like L2 and L3 caches when building a low-power processor [73], [74]. Figure 1.16 represents the power comparison of the SRAM-based and MRAM-based cache memory.

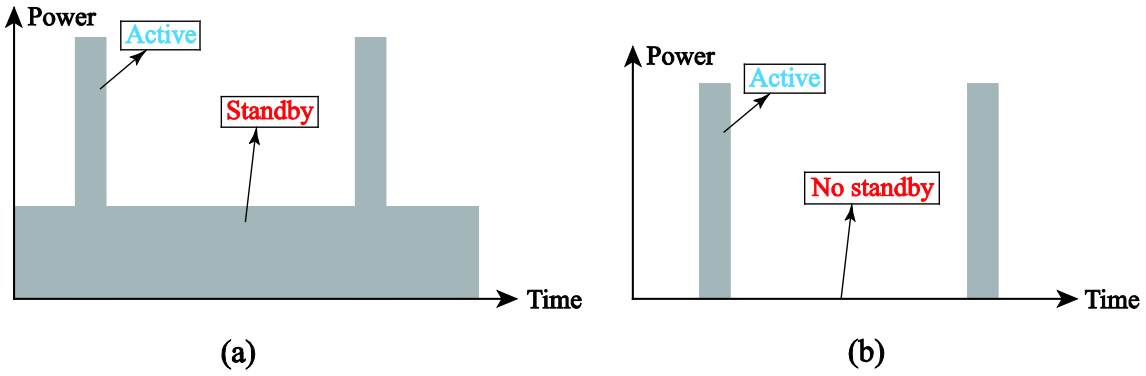


Figure 1.16 (a) Conventional SRAM-based cache memory (b) MRAM-based cache memory

As stated before, MTJ is the basic storage element of MRAM thanks to its fast access speed, easy integration with CMOS technology and good data retention time (around ten years [75]) etc. An MTJ is able to store 1-bit data by switching the resistance (low resistance or high resistance). Figure 1.17(a) shows the basic 1T1M memory cell and the STT-MRAM architecture [76]. One terminal of MTJ is connected to the bit line (BL) and another terminal is serially connected with a transistor. The gate and source terminals of the NMOS transistor are respectively connected to the word line (WL) and source line (SL). In order to write data into an MTJ, a writing current, which is larger than the critical current, is applied to the memory cell. Data reading is performed by applying a sensing current or a bias voltage between BL and SL. Apart from the 1T1M memory cell, there are some other cell structures proposed in literatures, for instance, 1T2M memory cell [77], 1T4M memory cell [78], 2T1M memory cell [79], 2T2M memory cell [80], 4T2M memory cell [81] etc.

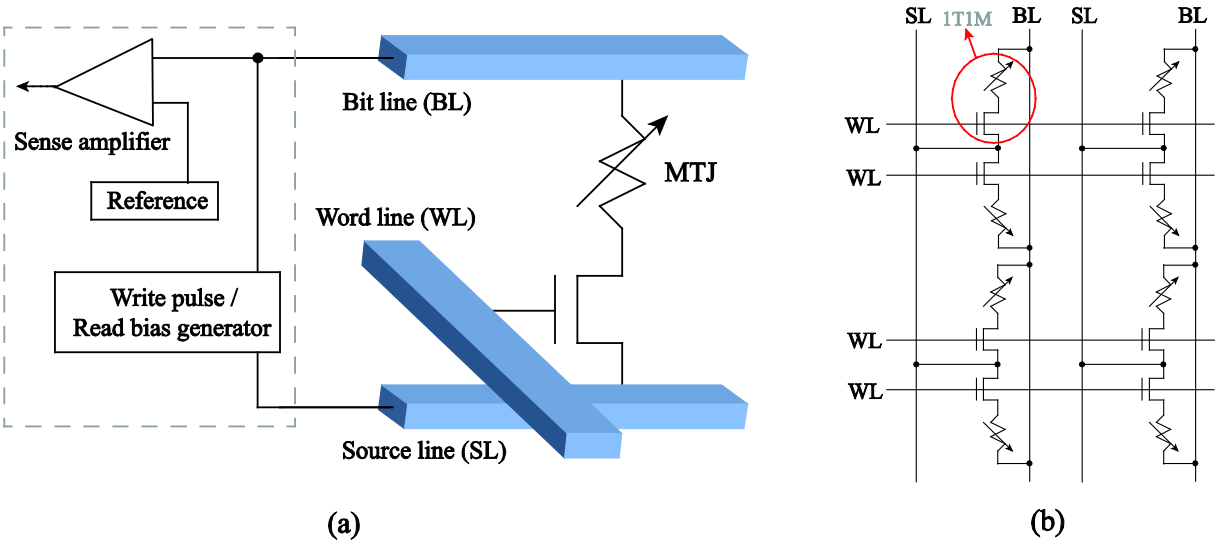


Figure 1.17 (a) 1T1M memory cell where one MTJ and one NMOS transistor are connected in series (b) MRAM architecture based on 1T1M cell

One drawback of the MRAM architecture illustrated in Figure 1.17(b) is the requirement of one transistor for each memory cell, which limits the storage density. In another MRAM architecture (cross-point architecture), MTJs are places at the cross points of two orthogonal lines (see Figure 1.18). This architecture promised high density but, however, suffers from sneak current issue, which disturbs the data reading. Some design considerations have been proposed to overcome this issue in [82]: 1) balanced sensing structure: the number of MTJs on two branches of the sense amplifier (SA) is the same, that is, one reference MTJ and M storage MTJs; 2) parallel data reading: all the SAs (N in Figure 1.18) operate in parallel.

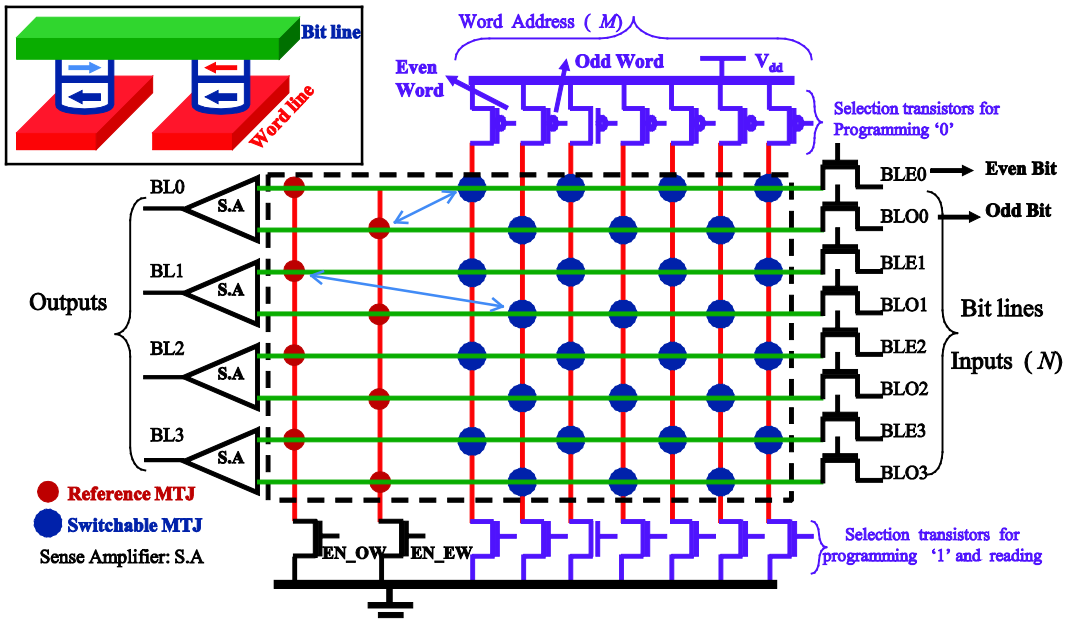


Figure 1.18 Schematic of the cross-point architecture for MRAM [82]. A cross-point array of MTJs is for data storage and another cross-point array is reference MTJs.

1.3.2 Non-volatile logic circuits

1.3.2.1 Logic-in-memory

Today's computing systems are mainly built on John von Neumann architecture [83]. As shown in Figure 1.19(a), logic and memory are separate functions, and they are connected through complex interconnections with a relatively long transfer distance. This usually results in long transfer delay (or low operation speed) and high transfer power dissipation (e.g., ~ 1 pJ/bit/mm). Even if the device size scales with the progress of CMOS technology, the interconnections are not shortened neither at a better speed. Besides, since the memories (e.g., SRAM) are volatile, they always need power to keep the computing data in stand-by state. Indeed, subthreshold and gate leakage currents are increasing, and high power issues have

become the main drawbacks of CMOS logic circuits as technology node shrinks below 45 nm. For this reason, reduction of static and dynamic power as well as the interconnection delay becomes two major objectives for the next generation computing system.

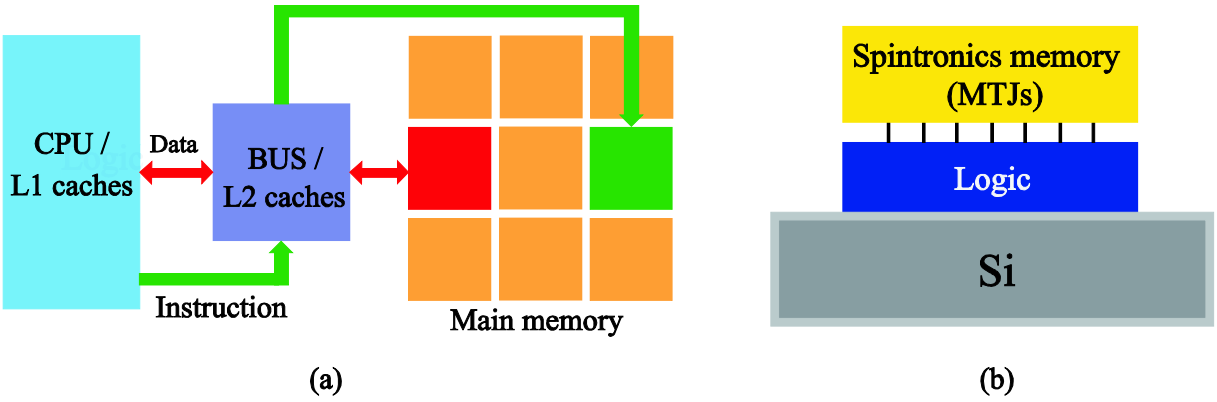


Figure 1.19 (a) Diagram of the classic Von Neumann architecture. Memory and logic chips are separated and connected by bus and cache memories. (b) 3-D hybrid logic structure

Logic circuits based on logic-in-memory (LIM) architecture can overcome the performance bottleneck of CMOS-only logic circuits. The concept of LIM was firstly introduced in 1969 [84]. Each cellar array combines both memory and logic for lower power and higher access speed. Since the discovery of non-volatile memories that have both memory and logic abilities, LIM becomes a promising architecture to build non-volatile logic functions. The LIM architecture has many advantages over the Von Neumann architecture:

- Non-volatile memories such as MTJs can be easily deposited above the logic-circuit plane by means of three-dimensional (3-D) back-end integration. This allows reducing the global routings and significantly shortening the distance between memory and logic chips from ~ mm to ~ µm. The total computing speed is thus accelerated and the dynamic transfer power consumption is reduced.

- The storage and the logic operation elements are merged into the same spintronics devices. Spintronics devices do not occupy extra area, so the die area is further reduced.

- The storage elements and the logic circuits are connected by vertical vias. The simpler interconnect paths allow much lower capacitance and dynamic power dissipation.

- Since the storage elements are non-volatile, the temporarily unused blocks can be powered off without loss of data for saving standby power consumption. Data can be

instantaneously recovered during the operation procedure, thus the LIM-based logic is suitable for the “normally-off” and “instant-on” system.

In order to fully take advantages of the non-volatile logic circuits, the spintronics devices should combine the features of high read/write speed, unlimited endurance, small size and compatibility with CMOS technology. The aforementioned progress of MTJ devices makes them suitable to be integrated with conventional CMOS-based logic and memory circuits. Recently, innovative circuits based on hybrid MTJ/CMOS circuits have been presented [85]. For instance, magnetic look-up-table (MLUT) [86], [87] and magnetic flip-flop (MFF) [88], [89] were introduced for reconfigurable logic such as field-programmable gate arrays (FPGA). As a typical example of MTJ-based logic circuit using LIM architecture, the first test chip of magnetic full-adder was fabricated by Matsunaga et al. in 2008 [90].

1.3.2.2 Other spin-based logic circuits

Except for the two-terminal and three-terminal MTJ devices, there are many other spintronics devices for logic applications. Here, we introduce two of them that attract much interest and are under intensive investigation by different research groups in the world, domain wall based logic (DWL) and all-spin logic (ASL).

- ***Domain wall based logic (DWL)***

Domain wall logic (DWL) was firstly proposed in [91], the authors use domain wall as the transition edge in a changing signal. A domain wall (DW) is a mobile interface separating two regions of oppositely aligned magnetization. It can be propagated with external magnetic field acting as both clock and power supply. The magnetization in the magnetic nanowire has two opposite directions along with the long axis, representing a binary data either ‘0’ or ‘1’. Different logic functions have been implemented, e.g., NOT gate by using a cusp-shaped planar nanowire and 2-input AND gate as shown in Figure 1.20(a) and (b), respectively. Besides, routing functions such as cross-over (see Figure 1.20(c)) and pass over (see Figure 1.20(d)) can also be developed using magnetic nanowires.

The logic circuits based on field-induced DW motion have the drawbacks of low speed and high power consumption [92]. Current-induced domain wall (CIDW) motion is a solution to overcome these issues. The concept of CIDW motion was proposed by Berger in 1978 [93]. Racetrack memory (RM) based on CIDW motion is shown in Figure 1.21(a) [94]. Data are

stored in a ferromagnetic film strip separated by DWs. A write head aims at writing data into the domain by injecting writing current. A read head away from the write head is used to detect the data stored in the domain above the read head. The domain can be moved (from the write head to the read head) by a steady current flowing along the strip, allowing write and read operations in sequence. Parkin et al. proposed the U-shape memory strip for high density requirement, where read and writes heads are at the bottom (see Figure 1.21(b)) [95].

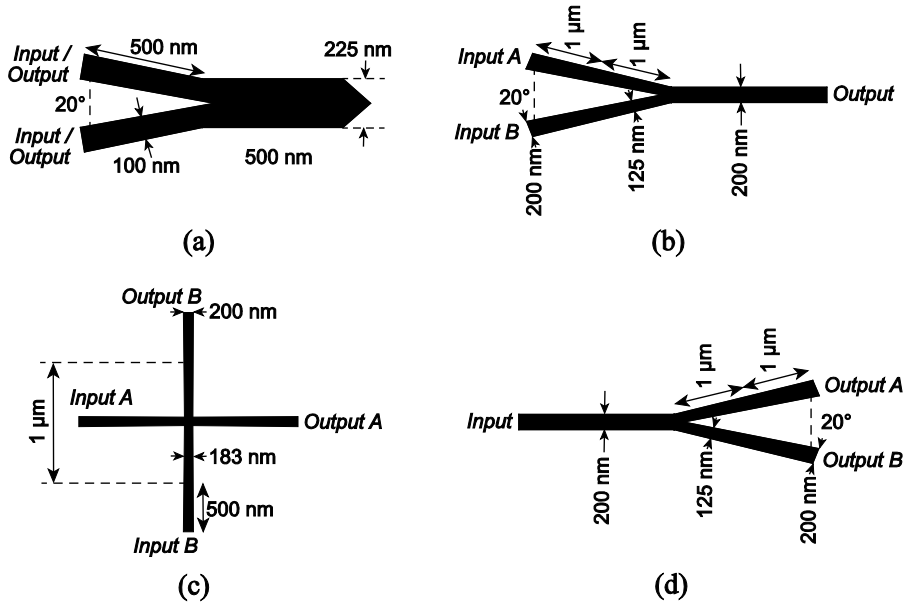


Figure 1.20 Domain wall logic and routing functions (a) NOT gate (b) AND gate (c) cross-over, which allows two signals to pass over each other whiout interference (d) fan-out, which makes two identical copies of an input signal [91]

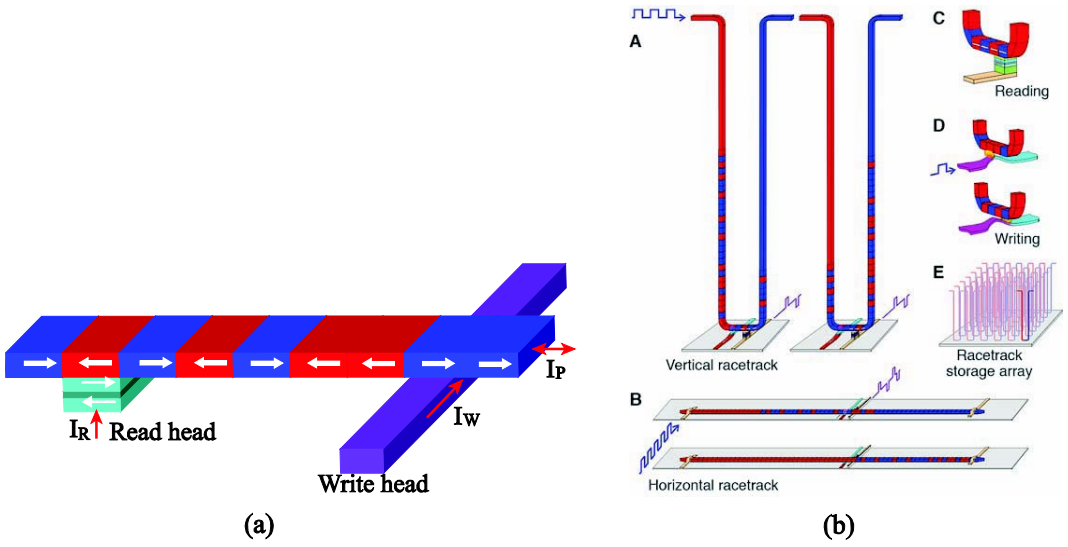


Figure 1.21 (a) Racetrack memory based on current-induced domain wall motion includes a read head, a write head and a magnetic strip. I_W, I_R and I_P represent write current, read current and propagation current for domain move, respectively. (b) Schematic of the U-shape racetrack memory [95].

The first prototype of RM was fabricated at 90 nm technology node by IBM in 2011 [96]. Based on the CIDW motion, reconfigurable logic and full-adder (FA) were realized [97], [98]. As it can be seen in Figure 1.22, all the input and output operands of the FA (A, B and C_i) are stored in the RMs, performing fully non-volatile logic function. Read and write operations are completely irrelevant since they are performed with different read or write head (see Figure 1.22). Moreover, the RMs provide high scalability and they can be fabricated above the CMOS circuit plane. However, it faces the challenges of propagation latency and fabrication to make the strip uniform (e.g., pinning defects).

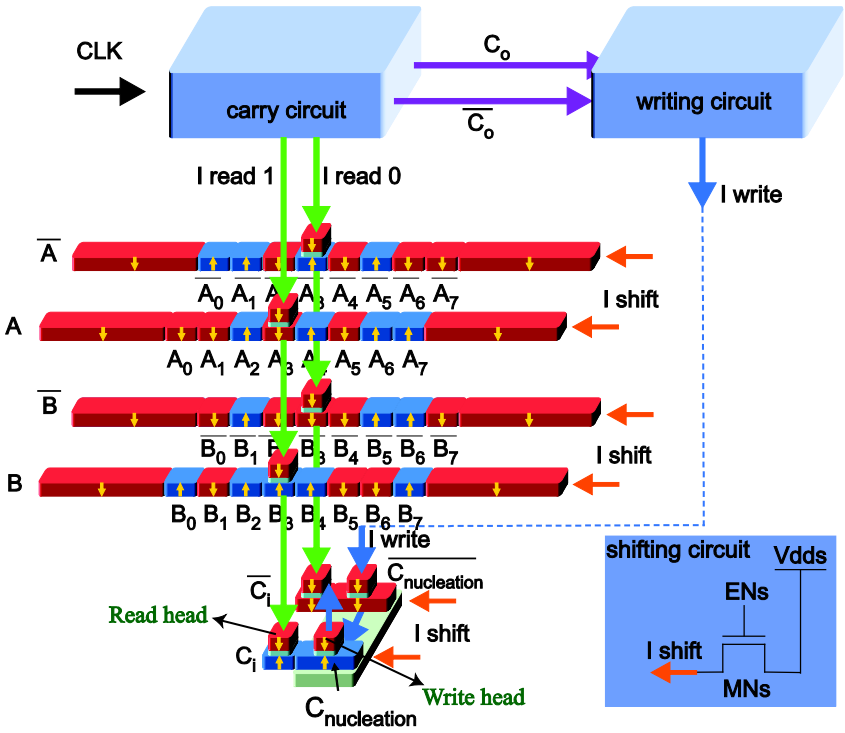


Figure 1.22 Schematic of the magnetic full-adder based on racetrack memory

- ***All-spin logic (ASL)***

All spin logic (ASL) was proposed in 2010 [99]. When considering the basic STT-based ASL device, the left nanomagnet acts as a transmitter (or input nanomagnet) while the right nanomagnet (or output nanomagnet) acts as a receiver (see Figure 1.23(a)). The information is transmitted through a spin channel which connects two nanomagnets. When a voltage is applied to the input nanomagnet, a spin current is generated and travels through the spin channel to the output nanomagnet, generating a torque large enough to switch the right nanomagnet between two stable states.

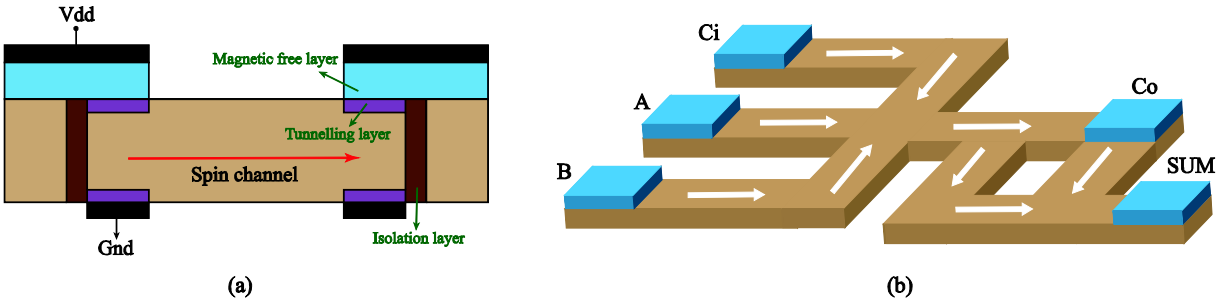


Figure 1.23 (a) All-spin logic (ASL) device (b) Layout of the ASL-based full-adder

Recently, complex logic circuits using multiple ASL devices as inputs have proposed based on spin majority evaluation, for instance, ASL-based full-adder shown in Figure 1.23(b) [100], [101]. ASL has the advantages of non-volatility, high density and low voltage. It also performs low-power operation since no charge current is needed in the information communication. However, there are still many issues to be resolved such as limited spin diffusion length of spin channel and high static power dissipation [102]. Even though the use of “clock” in [103] can largely reduce the static power consumption, it needs additional circuits and energy for clock controlling.

1.4 Conclusion

In this chapter, we reviewed the state-of-the-art of MTJ, which is the basic spintronics device used in our work. TMR ratio is one of the key factors that we will consider in the design of non-volatile logic circuits in Chapter 3. With the progress in the device fabrication, higher TMR ratio can be achieved, leading to reliable sensing against the process variations. The resistance value of MTJ (~ kΩ) is the order of resistance of the CMOS transistors. Furthermore, MTJ devices can be fabricated above the CMOS-based integrated circuits by 3-D back-end process. All these advantages of MTJ make it easier to be integrated with CMOS technology and provide a way towards hybrid MTJ/CMOS design.

We have investigated different writing approaches, their advantages and disadvantages for switching the state of MTJ. In Chapter 2-4, STT is used as the main writing mechanism and SHE is a feasible way to improve the performance (writing speed and power consumption) of non-volatile logic circuits.

LIM as well as its merits in logic design help to understand the work in Chapter 3. Finally, other related works in non-volatile memory and logic design were introduced to complete the state of the art, but they are not used in our approaches.

Chapter 2 Hybrid MTJ/CMOS circuit design

As presented in Chapter 1, spin transfer torque based magnetic tunnel junction with perpendicular magnetic anisotropy (PMA STT-MTJ) is currently considered as one of the most promising spintronics devices for both memory and logic applications. In this chapter, the compact model of PMA STT-MTJ, which is used in the whole circuit design of this thesis, is described and validated. This model enables designers to integrate MTJs in CMOS circuits and easily perform simulations.

Reading and writing circuits are introduced in the second section. They are the basic components for measuring and switching the magnetization configuration of MTJ.

The third section presents the architectural design and a comparative study of multi-context (or multiple bits) hybrid MTJ/CMOS circuit with a particular focus on reliability investigation. The multi-context hybrid circuit includes multiple non-volatile bits forming configuration plane for fast switching between contexts. It provides further area-efficient property owing to the 3-D integration of multiple MTJs above the CMOS logic circuits. Finally, design considerations and strategies are presented to further optimize the reliability performance.

The fourth section reports the design of an embedded MRAM as well as its peripheral circuits. MRAM is a non-volatile random access memory, where data are stored in spintronics devices such as MTJs. Its non-volatility allows the system to be easily powered off in “idle” state, thus the standby power consumption is dramatically reduced. By using the optimized model according to the MTJ technology, simulations are performed to validate its functionality and evaluate its performances.

2.1 Compact model of STT-based MTJ with perpendicular magnetic anisotropy (PMA STT-MTJ)

In order to design novel memory and logic circuits based on hybrid MTJ/CMOS technology, STT-MTJ spice-compatible model is necessary. A CoFeB/MgO/CoFeB PMA STT-MTJ compact model was developed by the NANOARCHI group at IEF (Institut d'Electronique Fondamentale) based on the in-depth understanding of fundamental physical mechanisms and experimental measurements [104], [105]. It provides a feasible way to integrate the MTJ signals with CMOS circuits, allowing to perform electrical simulation of the STT-MTJ based hybrid circuits to validate their functionalities and performances such as speed, power consumption, etc. Figure 2.1 illustrates an example MTJ stack, where CoFe/MgO/CoFe are the three main layers. The bottom CeFeB layer is deposited on a Ta/Ru/Ta buffer layer while the top CeFeB layer is deposited on a Ta buffer layer.

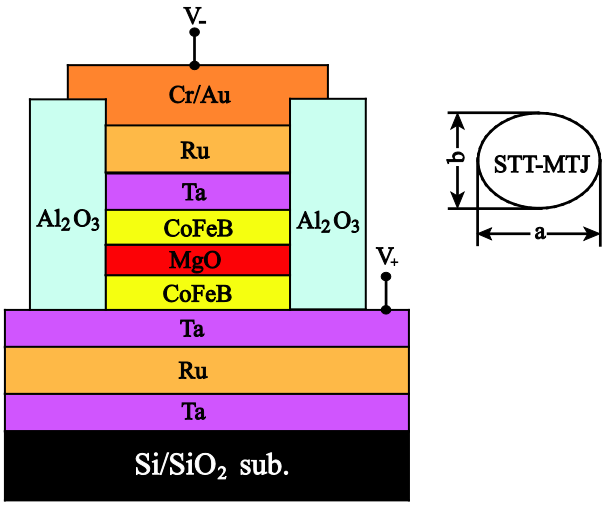


Figure 2.1 Vertical structure of the PMA STT-MTJ stack [58]

The main physical models integrated in the compact model are: 1) the MgO barrier tunnel resistance model; 2) the TMR model depending on the bias voltage; 3) the STT switching models including static model for calculating the critical current, dynamic model for calculating the switching time as well as stochastic model. Verilog-A is chosen to be the modeling language thanks to its compatibility with standard CAD tools like Cadence and programing flexibility. In order to meet different design requirements, designers can easily change the variable parameters in the prospective of flexible hybrid MTJ/CMOS circuit design. This model will be validated in Section 2.1.3 through Direct Current (DC) simulation, transient simulation as well as Monte-Carlo (MC) simulation.

2.1.1 Physical models of PMA STT-MTJ

2.1.1.1 MgO barrier tunnel resistance model

The simplified resistance equation is obtained from the first physical model of tunneling conductance proposed by Brinkman [106].

$$\frac{G(V)}{G(0)} = 1 - (\frac{A_0 \triangle \varphi}{16 \bar{\varphi}^{3/2}}) eV + (\frac{9}{128} \frac{A_0^{\,2}}{\bar{\varphi}})(eV)^2 \qquad \text{Eq. 2.1}$$

where $G(0)$ is the conductance with zero bias voltage, V is the bias voltage, e is the electron charge, $\bar{\varphi}$ is the potential energy barrier height (0.4 for MgO [31]), $\triangle \varphi = \varphi_2 - \varphi_1$, which should be 0 as the oxide barrier is symmetric. $A_0 = 4(2\,\mathrm{m})^{1/2}\,\mathrm{t}_{ox}/3\hbar$, m is the electron mass, $\hbar$ is the Planck constant, t_{ox} is the thickness of oxide barrier.

The low resistance of MTJ in the parallel state can be expressed as follows:

$$R_P = R_L = \frac{t_{ox}}{332.2 \times \bar{\varphi}^{1/2} \times Area} \times \exp(1.025 \times t_{ox} \times \bar{\varphi}^{1/2}) \qquad \text{Eq. 2.2}$$

where $Area$ is the area of MTJ, which depends on the MTJ shape. And for this model, there are three shapes (i.e., square, ellipse and round) to be chosen by designers according to the technology requirements, which will be detailed in Section 2.1.2.

2.1.1.2 TMR model

The tunnel magnetoresistance (TMR) ratio is an importance factor who determines the speed as well as margin for detecting the state of MTJ. As described in Section 1.2.1, TMR ratio is defined as $TMR = (R_{AP} - R_P)/R_P$, characterizing the amplitude of MTJ resistance change. It is found that the real TMR ratio is not a constant but strongly depends on the bias voltage. According to the theory in [107], the real value of TMR ratio is:

$$TMR_{real} = \frac{TMR(0)}{1 + \frac{V_{bias}^2}{V_h^2}} \qquad \text{Eq. 2.3}$$

where $TMR(0)$ is the TMR ratio with zero bias voltage. V_h is the bias voltage when

TMR_{real} is equals to 0.5×TMR(0).

The high resistance of the MTJ in anti-parallel state can be defined as:

$$R_{AP} = R_H = R_P \times (1 + TMR_{real}) \quad \text{Eq. 2.4}$$

2.1.1.3 Static model of STT switching mechanism

As mentioned in Chapter 1, the relative magnetization directions of two ferromagnetic layers will change from parallel to antiparallel or vice versa only when the current driven by STT effect exceeds the critical current, denoted as I_{C0}. The static behavior of STT switching mechanism mainly relies on the calculation of the critical current, as shown in Eq. 2.5 [58]:

$$I_{C0} = \alpha \frac{\gamma e}{\mu_B g}(\mu_0 M_S) H_K V = J_{C0} \times Area = 2\alpha \frac{\gamma e}{\mu_B g} E \quad \text{Eq. 2.5}$$

where α is the magnetic damping constant, γ is the gyromagnetic ratio, μ_B is the Bohr magneton, M_S saturation magnetization, H_K is the perpendicular magnetic anisotropy, V is the volume of the free layer, J_{C0} is the critical current density, E is the barrier energy, g is the spin polarization efficiency factor.

Since the switching of MTJ is determined by the critical current, the smaller the size of MTJ is, the smaller the writing current required for changing the magnetization direction of the free layer. It should be noted that the critical current needed to switch the MTJ from parallel state to anti-parallel state ($\mathrm{P} \rightarrow \mathrm{AP}$) is different from that to switch from anti-parallel state to parallel state ($\mathrm{AP} \rightarrow \mathrm{P}$), due to different spin polarization efficiency factor in two cases. Thus, the spin polarization efficiency factor g can be obtained with the following equations to describe the asymmetric critical current [108]:

$$g = g_{sv} \pm g_{tunnel} \quad \text{Eq. 2.6}$$

$$g_{sv} = [-4 + \frac{(P^{-1/2} + P^{1/2})^3 (3 + \cos\theta)}{4}]^{-1} \quad \text{Eq. 2.7}$$

$$g_{tunnel} = \frac{P}{2(1 + P^2 \cos\theta)} \quad \text{Eq. 2.8}$$

where g_{sv} and g_{tunnel} are the spin polarization efficiency values in a spin valve and tunnel junction nanopillars, which are predicted by Slonczewski. P is the spin polarization percentage of the tunnel current, θ is the angle between the magnetization of the two ferromagnetic layers (i.e. free layer and reference layer) [43], [109].

With the progress of technology, measuring the asymmetric critical current for STT switching is less and less obvious from experiments. One recent experiment conducted by IBM showed the same g value for both switching cases electrodes [110], which can be described as follows:

$$g = \frac{\left[TMR \times (TMR + 2)\right]^{1/2}}{2(TMR + 1)} \qquad \text{Eq. 2.9}$$

In this model for hybrid MTJ/CMOS circuit design, the spin polarization efficiency factor from Eq. 2.5 is defined by Eq. 2.9.

2.1.1.4 Dynamic model of STT switching mechanism

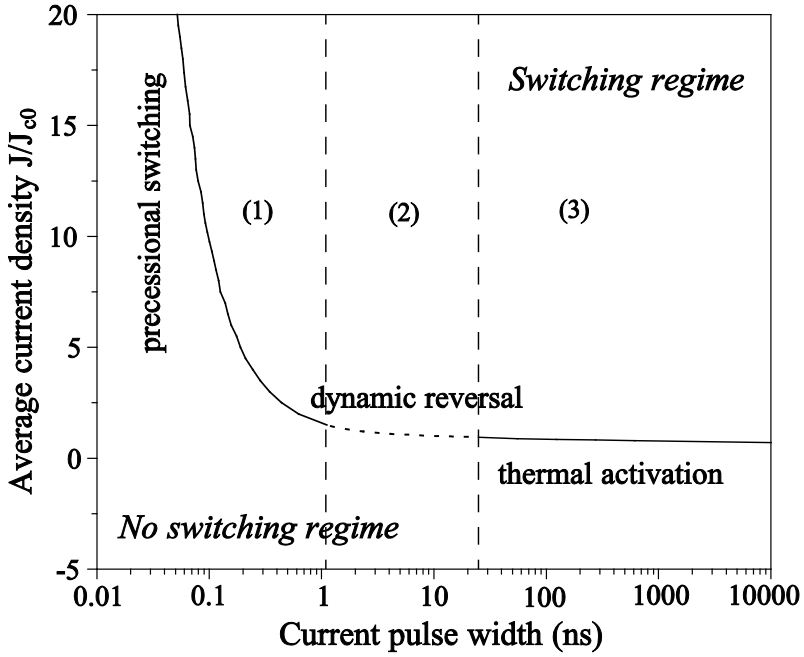


Figure 2.2 Phase diagram of MTJ switching driven by spin-transfer torque (STT) [111]

The dynamic model of STT switching mechanism studies the influence of switching current (or writing current I_{write}) on the switching duration τ. Based on the relationship between the switching current I_{write} and the critical current I_{C0}, the complex switching process can be

divided into three regimes, as can be seen in Figure 2.2, including: (1) precessional switching regime when $I_{write} > I_{C0}$ with short switching delay, (2) dynamic reversal regime when $0.8I_{C0} < I_{write} < I_{C0}$, (3) thermal activation regime when $I_{write} < 0.8I_{C0}$ with relative long switching delay [111].

The dynamic reversal regime (i.e., regime (2) presented by a dotted line in Figure 2.2 is not integrated in this model due to a lack of clear theories and experimental results related to the range from $0.8I_{C0}$ to I_{C0}. The average switching durations of the other two regimes can be expressed by the Néel-Brown model [112], [113] (for thermal activation regime) and the Sun model [114] (for precessional switching regime), shown in Eq. 2.10 and Eq. 2.11.

- Néel-Brown model for thermal activation regime ($I_{write} < 0.8I_{C0}$):

$$\tau = \tau_0 \exp(\frac{E}{k_B T}(1-\frac{I_{write}}{I_{C0}})) \qquad \text{Eq. 2.10}$$

where τ_0 is the attempt period. k_B is the Boltzmann constant, T is the temperature.

- Sum model for precessional switching regime ($I_{write} > I_{C0}$):

$$<\tau>=\frac{C+\ln(\frac{\pi^2\xi}{4})}{2}\times\frac{em_m(1+P^2)}{\mu_B P}\times\frac{1}{I_{write}-I_{C0}} \qquad \text{Eq. 2.11}$$

where $C \approx 0.577$ is the Euler's constant, $\xi = E / k_B T$ is the thermal stability factor, m_m is the magnetic moment of free FM layer, P is the tunneling spin polarization percentage.

From the above equations, we can find that if the writing current I_{write} increases or the critical current I_{C0} decreases, the switching time is reduced. When switching the state of MTJ by CMOS-based circuit, large transistors are usually necessary to ensure high switching current, which results in high area overhead. This trade-off between the writing speed and the circuit size will be discussed in Section 2.2.2.

2.1.2 Spice model of PMA STT-MTJ

Verilog-A [115] is chosen as the modeling language to create an interface between the physical models and the electrical simulators (e.g., Spectre [116], Eldo [117], etc. under the

platform of Cadence). Figure 2.3 summarizes all the aforementioned physical models integrated in the PMA STT-MTJ compact model in order to obtain the following outputs: the resistances of MTJ in two states (parallel R_P or anti-parallel R_{AP}), the critical current I_{C0} and the switching duration τ. MgO barrier tunnel resistance model and TMR model are used to calculate R_P and R_{AP} with a input bias voltage V_{bias}. The static and dynamic models determine I_{C0} and τ. Only when the current pulse generated by the CMOS-based writing circuit exceeds I_{C0} being larger than τ, the state of MTJ can be changed.

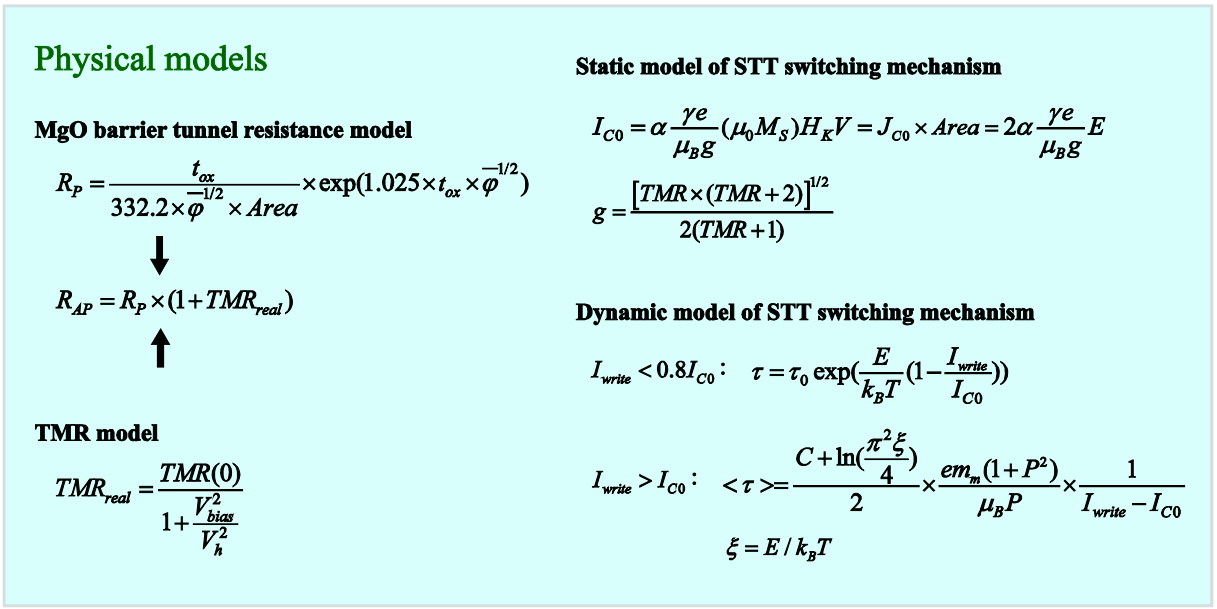


Figure 2.3 Physical models integrated in the PMA STT-MTJ model

There are three different types of parameters including general constants such as electron mass (m), MTJ technology parameters such as out of plane magnetic anisotropy (H_K) and device parameters such as the size of MTJ, which are listed Table 2.1. When we design the hybrid MTJ/CMOS circuits by using this model, some parameters ($Area(a,b)$, $TMR(0)$, t_{ox}, t_f) are changeable through the graphical user interface (i.e., the Edit Component CDF form) according to different applications. Other parameters are unchangeable. It should be noted the shape of MTJ can be chosen from “Square”, “Ellipse” and “Round”. The area of MTJ can be calculated with the parameters a and b, i.e., $Area = a \times b$ for square MTJ, $Area = \pi \times a \times b / 4$ for elliptical MTJ and $Area = r^2$(or a^2) for round MTJ.

Besides, the process variations (TMR, t_{ox}, t_f) and stochastic behavior (τ) of MTJ are also integrated in this model by using the random functions (e.g., “$rdist_uniform” for uniform

distribution, "$rdist_normal" for normal distribution) and statistical block. The source files including the source code of the PMA STT-MTJ model can be easily downloaded and then used for hybrid circuit design and simulation from [118].

Table 2.1 Parameters in the STT-MTJ model

Parameters	Descriptions	Default values	Unit
General constants			
e	Elementary charge	1.6×10^{-19}	C
m	Electron mass	9.1×10^{-31}	Kg
K_B	Boltzmann constant	1.38×10^{-23}	J/K
μ_B	Bohr magneton constant	9.27×10^{-28}	J/Oe
$\hbar$	Plank's constant	1.0545×10^{-34}	J·s
μ_0	Permeability of free space	1.25663×10^{-6}	H/m
C	Euler's constant	0.577	
MTJ technology parameters			
T	Temperature	300	K
α	Gilbert damping coefficient	0.027	
γ	Gwyromagnetic constant	1.76×10^{7}	Hz/Oe
P	Electron polarization percentage	0.52	
H_K	Out of plane magnetic anisotropy	1433	Oe
M_S	Saturation field in the free layer	15800	Oe
$\overline{\varphi}$	Oxide layer energy barrier height	0.4	eV
V_h	Voltage bias when TMR_{real} equals to 0.5TMR(0)	0.5	V
τ_0	Attempt period	0.87	ns
RA	Resistance area product	5 (5-15)	$\Omega\cdot\mu m^2$
Device parameters			
SHAPE	Shape of MTJ	Ellipse	
a	Length of MTJ	40	nm
b	Width of MTJ	40	nm
t_{ox}	Thickness of the oxide layer	0.85 (0.6-1.2)	nm
t_f	Thickness of the free layer	1.3 (0.8-2)	nm
TMR(0)	TMR value with zero volt bias voltage	150% (50%-600%)	

Figure 2.4 shows the MTJ symbol with a top pin $T1$ connected with the reference FM layer and a bottom pin $T2$ connected with the free FM layer. The virtual pin *State*, which is not a real pin for a two-terminal MTJ device, is used to identify the magnetization configuration of MTJ by connecting a resistance of 1 Ω. The output of this pin (V_{state}) will be 0 V (or 1 V) if the MTJ is in parallel (or anti-parallel) state. When a current (indicated by the red arrows) passing through the MTJ exceeds the critical current, the state of MTJ will change from parallel (P) to anti-parallel (AP) if the current flows from the top (from $T1$ to $T2$), or from AP to P if the current flows from the bottom (from $T2$ to $T1$).

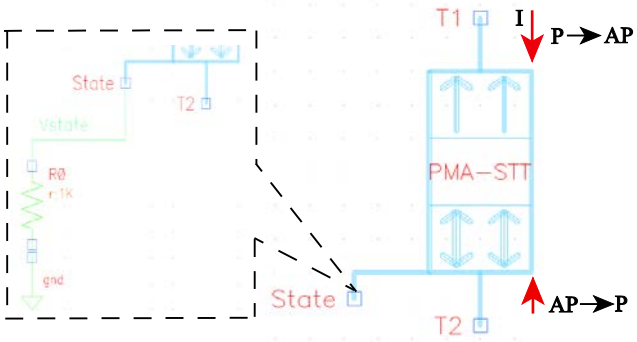


Figure 2.4 Symbol of the MTJ model

2.1.3 Simulation of the PMA STT-MTJ model

Different simulations are performed to validate the behaviors (static, dynamic and stochastic) of the MTJ model, including DC simulation, transient simulation and MC simulation. The parameters of the MTJ used for simulations are listed in Table 2.1. Figure 2.5 illustrates the simulation framework from the integration of MTJ model to the output of simulation results.

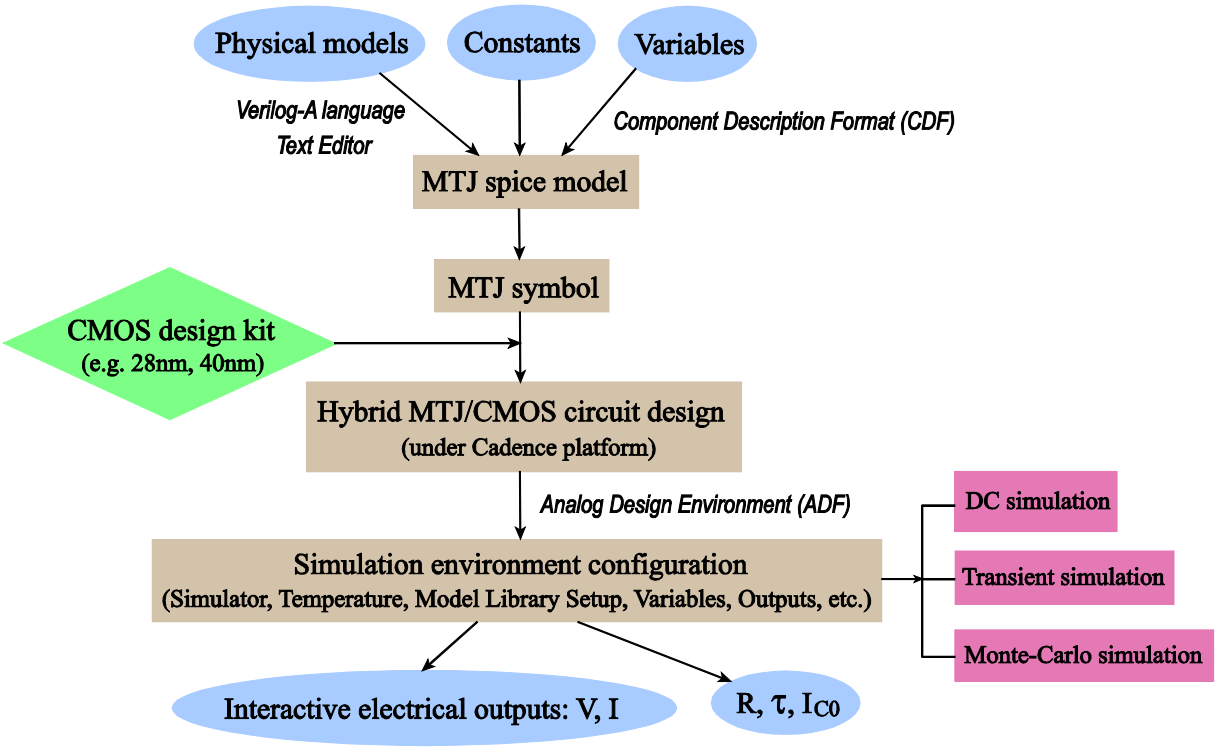
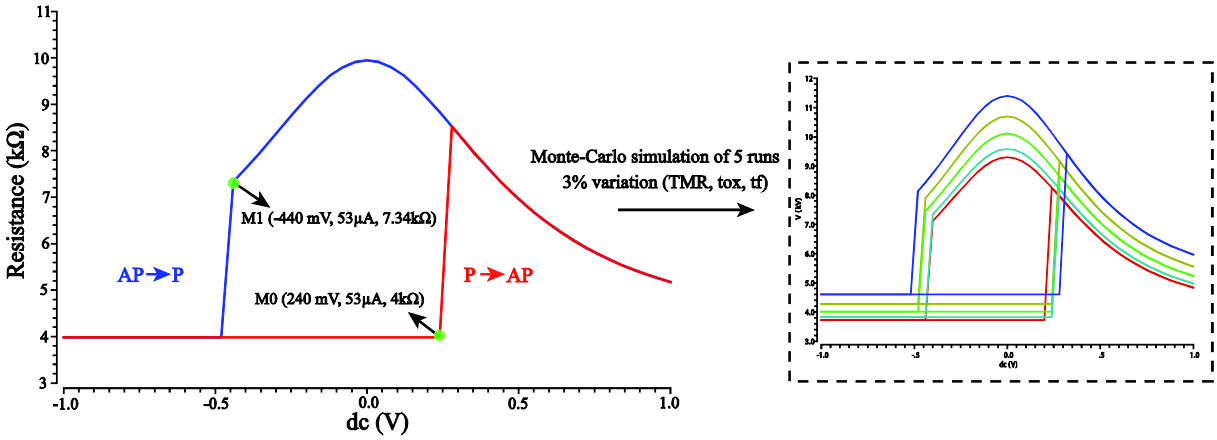


Figure 2.5 Simulation framework

Figure 2.6(a) shows the DC simulation of a simple MTJ with a supply voltage varying between -1 V and 1 V. The critical current for STT writing (I_{C0}) is estimated to be ~50 µA.

The state of MTJ is switched from parallel (P) to anti-parallel (AP) at M_0, and from AP to P at M_1. Figure 2.6(b) shows the MC simulation waveforms, taking into account 3% process variations of the MTJ's key parameters such as TMR ratio, oxide barrier thickness and free layer thickness. As an example, 5 runs of MC simulation are performed to quickly show that the process variations can result in the fluctuation of MTJ resistance and critical switching current, which is important and should be considered for the reliability analysis of hybrid MTJ/CMOS circuits.



(a) (b)

Figure 2.6 (a) DC simulation of the MTJ model (b) Monte-Carlo simulation model with 3% variation of TMR, t_{ox}, t_f following normal distribution

Figure 2.7 shows the transient simulation to validate the dynamic model. A voltage pulse V (from -1 V to 1V) is applied to generate a bi-directional current I. The state of MTJ is switched from parallel to anti-parallel, and then back to parallel with a certain writing delay ($t_{P\to AP} \approx 1.1\,ns, t_{AP\to P} \approx 1.6\,ns$). The writing time $t_{P\to AP}$ and $t_{AP\to P}$ are different because the resistances of MTJ in two states are different, resulting in different writing current.

MC simulations are performed in which the parameters of MTJ (i.e., TMR ratio, t_{ox} and t_f, see Figure 2.8(a)) or the switching duration τ (see Figure 2.8(b)) follow normal distribution with 3% variation. As shown in Figure 2.8(a), only the parameter variations are taken into consideration, which results in resistance variation and hence the writing current I_{write} is variable. According to Eq. 2.10 and Eq. 2.11, τ is inversely proportional to I_{write}. Therefore, τ is no longer constant. Figure 2.8(b) shows the stochastic behavior of the MTJ model. The switching duration τ is distributed around the average delay calculated by Eq. 2.11.

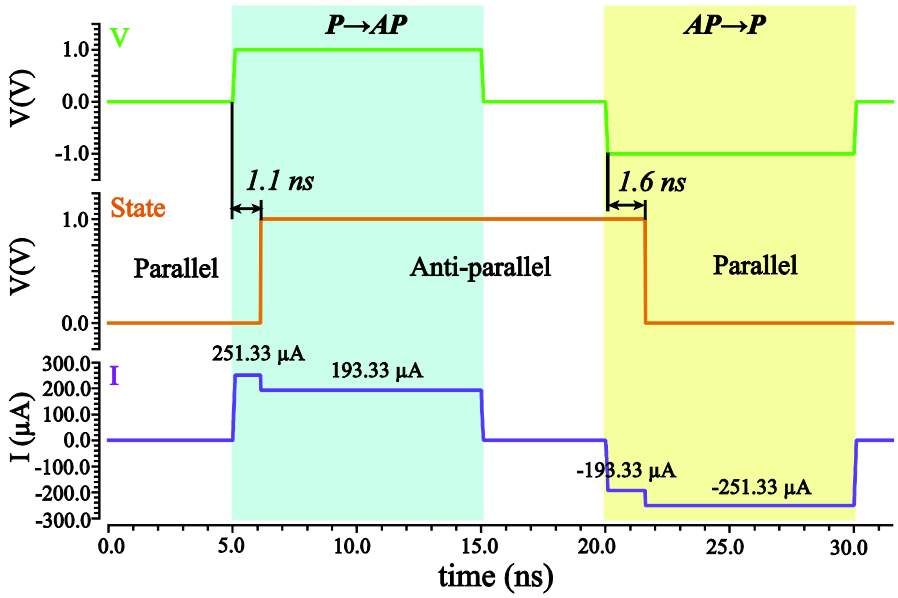
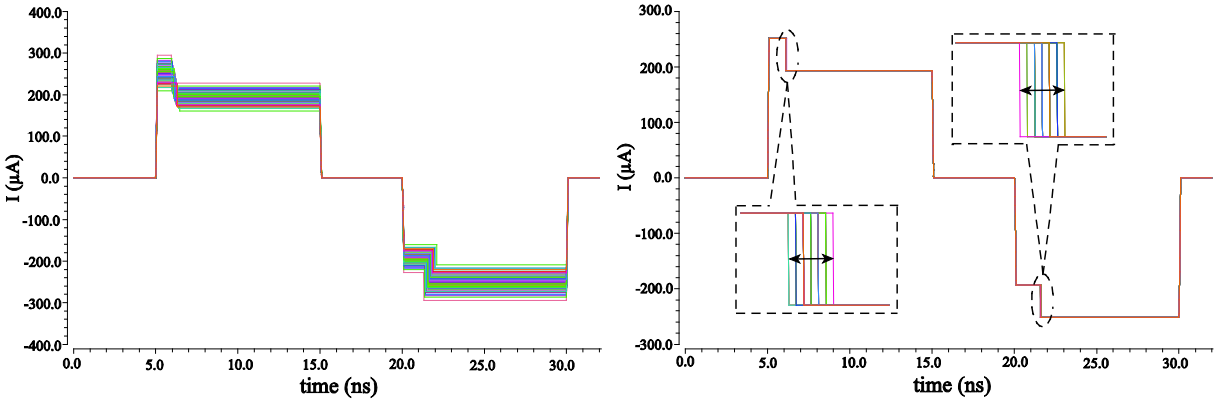


Figure 2.7 Transient simulation of the MTJ model



(a) (b)

Figure 2.8 Monte-Carlo simulation (100 runs) of STT writing operation with (a) process variations of parameters including TMR, t_{ox}, t_f (b) stochastic behaviors

In this section, the PMA STT-MTJ compact model was introduced and validated. We then integrate it with CMOS circuits to perform reading and writing operations, which is the basis of the hybrid MTJ/CMOS memory and logic circuit design.

2.2 MTJ reading and writing circuits

2.2.1 MTJ reading circuit

As mentioned in Chapter 1, an MTJ can be considered as a two-value resistor with low resistance (R_P) for parallel state or high resistance (R_{AP}) for anti-parallel state due to the TMR effect. This characteristic allows MTJs to be embedded into a current-mode sense amplifier [119] that detects the MTJs' magnetic configurations and amplifies them to logic outputs. It was confirmed in [23] that pre-charge sense amplifier (PCSA) performs the best sensing speed, power consumption, area overhead and reliability, when comparing to other current-mode sense amplifiers. Consequently, PCSA is used as the reading circuit in our hybrid MTJ/CMOS circuit design.

2.2.1.1 Structure of the reading circuit

Figure 2.9 shows the schematic of the MTJ reading circuit based on current-mode sense amplifier (PCSA). The seven-transistor (7T) circuit consists of a pre-charge sub-circuit (transistors P_{0-1}), a discharge transistor (transistor N_2) and a pair of cross-coupled inverters (transistors P_2 / N_0 and P_3 / N_1). Two MTJs in complementary sates are placed in two branches of the sense amplifier and store binary data. Data stored in MTJs is detected and amplified at output nodes Qm and $\overline{Qm}$.

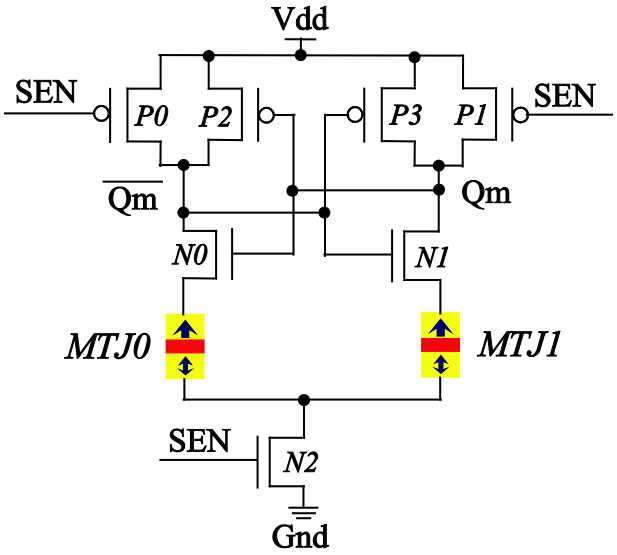


Figure 2.9 Schematic of the pre-charge sense amplifier (PCSA) for detecting the configurations of the embedded MTJs and amplifying to logic signals

The reading circuit is symmetric except that the resistances of $MTJ0$ and $MTJ1$ are different. The resistance difference between the left and right branches can be expressed by Eq. 2.12, which determines the sensing margin.

$$\Delta R = \left| R_L - R_R \right| = R_{AP} - R_P = R_P \times TMR \qquad \text{Eq. 2.12}$$

According to the value of control signal SEN, the PCSA circuit operates in two phases: pre-charge phase and evaluation phase.

- *Pre-charge phase* ($SEN = '0'$): Both nodes Qm and $\overline{Qm}$ are pulled up to Vdd through PMOS transistors P_{0-1}. No current flows between V_{dd} and the ground since the discharge transistor N_2 remains OFF (see Figure 2.10 (a)).

- *Evaluation phase* ($SEN = '1'$): N_2 is turned ON, enabling the reading current (I_0 and I_1) to pass through both MTJs (see Figure 2.10 (b)). Qm and $\overline{Qm}$ begin to discharge at different speeds due to the resistance different between $MTJ0$ and $MTJ1$. We assume that MTJ_0 and MTJ_1 are respectively initialized to parallel state and anti-parallel state, and hence the resistance of the left branch is smaller than that of the right one ($R_L < R_R$). In this case, I_0 is larger than I_1, $\overline{Qm}$ reaches more quickly the threshold voltage of the PMOS transistor P_3 than Qm. Then Qm will be pulled up to V_{dd} or logic '1', while $\overline{Qm}$ will continue to discharge to Gnd or logic '0' (see Figure 2.10 (c)).

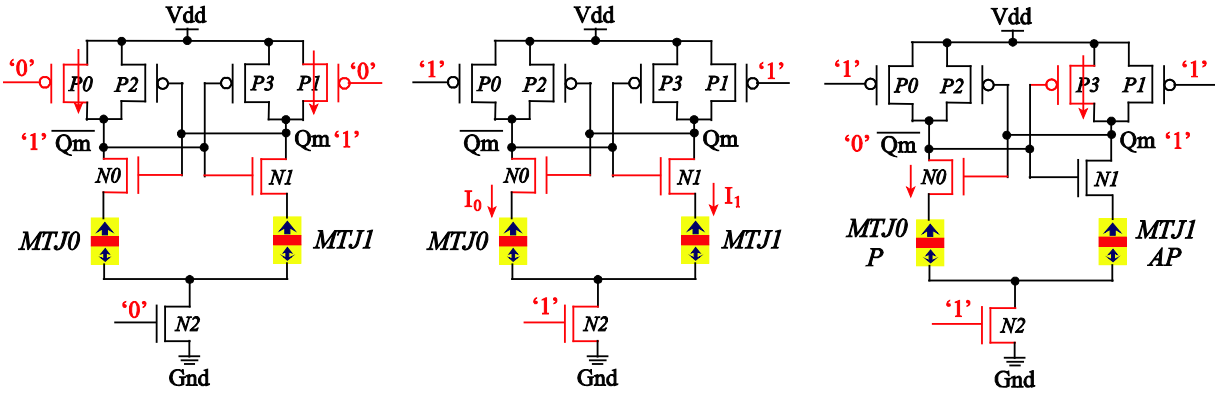


Figure 2.10 Three states for the sensing operation of the PCSA-based reading circuit

We can conclude that a part of the energy is stored in the node capacitors (Qm and $\overline{Qm}$) during the pre-charge phase and is dissipated to the ground during the evaluation phase. Another part of the energy is dissipated for pulling one node to V_{dd} at the end of the sensing operation.

2.2.1.2 Simulation and performance analysis of the reading circuit

Figure 2.11 shows the simulation of one sensing operation with MTJ_0 in parallel state and MTJ_1 in anti-parallel state. All the transistors are kept in the minimum size (80 nm×30 nm at CMOS 28 nm technology). Other MTJ parameters are listed in Table 2.1. Before T_0, this circuit operates in pre-charge phase. Both output nodes Qm and $\overline{Qm}$ are pull up to V_{dd} (here in the simulation, it is 1 V) or logic '1'. The evaluation phase starts at T_0 with SEN switching from 0 V to 1 V. It can be seen that $\overline{Qm}$ reaches the threshold voltage faster than Qm at the time T_1, opening the transistor P_3 (see Figure 2.9) and pulling Qm to V_{dd}. During the period from T_1 to T_2, the voltage of Qm rises from 505.92 mV to 900 mV (0.9 V_{dd}) while $\overline{Qm}$ continues to discharge until 53.2 mV. After T_2, the output nodes stays in stable logic states ($Qm = '1'$ and $\overline{Qm} = '0'$).

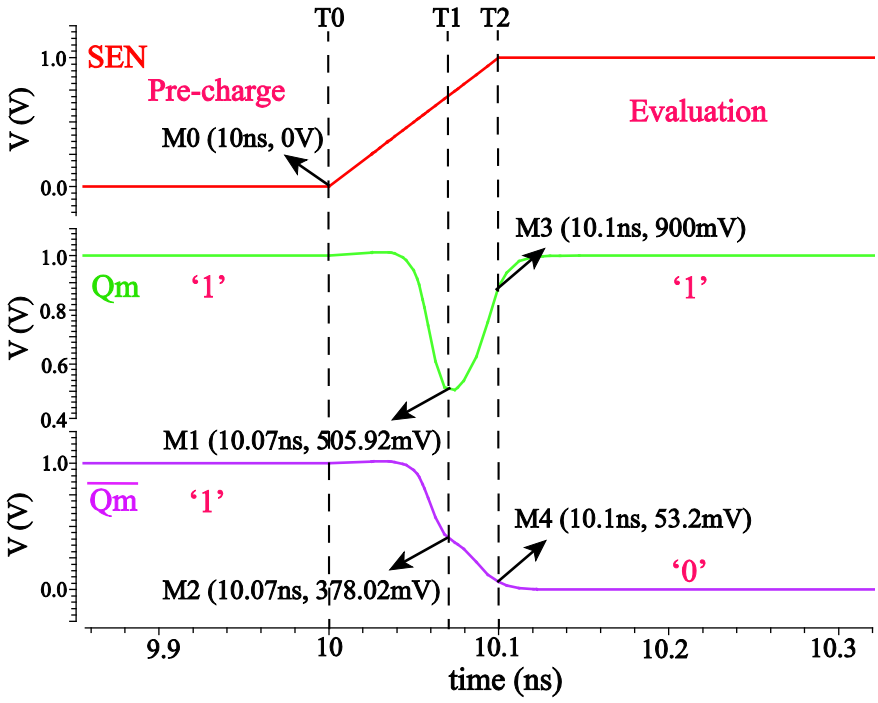


Figure 2.11 Simulation of the PCSA-based reading circuit

More simulated results show that the PCSA-based sensing circuit has a reading delay smaller than 200 ps and the energy dissipation as low as almost 2 fJ with a SEN frequency of 500 MHz. The reading time can be further reduced by increasing the size of transistors or the TMR ratio of MTJ. The advantages of high speed and low power make the PCSA circuit suitable for hybrid MTJ/CMOS circuit design. Moreover, thanks to the dynamic sensing and small currents passing through MTJs ($I_{0\_pick} \approx 8.1\,\mu A$ and $I_{1\_pick} \approx 3.3\,\mu A$), which are much lower than the switching critical current (~50 μA), an erroneous writing during sensing operation can be avoided.

2.2.1.3 Reliability analysis of the reading circuit

The PCSA circuit greatly reduces the chip failure thanks to the low reading currents and short sensing delay [120]. However, it is still sensitive to variations of CMOS process and MTJ process. For instance, with $TMR = 150\%$ and all CMOS transistors in minimum size, 23 errors (which means that Qm is different from the sensing result shown in Figure 2.11) have been observed through the MC simulation of 100 runs (see Figure 2.12). Unlike classic memory circuit design, where complex error correction blocks (ECB) can be easily employed [121], it is rather difficult to embed ECB in logic designs while keeping high speed, low area and power efficiency. Therefore, different optimizations should be investigated to meet the requirement of nearly "zero" errors in the non-volatile logic circuits for practical applications.

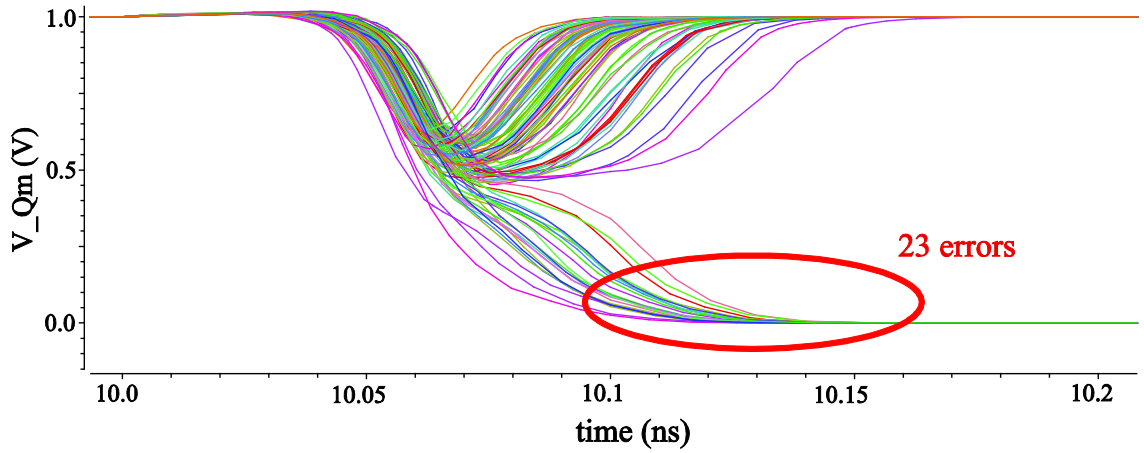


Figure 2.12 Monte-Carlo simulation of PCSA-based reading circuit (100 runs)

The most efficient methods to improve the reliability of the reading circuit are:

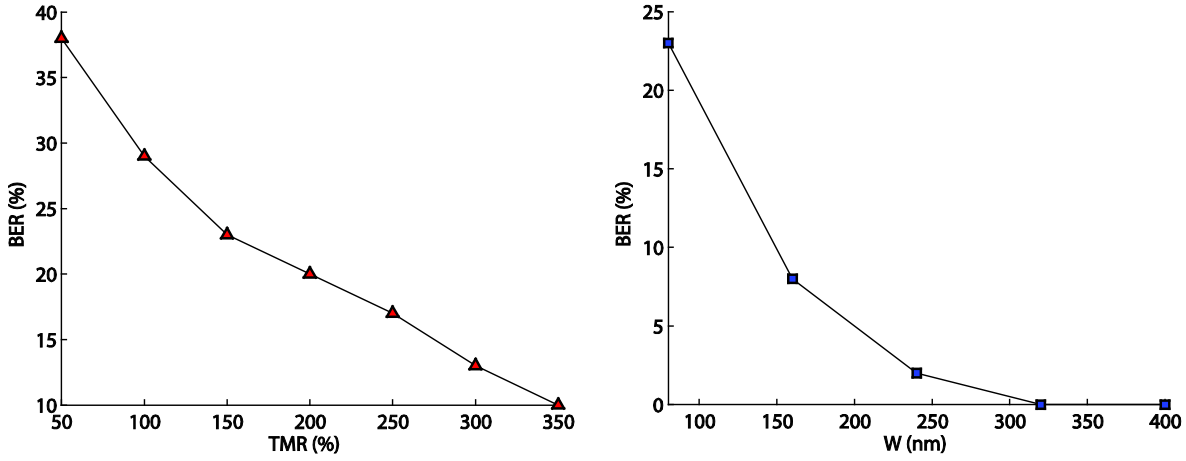
1) *Increasing the value of TMR ratio*. From Eq. 2.4 and Eq. 2.12, we can tell that limited TMR ratio results in low resistance and small sensing margin of the PCSA circuit.

Therefore, larger TMR ratio is required for reliable sensing. With the recent progress, TMR ratio can reach more than 600% with MgO barrier at room temperature [4]. As shown in Figure 2.13, the bit error rate (BER) decreases from 38% to 10% for a TMR ratio increasing from 50% to 350%. Here, BER represents the error percentage of a circuit when performing the MC simulations, which can be expressed as:

$$BER = \frac{N_{error}}{N_{simu}} \qquad \text{Eq. 2.13}$$

where N_{error} is the number of output errors when the sense amplifier is in stable logic state, N_{simu} is the total number of MC simulation runs (100 in Figure 2.13).

2) *Increasing the width of CMOS transistors*. As discussed above, PCSA-based reading circuit achieves ultra-low currents to avoid erroneous writing. However, this also leads to low sensing margin, denoted as $\Delta I = \left| I_0 - I_1 \right|$, and relatively high sensing errors. By increasing the transistor size, the resistances in two branches can be decreased, and in turn the sensing margin can be increased at the expense of more area overhead. As shown in Figure 2.13, the output errors is less than 1% when the width of transistors is four times (4X) larger than the minimum size ($W = 80nm$).



(a) (b)

Figure 2.13 Bit error rate (BER) with respect to (a) the TMR ratio (b) the width of the transistors in the PCSA-based reading circuit

Other methods for reducing the sensing errors of multi-bit hybrid MTJ/CMOS circuit will be introduced in Section 2.3.3.

2.2.2 MTJ writing circuit

In the PCSA-based reading circuit, two MTJs are always in different states. Therefore, they should both be switched to the opposite state when reconfiguring the non-volatile data. In order to realize the writing operation, a bi-directional writing current I_{write} should be generated by a CMOS-based circuit. In the following paragraphs, we will present in details two writing structures as well as the logic gate implementation for controlling the direction of writing current.

2.2.2.1 Structures of the writing circuit

The four-transistor (4T) writing circuit is illustrated in Figure 2.14(a). It is mainly composed of two PMOS transistors and two NMOS transistors, i.e., P_{0-1} and N_{0-1}. Two MTJs are serially connected by the electrodes $T2$. During the writing operation, only one PMOS transistor (e.g., P_0) and one NMOS transistor (e.g., N_1) will be open, generating a writing current passing from V_{dda} to the ground.

The six-transistor (6T) writing circuit is illustrated in Figure 2.14(b). It has three PMOS transistors and three NMOS transistors, i.e., P_{0-2} and N_{0-2}. The electrode $T2$ of MTJ_0 is connected with the electrode $T1$ of MTJ_1. When performing the writing operation, P_{0-1} and N_2 are turned ON while N_{0-1} and P_2 are turned OFF, or vice versa.

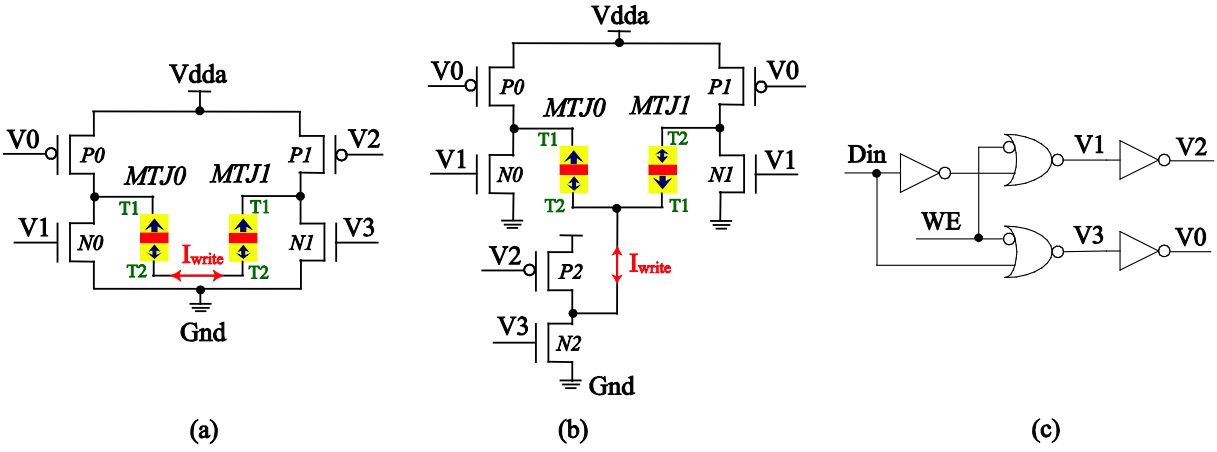


Figure 2.14 (a) 4T writing circuit (b) 6T writing circuit (c) Logic gate part for controlling the activation and the direction of writing current

As mentioned above, the direction of the writing current is controlled by opening the corresponding transistors through four bias voltages V_{0-3}. In order to simplify the design,

these voltages are generated by two signals and a logic gate part, containing three NOT logic gates and two NOR logic gates, as shown in Figure 2.14(c). WE is the activation signal and Din determines the direction of the writing current. The corresponding truth table is shown in Table 2.2. For both 4T and 6T writing circuits, MTJ_0 and MTJ_1 are switched to:

- anti-parallel state and parallel state if $V_0 = V_1 = '0'$ and $V_2 = V_3 = '1'$;
- parallel state and anti-parallel state if $V_0 = V_1 = '1'$ and $V_2 = V_3 = '0'$.

There are three combinations of signals WE and Din:

- If $WE = '0'$, all the transistors are closed and there is no current passing through the MTJs regardless the value of Din.
- If $WE = '1'$ and $Din = '0'$, V_{0-1} will be '0' and V_{2-3} will be '1'. For the 4T writing circuit, only P_0 and N_1 are open, creating a current flows from the top of MTJ_0 to the top of MTJ_1. For the 6T writing circuit, P_{0-1} and N_2 are open while other transistors are closed. A current passing from $T1$ to $T2$ of MTJ_0 and another current passing from $T2$ to $T1$ of MTJ_1 will be created. After a certain delay, MTJ_0 will be switched to anti-parallel state and MTJ_1 will be parallel state.
- If $WE = '1'$ and $Din = '1'$, a reverse current will be generated MTJ_0 and MTJ_1 in both writing circuits will be switched to parallel and anti-parallel state, respectively.

Table 2.2 Operation mechanism of the full writing circuit

Inputs		**Intermediate signals**				**MTJ state**	
WE	**Din**	**V_0**	**V_1**	**V_2**	**V_3**	**MTJ_0**	**MTJ_1**
0	×	1	0	1	0	–	–
1	*0*	0	0	1	1	AP	P
1	*1*	1	1	0	0	P	AP

2.2.2.2 Simulation and performance analysis of the writing circuits

We simulate the writing circuit with a supply voltage V_{dda} of 1.2 V. All the three cases listed in Table 2.2 are included in Figure 2.15. Switching takes place only when $WE = '1'$ and Din controls the switching from P to AP or from AP to P. When comparing the two writing

structures, the 4T writing circuit has simpler structure with two less CMOS transistors. However, the 6T writing circuit can be integrated for energy saving since it generates larger writing current and smaller writing time than those of the 4T writing circuit when keeping the same circuit area.

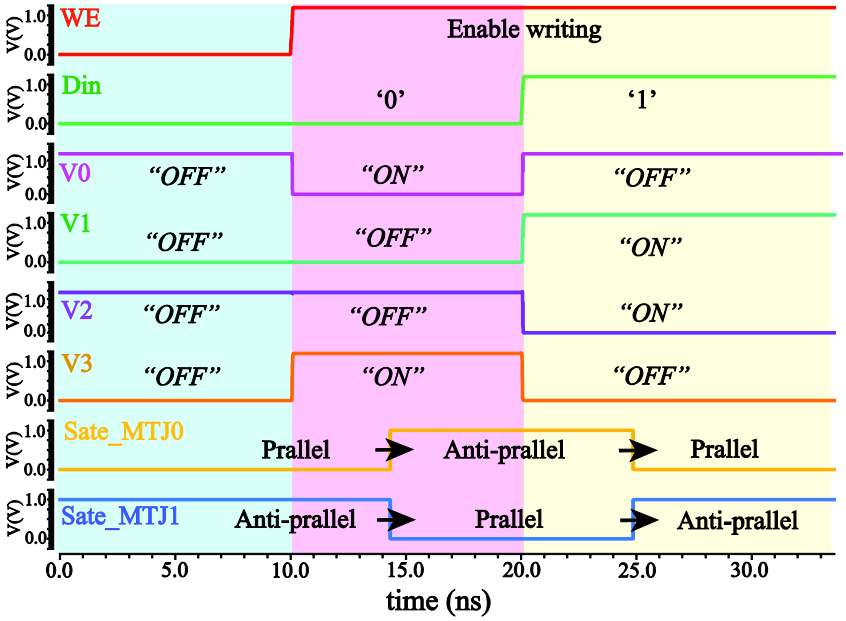


Figure 2.15 Simulation of the writing circuit. "ON" or "OFF" means that corresponding transistor is open or closed.

In the hybrid logic circuit, long writing delay of MTJ (when compared to the reading delay ~ ps) is an obstacle for achieving the requirement of high frequency operation. Reducing of the critical current is one solution to reduce the writing time. Another method is to increase the writing current. From the circuit-level, there are two ways to achieve higher writing current:

1) Increasing the supply voltage $Vdda$. According to Eq. 2.14, a higher supply voltage leads to larger writing current, and thus less time is needed to switch the state of MTJs. However, the power consumption is higher.

2) Increasing the transistor size. The resistance of NMOS and PMOS transistors in ON state is inversely proportional to the width (W) [8]. Therefore, we can both increase the writing current and reduce the writing delay at the expense of more area overhead.

$$I_{write} \approx \frac{V_{dda}}{R_p + R_n + R_{AP} + R_P} \qquad \text{Eq. 2.14}$$

where R_n and R_p are the resistances of NMOS and PMOS transistor in ON state.

2.2.3 Full hybrid MTJ/CMOS circuit

By combining the PCSA-based reading circuit and the writing circuit, hybrid MTJ/CMOS logic circuits can be designed. As shown in Figure 2.16, the 1-bit data stored in a pair of MTJs can be sensed by the PCSA circuit and written by the 4T or 6T writing circuit. In order to realize the writing operation without disturbing the outputs, two separating transistors N_3 and N_4 contribute to insulating the MTJ cells from the sensing part and thus preventing the writing current from passing through the sensing part during this phase.

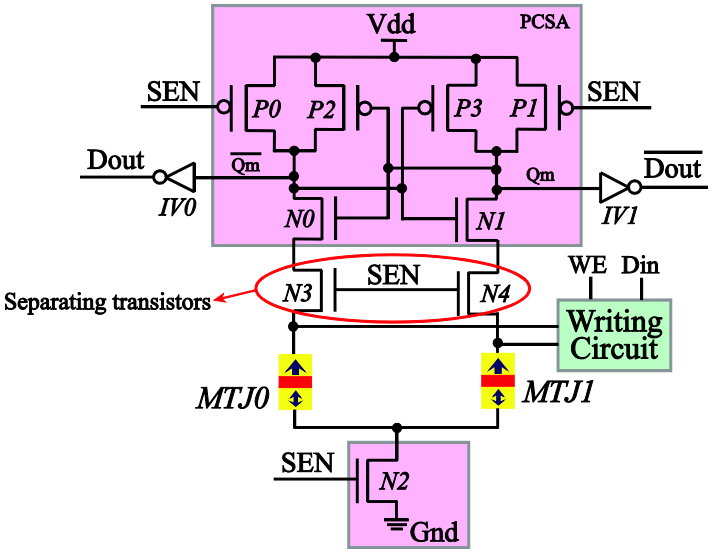


Figure 2.16 Full schematic of the reading/writing circuit

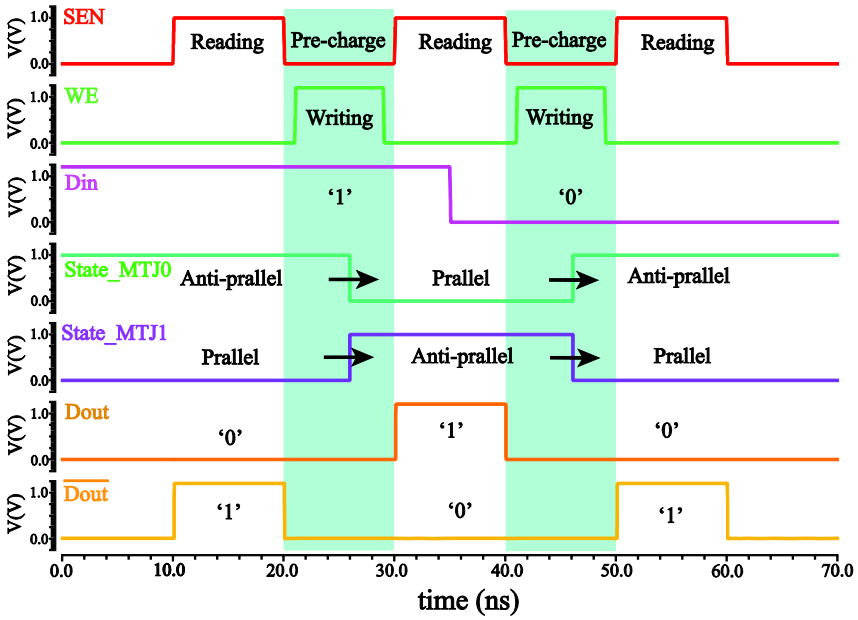


Figure 2.17 Simulation of the full reading/writing circuit

Figure 2.17 shows the full simulation of both sensing and switching operations. It should be noted that reading and writing share the same current path. Therefore, the switching operation should be performed when $SEN = '0'$, that is to say, during the pre-charge phase. In Figure 2.17, the state of $MTJ0$ is switched from AP to P and then back to AP, whereas that of $MTJ1$ is switched from P to AP and then back to P. $D_{out} = '0'$ is obtained during the first and third reading phases (or discharge phases) and $D_{out} = '1'$ during the second reading phase. The way to separate the current path for reading and writing operations by the three-terminal MTJ device will be introduced in Section 3.3.3.2.

2.3 Multi-context hybrid MTJ/CMOS circuit

The hybrid technology has shown its potential applications in both memory and logic implementations for energy saving because the unused blocks can be completely powered off without data lose. However, the aforementioned hybrid circuit with peripheral write/read circuit can only store and write 1-bit non-volatile data. It performs relative low density and the data is vulnerable. Multi-context (or multiple bit) hybrid logic architecture (see Figure 2.18), which has multiple non-volatile bits forming configuration plane for fast switching between contexts, has drawn much attention in logic designs [122], [123]. It provides further area-efficient property owing to the 3-D integration of multiple MTJs above the CMOS logic circuits. Moreover, the data security can be improved compared to the single-bit MTJ logic circuits. Data can be stored in two or more MTJs embedded in the same circuit. When an error occurs, the system can retrieve data from the nearby memory cell in which the same data was stored.

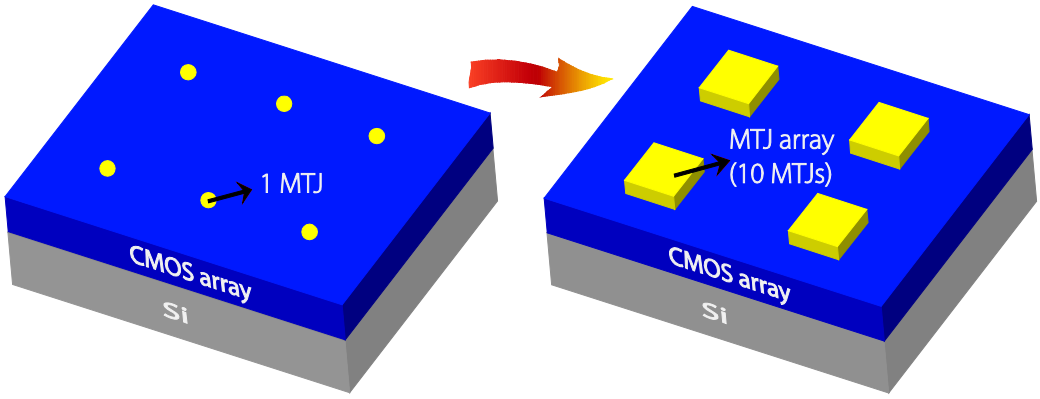


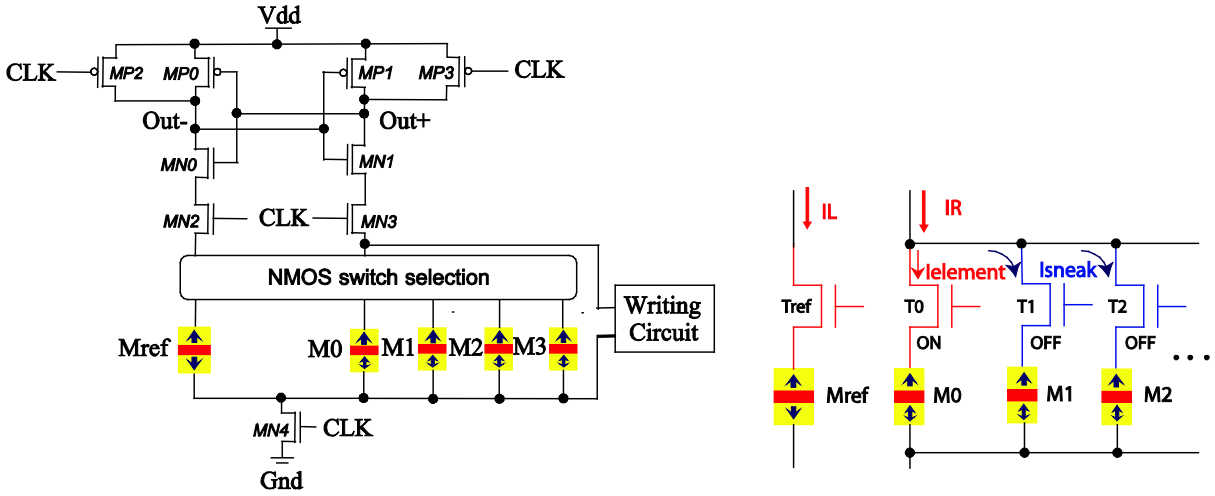
Figure 2.18 3-D structure of hybrid MTJ/CMOS integrating several memory cells (MTJs)

In the following paragraphs, multi-context hybrid MTJ/CMOS circuit will be proposed. Its advantages and disadvantages will be discussed, followed by structure-level and circuit-level optimizations.

2.3.1 Asymmetric structure based on pre-charge sense amplifier (asym-PCSA) and its reliability issues

Figure 2.19(a) shows the basic multi-context hybrid MTJ/CMOS structure integrating four contexts, where PCSA is used to evaluate the logic result. Different from the traditional way that stores 1-bit data in a couple MTJs with complementary states, this structure uses a reference MTJ (M_{ref}) to detect the non-volatile data stored in the storage MTJs (M_{0-3}). The

reference and storage MTJs keep the same round shape. Less power is consumed to switch only one MTJ during the writing operation instead of two. Switching between four contexts can be achieved by configuring a 2 to 4 (2-4) decoder, allowing only one NMOS selection transistor to be open while the other three selection transistors are kept closed. Therefore, one out of four MTJs (e.g, M_0) is selected. The resistance of the reference MTJ (R_{ref}) should be between R_P and R_{AP}. In our design, this is realized by using a reference MTJ whose diameter (i.e., 40 nm) is larger than that of the storage MTJs (i.e., 32 nm). This reference MTJ should always keep in anti-parallel configuration.



(a) (b)

Figure 2.19 (a) Schematic of multi-context hybrid MTJ/CMOS asymmetric structure based on PCSA (asym-PCSA) (b) Sneak paths problem in the asym-PCSA structure

This structure exhibits the performances of ultra-low power, area-efficiency and fast access speed. A low-power magnetic flip-flop (MFF) based on this structure was proposed in one of our publications [88]. The full schematic of this MFF is shown in Appendix A. However, the asym-PCSA structure faces several critical reliability issues:

a) Asymmetric sensing operation: In such structure, all the storage MTJs are placed on the same side while only one reference MTJ is placed on the other side. Therefore, several sub-branches are connected to each other as illustrated in Figure 2.19(b). During the reading operation, except for the current ($I_{element}$) flowing through the addressed MTJ sub-branch (e.g., M_0), some sneak currents (I_{sneak}) flowing through the closed sub-branches (e.g., M_{1-3}) are not negligible due to parasitic capacitances. The functionality of PCSA basically depends on the differential current of two paths. These

sneak currents might drastically affect the current difference between the two branches, leading to wrong evaluation of the logic result and significantly limit the sensing number of storage memory cells.

b) Highly scaled technology process: The increasing process variations in ultra-deep submicron technology (e.g., 28 nm) result in significant deviation of both MTJ and CMOS transistor parameters, leading to big offset of the sensing circuit [124].

These issues are difficult to overcome and they can completely perturb the sensing operation if there are not any mitigation solutions in ultra-deep submicron technology. For this reason, we will propose some optimization methods in the following sections.

2.3.2 Structure-level optimization

2.3.2.1 PCSA based symmetric structure (sym-PCSA)

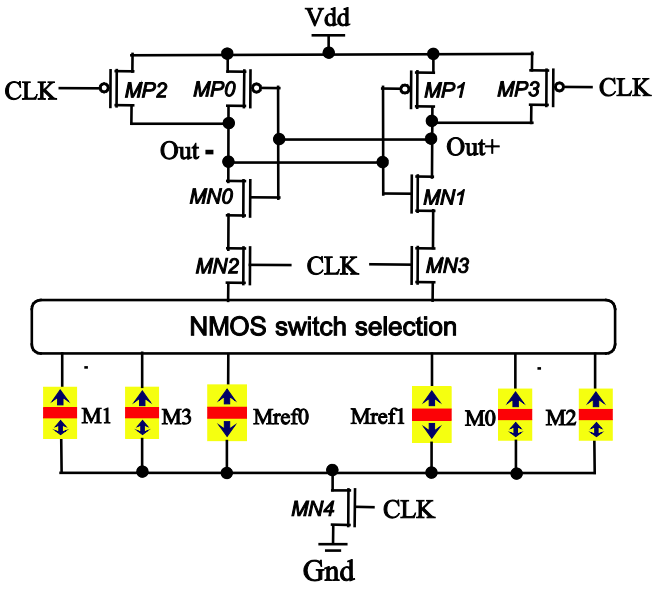


Figure 2.20 Schematic of multi-context hybrid MTJ/CMOS symmetric structure based on PCSA (sym-PCSA)

To overcome the asymmetric sensing problem and mitigate the influence of the aforementioned sneak currents, we propose a symmetric sensing structure. In such configuration, there are M storage MTJs (e.g., M=2 in Figure 2.20) and a reference MTJ on each side of the sense amplifier. It should be noted that M_{ref0} and M_{ref1} shares the same size and configuration. During the sensing operation, only a reference MTJ (e.g., M_{ref0}) on one side and a storage MTJ (e.g., M_0) on the opposite side are selected.

This design allows the disturbance of the total sneak currents from two branches to be drastically mitigated (see Eq. 2.15 and Eq. 2.16). Thanks to the balanced structure, the sneak currents of the closed sub-branches (e.g., M_{ref1} and M_{1-3}) on both sides are nearly the same. Therefore the sensing currents of the left branch I_L and the right branch I_R depend mainly on the currents passing through the selected paths (e.g., M_{ref0} and M_0).

$$I_L = I_{ref0} + \sum_{i=0}^{N-1} I_{sneak\_L\_i} \qquad \text{Eq. 2.15}$$

$$I_R = I_0 + \sum_{i=0}^{N-1} I_{sneak\_R\_i} \qquad \text{Eq. 2.16}$$

where N is the number of the closed sub-branches on one side of the structure (e.g., $N = 2$ in Figure 2.20), I_L and I_R are the currents passing through the two branches of the structure. I_{ref0} is the current passing through M_{ref0} and I_0 is the current passing through M_0. $I_{sneak\_L\_i}$ and $I_{sneak\_R\_i}$ are the sneak currents passing through the closed sub-branches on both sides.

This symmetric structure may improve significantly the scalability of the hybrid architecture. However, it has no impact on reliability, which is mainly dominated by the sensing circuit (i.e., PCSA) in the ultra-deep submicron technology. Therefore we will further propose a higher reliability sensing circuit based on the symmetric structure.

2.3.2.2 Symmetric structure based on separate pre-charge sense amplifier (sym-SPCSA)

To overcome the scaled technology process issues, double-tail sensing amplifiers are proposed [125], e.g., separated pre-charge sense amplifier (SPCSA). They achieve indeed better reliability performance. However, few solutions have been designed particularly for ultra-deep submicron hybrid MTJ/CMOS logic circuits. In our work, SPCSA is first used for reliable reading of non-volatile data stored in MTJs [126]. In this sub-section, a new multi-context sym-SPCSA structure, which combines the symmetric structure and SPCSA, is proposed (see Figure 2.21). The main difference between the PCSA and SPCSA is that SPCSA with two discharge tails allows separating the discharge phase from the evaluation phase. In addition, with two inverters (IV_0 and IV_1) inserted between the discharge part and the evaluation part, the small current difference (due to limited TMR ratio) in the discharge phase is amplified

before entering the evaluation phase. Thus, the sensing margin (denoted as the voltage different or the current difference which will be described in the following paragraphs) is greatly increased to tolerate the process variations.

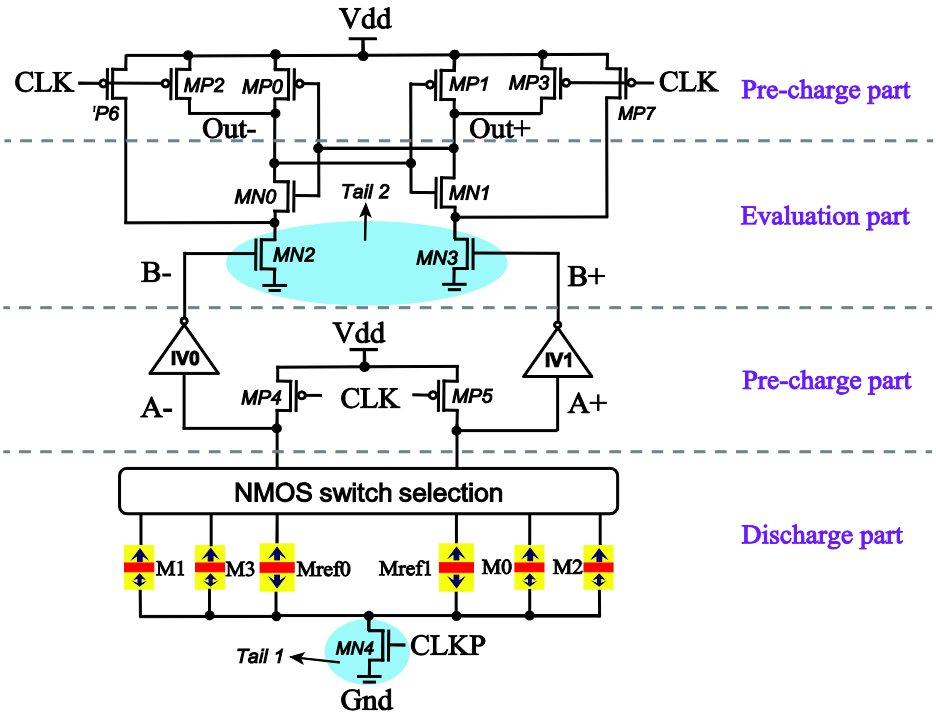


Figure 2.21 Schematic of multi-context hybrid MTJ/CMOS symmetric structure based on separated pre-charge sense amplifier (sym-SPCSA), which has three parts: pre-charge part, evaluation part and discharge part.

Three phases (pre-charge phase, discharge phase and evaluation phases) for one sensing operation of the sym-SPCSA structure can be described as follows:

- During the pre-charge phase ($CLK = CLKP = '0'$), MN_4 are closed while MP_{2-5} are turned ON and they charge both nodes A ($A+$ and $A-$) and Out ($Out+$ or $Out-$) to V_{dd}. Nodes B ($B+$ and $B-$) are then discharged to the ground through inverters IV_0 and IV_1. Transistors MN_{2-3} are then turned OFF. Therefore, there is no current passing from V_{dd} to the ground in both discharge and evaluation parts.
- During the discharge phase ($CLK = CLKP = '1'$), both $A+$ and $A-$ nodes begin to discharge but with a different time rate. That is because the sensing current is inversely proportion to the resistance of MTJ and the addressed MTJs on two sides have different resistance values. As a result, a differential voltage (ΔA) between $A+$ and $A-$ is created, which generates, after the propagation delay of the inverters, a differential voltage at the B nodes (ΔB). This leads to a different turn-on time for

the transistors MN_2 and MN_3.

- During the evaluation phase, MN_2 and MN_3 keep ON, enabling the output nodes to discharge. Once one of the output nodes ($Out+$ or $Out-$) reaches the threshold voltage of the back-to-back cross-coupled inverter (MN_0/MP_0 or MN_1/MP_1), the other output ($Out-$ or $Out+$) will be pulled up to V_{dd} (logic "1"), and this specific output continues to discharge to the ground (logic '0'). In this way, output stage generates the small voltage difference ΔB into digital signals.

It should be noted that the separating transistors, which are necessary in the PCSA based structures, are not required in the SPCSA based structure. Once the discharge is finished, $CLKP$ can be reset to logic '0' to close MN_4. As MP_{4-5} are also closed ($CLK = '1'$), the MTJ sub-branches can be completely separated from the sense amplifier. Therefore, writing and reading can take place in the same phase with a delay controlled by CLK and $CLKP$.

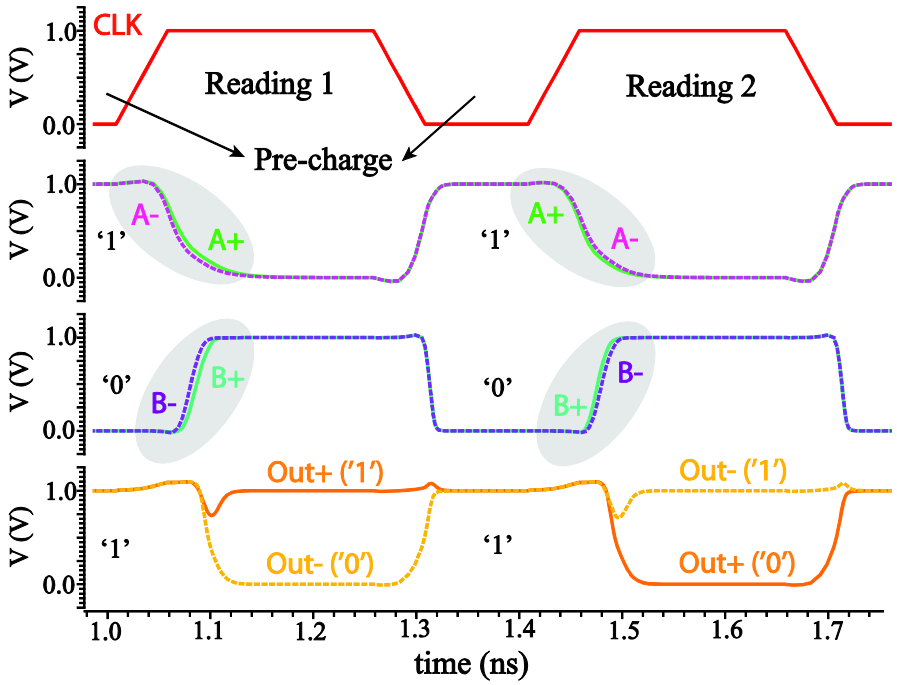


Figure 2.22 Signal behavior of the multi-context sym-SPCSA circuit

The simulated waves, where the voltage differences ΔA and ΔB are clearly illustrated, are shown in Figure 2.22. It confirms the two reading operations of data stored in M_0 (in antiparallel state) and M_2 (in parallel state). 1) During the pre-charge phases, nodes A and Out are pre-charged to 1 V while nodes B are pulled down to 0 V. 2) During the first reading phase (Reading 1), $A-$ discharges faster than $A+$, and thus $B-$ reaches the

threshold of NMOS transistor faster than $B+$. As a result, logic '1' and '0' can be read at the output $Out+$ and $Out-$. 3) During the second reading phase (Reading 2), $B+$ reaches the threshold of NMOS transistor faster since $A+$ discharges in a faster speed. Finally, $Out+=\text{'0'}$ and $Out-=\text{'1'}$ are obtained.

2.3.2.3 Comparative discussion

Transient and MC statistical analyses are performed by using the MTJ model and a STMicroelectronics 28 nm bulk CMOS design kit [127] to exhibit their functionalities and effectiveness. Some major performances (e.g., delay, energy, size, scalability and reliability) of the three structures are obtained and concluded in Table 2.3.

All three structures can operate at a high frequency as they maintain a propagation delay lower than 200 ps, thanks to the fast dynamic sensing approach. Besides, they perform low sensing power, which reaches nearly negligible level (~ fJ). The asymmetric structure (asym-PCSA) exhibits poor scalability and at most five MTJs can be integrated, while the symmetric structures (sym-PCSA and sym-SPCSA) show good prospect in embedding a large number of MTJs, e.g., 32 MTJs. The sym-SPCSA structure shows almost half less error rate and 14.2% smaller sensing time compared to the asym-PCSA structure, with all transistors kept in the minimum size. Thus it exhibits the best reliability and sensing speed. However, its reading energy is nearly four times larger than the PCSA-based structures (asym-PCSA and sym-PCSA) due to its two current paths.

Table 2.3 Comparison of three multi-context hybrid MTJ/CMOS structures

Performances		**asym-PCSA**	**sym-PCSA**	**sym-SPCSA**
Delay time (ps/bit)		160	162.7	139.6
Energy (fJ/bit)		1.21	1.24	5.32
Size		14T	15T	23T
MTJ number limitation		< 6 MTJs	> 30 MTJs	> 30 MTJs
Bit error rate (BER)	MTJ_AP	30.4%	29.8%	15%
	MTJ_P	32.2%	34.6%	19.5%
	Average BER	31.3%	32.2%	17.25%

More specifically, we focus mainly on the reliability performance. The sensing BER values in Figure 2.23 are the average values for detecting MTJ in two states. It is a crucial parameter to

evaluate the robustness of a hybrid logic circuit. 1000 runs of MC simulation by considering the variations of CMOS transistors and MTJ model have been performed to obtain the curves, illustrating that sym-PCSA has similar performance compared to asym-PCSA, except that it liberates the limitation on the memory cells (< 6 MTJs). As can be seen, sym-SPCSA structure exhibits the best error rate when maintaining the same area overhead compared to the other two structures. In practical applications, we can choose from the three structures based on the application-oriented requirements to obtain the best performance trade-off including area, power, latency and reliability.

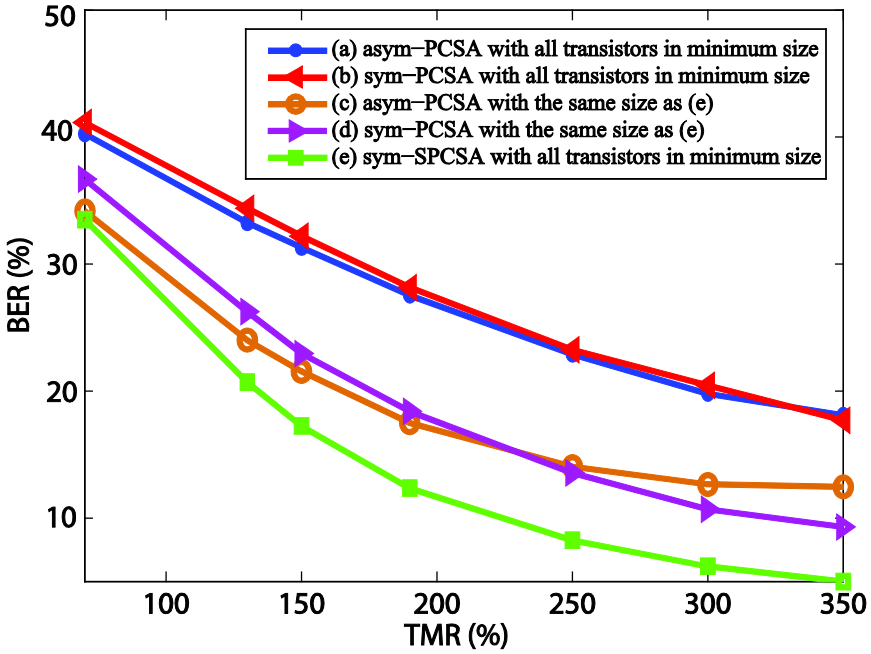


Figure 2.23 Sensing error rate reduces rapidly with the increase of TMR value

2.3.3 Circuit-level optimization

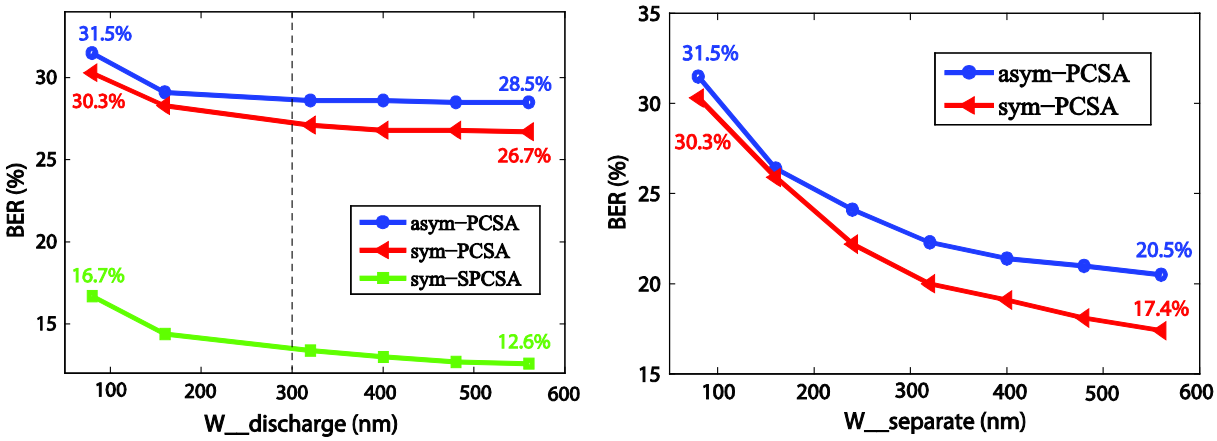
In the previous sub-sections, we have proposed two structures (i.e., sym-PCSA and sym-SPCSA) to improve the sensing scalability and reliablity. We propose several optimization methods to further improve the reliability of the multi-context hybrid MTJ/CMOS circuits including CMOS transistor sizing, dynamic reference MTJ selection and multi-Vt strategy.

2.3.3.1 CMOS transistor sizing

The CMOS transistors in the hybrid MTJ/CMOS logic structure, such as the discharge transistor and the separating transistors, play different roles and induce different reliability issues. In this part, we give some basic ideas of reliability improvement of the three structures

by varying the size of different transistors.

The resistance of NMOS transistor in the open state (or ON state) is inversely proportional to its width (W). Larger W, therefore, leads to lower resistance, providing larger sensing currents and sensing margin to overcome the offset or mismatch (caused by process variations). However, the resistance of the discharge transistor (MN_4 in Figure 2.19(a), Figure 2.20 and Figure 2.21) becomes too small to affect the sensing currents if W exceeds a certain value (~300 nm). Hence, BER becomes less sensitive to the size of the discharge transistor. This is confirmed by the simulation results shown in Figure 2.24(a). Figure 2.24(b) represents the BER with respect to the width of separate transistors (MN_2 and MN_3 in Figure 2.19(a) and Figure 2.20). It can be concluded that larger separating transistor size is expected to obtain less sensing errors for the PCSA based symmetric and asymmetric structures. It has been mentioned that the separate transistors are not necessary for the SPCSA structure, and therefore the curve of BER with respect to the separate transistors are neglected.



(a) (b)

Figure 2.24 Sensing bit error rate (BER) with respect to (a) the discharge transistor (b) the separating transistors

2.3.3.2 Dynamic reference MTJ selection

In conventional circuits, the reference cell is formed by putting in parallel two serially connected MTJs, as shown in Figure 2.25. This reference cell suffers loss of reliability as the four-MTJ structure has much more variations. Besides, it can be seen in Eq. 2.3 that the TMR ratio is variable and it decreases as the reading bias voltage V_{bias} increases. The four-MTJ reference cell has larger range of TMR ratio variation because two MTJs with opposite configurations are connected in series.

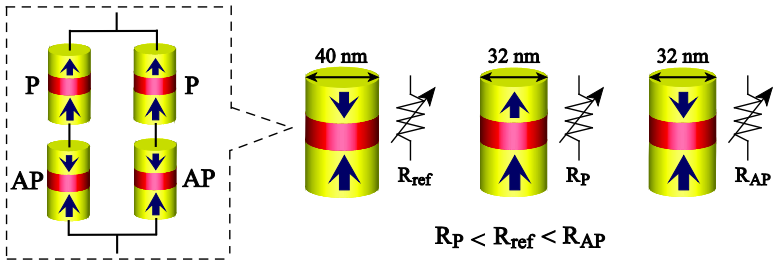


Figure 2.25 Resistance of the reference resistance corresponding to the intermediate resistance R_{ref}, the parallel low resistance R_P and the anti-parallel high resistance R_{AP}

In our proposed hybrid MTJ/CMOS design, instead, only one MTJ kept in anti-parallel state (i.e., M_{ref}) acts as the reference element. Based on equations Eq. 2.2 and Eq. 2.4, the resistance of M_{ref} is lower than the storage MTJ when M_{ref} has a larger diameter. This configuration allows keeping the same structure on both sides of the circuit, providing symmetrical sensing paths of both branches. In order to obtain the best sensing margin between I_L and I_R, the resistance value of M_{ref} should be equal to $R_{ref} = (R_{AP} + R_P)/2$. For example, R_P and R_{AP} of the storage MTJs with the diameter of 32 nm are 6.43 kΩ and 16.09 kΩ, respectively. Then, the size of M_{ref} should be set to 38 nm or 39 nm to obtain a resistance value approximate to 11.26 kΩ according to the simulations. This can be achieved by varying the surface of M_{ref}.

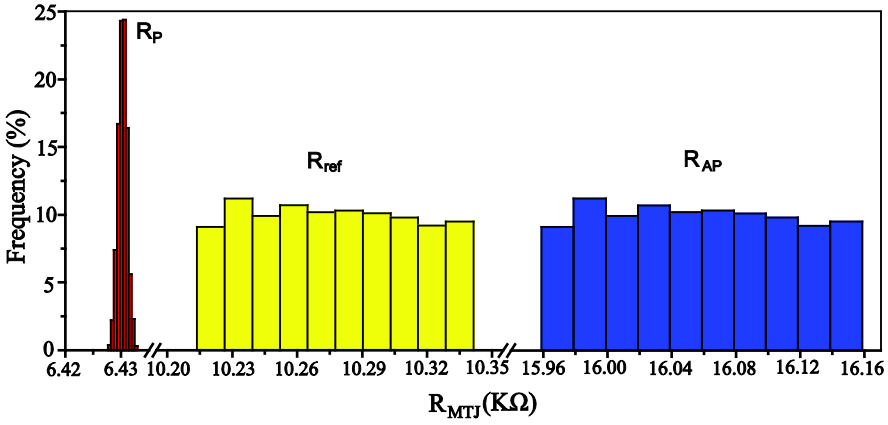


Figure 2.26 MTJ resistance (R_{MTJ}) distribution obtained from the Monte-Carlo simulation (1000 runs). R_P, R_{AP} and R_{ref} represent the resistances of storage MTJ in parallel state and anti-parallel state and reference MTJ, respectively.

However, the process variations make the resistances R_P, R_{AP} and R_{ref} deviate from the designed values. Figure 2.26 shows an example of MC simulation histogram for both high resistance and low resistance (R_{AP} and R_P) of the storage MTJ, and the resistance of the

reference MTJ (R_{ref}). The same bias voltage is applied to three types of MTJ, i.e., storage MTJ in parallel and anti-parallel state with a diameter of 32 nm, reference MTJ in anti-parallel state with a diameter of 40 nm. It can be seen that R_{AP} has a much larger distribution than R_P. Therefore the value of R_{ref} should be smaller than the designed value (that keeps R_{ref} in the exact middle of R_P and R_{AP}) to keep a BER balance to read MTJ in two states.

Table 2.4 shows that with the increasing size of M_{ref}, the errors for reading the MTJ in anti-parallel state slightly decreases while that for reading the MTJ in parallel state increases. This is because the resistance of the reference MTJ decreases when its diameter increases, creating a larger resistance difference between R_{ref} and R_{AP} ($\Delta R_1 = R_{AP} - R_{ref}$) and a smaller difference between R_{ref} and R_P ($\Delta R_2 = R_{ref} - R_P$). Thus a proper selection of the reference MTJ size is important to reduce the average BER, while keeping a similar BER for sensing MTJ in two states (parallel and anti-parallel). Table 2.5 presents the best M_{ref} choices of three structures for reliability enhancement.

Table 2.4 Simulations of three structures by varying the size of M_{ref}

M_{ref} size / nm	**R_{ref} / kΩ**	**Bit error rate (BER)**					
		asym-PCSA		**sym-PCSA**		**sym-SPCSA**	
		R_{AP}	R_P	R_{AP}	R_P	R_{AP}	R_P
35	13.45	44.6%	21.5%	41.1%	23.6%	31.8%	7.3%
36	12.71	41.4%	23.7%	38.6%	26.1%	28.1%	8.4%
37	12.03	38.1%	25.6%	36.3%	28%	24.3%	11.1%
38	11.41	35.6%	27.9%	*33.4%*	*29.8%*	20.4%	13.8%
39	10.83	*32.3%*	*30.5%*	*31.7%*	*33.2%*	*18.1%*	*16.7%*
40	10.3	*30.4%*	*32.2%*	29.8%	34.6%	*15%*	*19.5%*
41	9.8	28.7%	33.7%	28%	36.7%	23%	23%

Table 2.5 Best M_{ref} size of three structures

Structure	**Diameter of M_{ref}**	**Average BER**
asym-PCSA	40 nm	31.3%
sym-PCSA	38 nm	31.6%
sym-SPCSA	40 nm	17.25%

2.3.3.3 Multi-Vt design strategy

The threshold voltage (Vt) is one of the main CMOS parameters that affect device performances. As mentioned in [128], a single Vt design can no longer meet application goals in most 28 nm production SOC designs due to significant variations. Consequently, it becomes necessary to implement multi-Vt strategy in MTJ based logic circuit design. Two or three levels of Vt is a good choice for multi-Vt optimization. Implementing more than three levels Vt cells often introduces more challenges in variation control cross all signoff corners. This part of work uses the multi-Vt strategy, which assembles two types of CMOS transistors in the same structure, for the purpose of reliability optimization.

We have done a full study of the transistor combinations of two levels Vt cells, called “rvt” and “lvt”, of the three structures. Here, “rvt” represents regular transistor and “lvt” denotes low threshold voltage transistor, respectively. In order to obtain the best reliability of the three proposed structures, we look for the best multi-Vt combinations. For the asym-PCSA structure, higher reliability can be achieved when the NMOS transistors MN_{0-4} and PMOS transistors MP_{0-3} are of “rvt” type while the others are of “lvt” type. For the sym-PCSA structure, only the NMOS transistors MN_{0-3} and PMOS transistors MP_{0-3} are expected to use the type “rvt”. Moreover, all the transistors should use the low Vt type in order to perform good reliability for the sym-SPCSA circuit.

2.3.3.4 Combination of the three reliability optimization methods

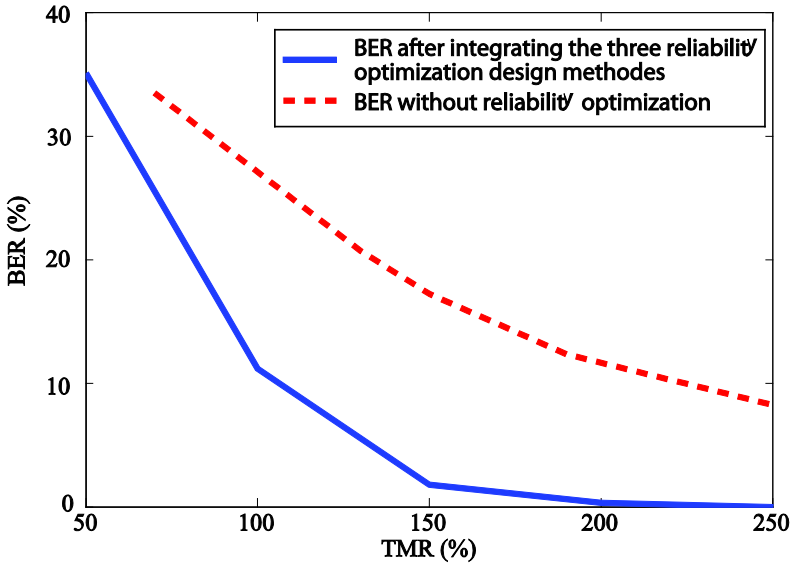


Figure 2.27 BER of the sym-SPCSA structure versus TMR ratio

More simulations have been done to realize further reliability enhancement by incorporating all the aforementioned methods in the multi-context circuits. The PCSA based structures are able to realize nearly zero sensing error with a TMR ratio of 200% by quadrupling the circuit area, while the sym-SPCSA structure performs area-efficiency as it only needs to double the area to achieve the "nearly zero sensing error" requirement for logic application. The optimized results of the sym-SPCSA structure are shown in Figure 2.27 with blue solid curve. The red dotted line presents the BER results without any optimization.

Other methods can be integrated to optimize the reliability performance of multi-context hybrid circuit. For example, we can use a couple of MTJs that are in complementary states (e.g., M_0 and M_1 in Figure 2.28) to store one bit. The sensing margin is maximized. However, the density is decreased since the number of MTJs is nearly doubled.

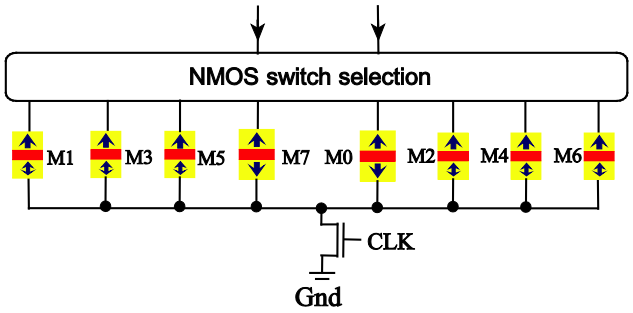


Figure 2.28 Schematic of the non-volatile storage part for reliable multi-context hybrid MTJ/CMOS circuit. Two MTJs in opposite states store 1-bit data.

Magnetic random access memory (MRAM) is one of the most important applications of MTJ devices. We implement the design of a novel MRAM, where the storage data is locally sensed by using the multi-context hybrid MTJ/CMOS circuit discussed in this section.

2.4 Design of 1KB magnetic random access memory using spin transfer torque switching mechanism (STT-MRAM)

DIPMEM project is leaded by CEA-LETI, and other partners involved in this project are SPINTEC, IEF, IM2NP, LIRMM and CMP. It aims at demonstrating the advantages of emerging non-volatile memories integrated in logic blocks in processor in terms of ultra-low power consumption, high reliability and data security. The ambition of DIPMEM project is to realize a demonstrator of an embedded processor with two resistive memory technologies, STT-MRAM and ReRAM. As part of this project, a 1 kilobyte magnetic random access memory using spin transfer torque switching mechanism (1KB STT-MRAM) for embedded processor is designed and validated by using the PMA STT-MTJ model, STMicroelectronics 28nm FDSOI CMOS design kit and STT-MTJ back-end process brought by SPINTEC.

2.4.1 MRAM architecture

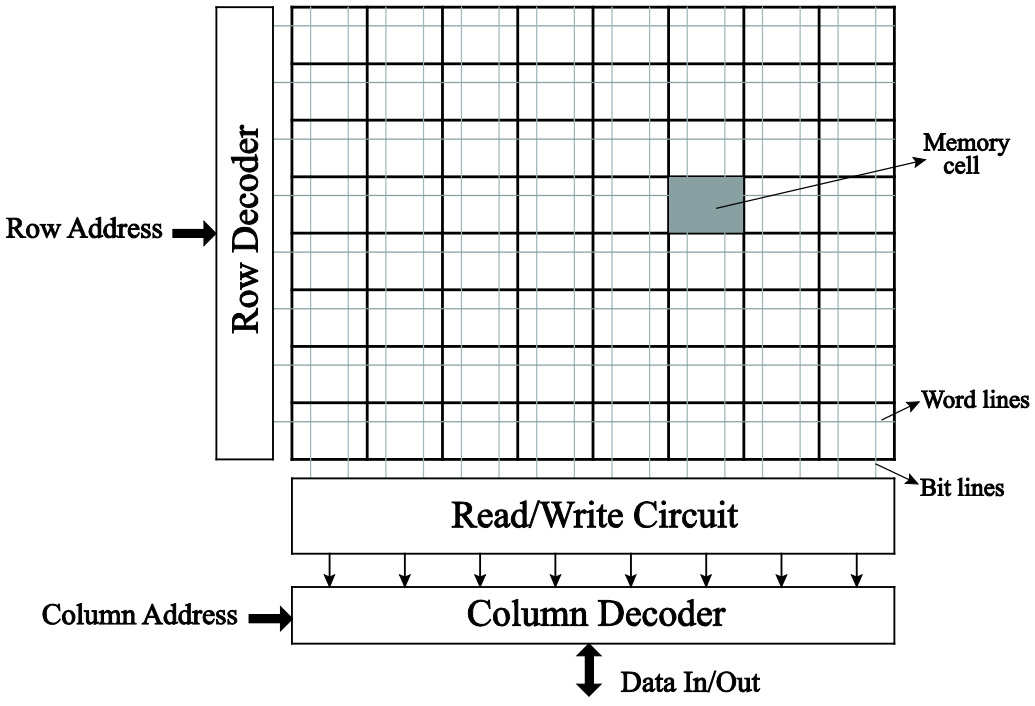


Figure 2.29 Memory array architecture

As illustrated in Figure 2.29, the general MRAM architecture is composed of a memory array and peripheral circuits [129], [130]. The peripheral circuits include sense amplifiers for data reading, write drivers for data programing and row/column decoders for word/bit selection. There are two main issues caused by this architecture: 1) there is a large number of logic gates in the complicated row/column decoder in order to address a specific cell in a large memory.

This increases not only the area overhead but also the addressing delay; 2) for a mass memory (> 1 Kb), long word/bit lines increase the propagation delay as well as read/write power consumption due to increasing line capacitance.

In order to deal with the aforementioned issues and realize high speed embedded memory, we propose a novel MRAM architecture shown in Figure 2.30. The 1KB memory array is divided into 64 subarrays, and each subarray (16B array) includes four lines and 32 columns (4-word-length-32-bit-width). The length of local lines is greatly reduced. The "Predecoder" block, whose inputs are 8-bit row address $Addr[0:7]$, generates signals $SA[0:3]$, $SB[0:3]$, $SC[0:3]$ and $SD[0:3]$. The "BL_select" block is activated only when $BE='1'$, selecting one byte to be read or written.

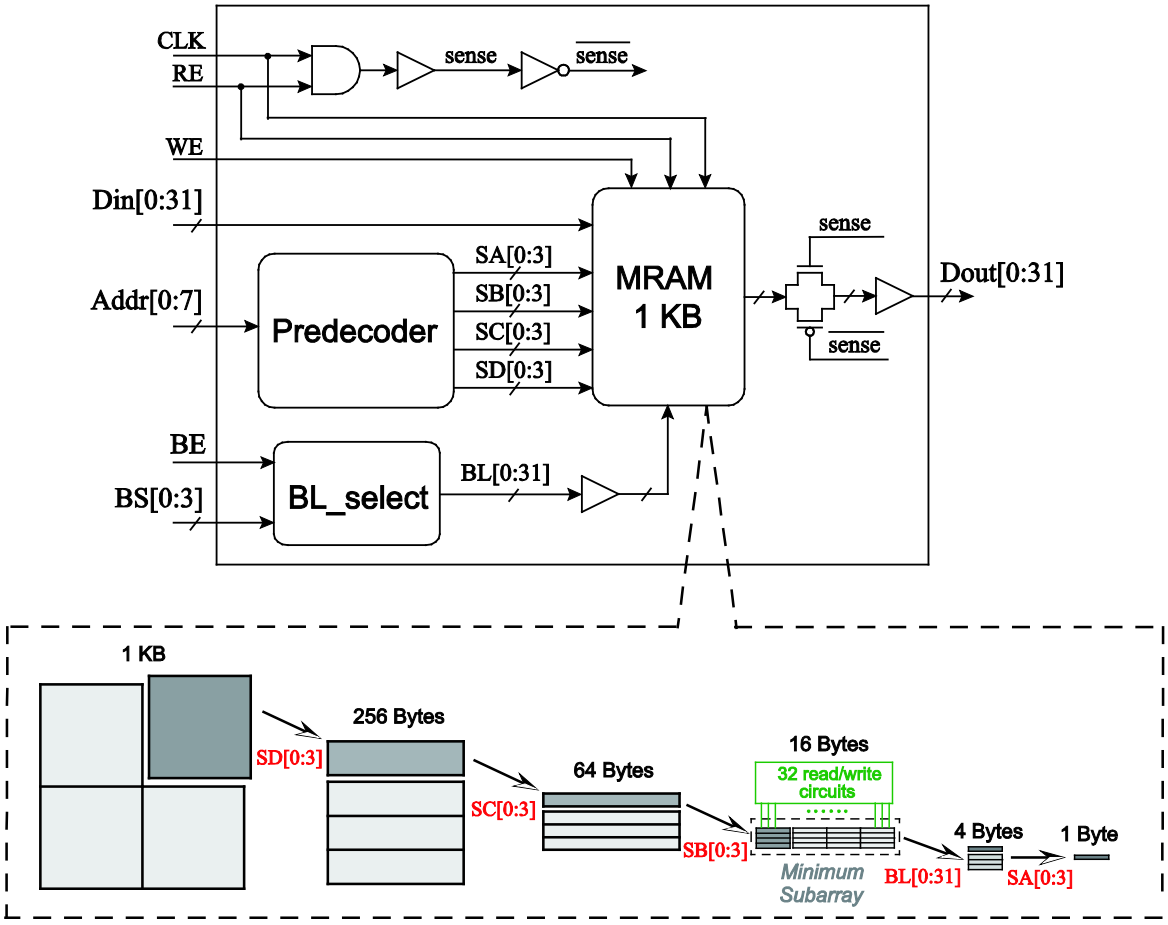


Figure 2.30 Structure of the proposed 1kB MRAM

During the read/write operation, only one 16B array is activated by intermediate signals $SD[0:3]$, $SC[0:3]$ and $SB[0:3]$. For instance, the first subarray is selected with $SD[0]='1'$, $SC[0]='1'$ and $SB[0]='1'$. The sense amplifier and the write drivers are no longer shared by all the storage cells in the same column, but by 4-bit memory cells in the same column of this

subarray. In other words, there are 32 read/write circuits in each 16B array. Signals $SA[0:3]$ are therefore used for selecting one out of four lines in the 16B array (for example, the first line is chosen when $SA[0]='1'$). Only 3-bit column address $BS[0:3]$ is necessary in this architecture, selecting one word to be read or programmed. In our design, the minimum red/write unit is one byte (or 8 bits). This number can be changed by designers (e.g., 16 bits or 32 bits) according to different applications, which will de detailed in the following subsection. One transmission gate and one CMOS buffer constitute the output driver.

There are two types of signals, i.e., control signals and data signals. Read/write control signals RE and WE specify the memory operations. Read operation takes place when $RE='1'$ and $CLK='1'$, that is, the intermediate signal $sense$ is '1'. Write operation (or programing operation) is performed when $WE='1'$. Other signals are listed in Table 2.6.

Table 2.6 List of control signals and data signals

Name	Description	IN/OUT
CLK	Synchronous clock signal	IN
RE	Enable reading	IN
WE	Enable writing	IN
BE	Enable selecting a word for reading/writing	IN
Addr<2:9>	Row address	IN
BS<0:3>	Column address	IN
Din<0:31>	Input data to be stored	IN
Dout<0:31>	Output read data	OUT

In the following, we will explain all the blocks in details: their functionality, circuit-level design and layout.

2.4.2 Memory blocks design

2.4.2.1 Memory unit

As mentioned before, 4-bit storage cells are embedded in the same local sense amplifier and share the same writing circuit. Figure 2.31 shows the transistor level schematic, which is designed based on the multi-context hybrid MTJ/CMOS circuit. It is composed of four parts:

- ***4-bit storage part***: each 2T/2MTJ bit cell includes a pair of MTJs (e.g., M_0 and M_1

in complementary states) to store 1-bit non-volatile data and two transistors (e.g., P_4 and P_8) to conduct/block the sub-branches. The control signals $WL[0:3]$ and the control mechanism will be described in Section 2.4.2.2.

- ***Read circuit***: PCSA is used to detect magnetization configuration of MTJs. One more discharge transistor N_3 is added to provide two-tail writing, that is, there are two writing currents passing through the MTJs places on two sides.

- ***Write circuit***: two 4T write circuits are used for generating different writing currents I_{write_L} and I_{write_R} to switch the magnetization of MTJs. WE and $BL[i]$ are active signals and $Din[i]$ controls the switching direction (P→AP or AP→P), where i (0→31) is the bit number.

- ***Output circuit***: The sensed output on node Qm is transmitted to D_{out} through an output circuit. $D_{out}[i]$ equals to Qm only when $BL[i]='0'$. Otherwise, the output is with high impedance.

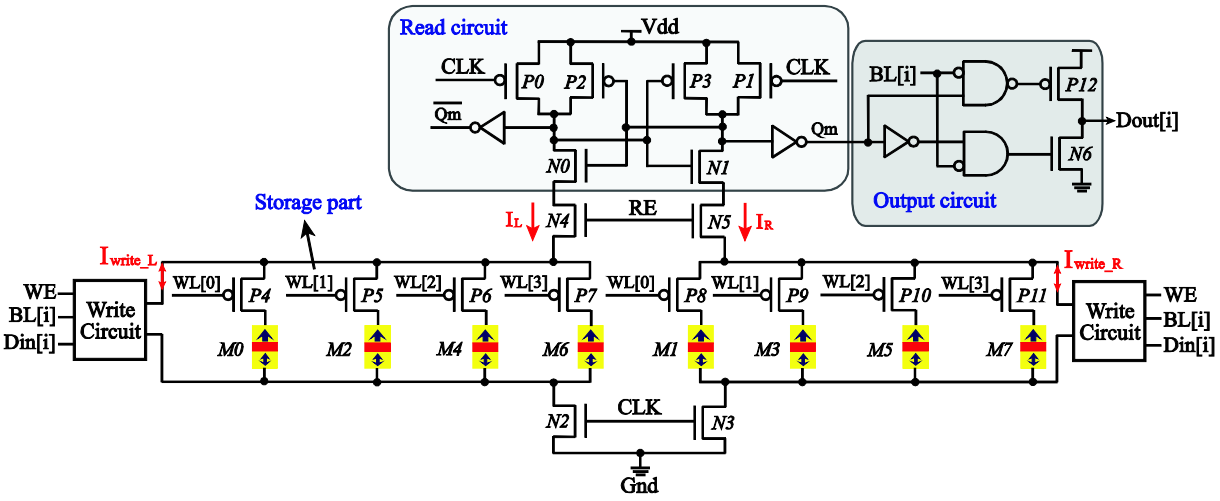


Figure 2.31 Schematic of the 1kB MRAM memory unit

MTJs can be fabricated in the back-end of the CMOS process from the metal level 6 (M6), last metal level of the STMiceoelectronics 28 nm FDSOI technology (see Figure 2.32(a)). The advantage of this fabrication process is that MTJs do not take much area. However, the contact necessary to connect MTJs with CMOS transistors is large due to the fabrication characteristics available at the laboratory SPINTEC. As shown in Figure 2.32(b), the connection layers “LIG_INF” and “LIG_SUP” are connected onto the bottom and top of MTJ. “VIA1_MAG” connect the M6 and “LIG_INF”, and “VIA2_MAG” connect “LIG_INF” and

"LIG_SUP".

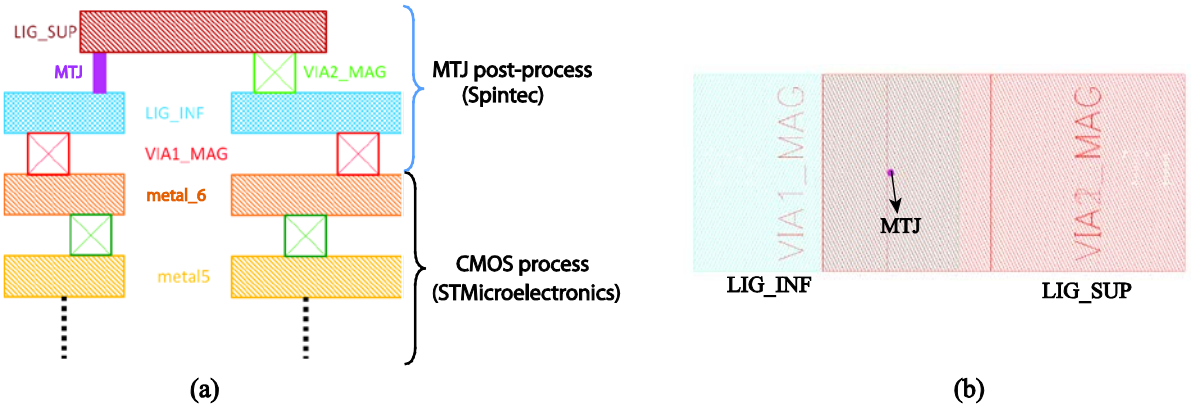


Figure 2.32 (a) Hybrid MTJ/CMOS process. MTJ is integrated above CMOS circuit from metal level 6 (M6) (b) Layout of the MTJ including MTJ nano-pillar, lower connection layer (LIG_INF) and upper connection layer (LIG_SUP)

Figure 2.33 is the full layout of this memory unit. Even though the CMOS circuit occupies small area, the layout area (68.755 μm^2×13.604 μm^2) is large, owing to large contacts "LIG_INF" and "LIG_SUP". Two verification tools, Design Rules Checking (DRC) and Layout Versus Schematic (LVS), are used for the layout of each block and then the full 1KB MRAM. DRC helps designers to verify whether the layout satisfies the design rules of CMOS process as well as the MTJ back-end process, for instance, the minimum space between two metal layers. LVS allows designers to confirm whether the designed layout corresponds to the original circuit schematic.

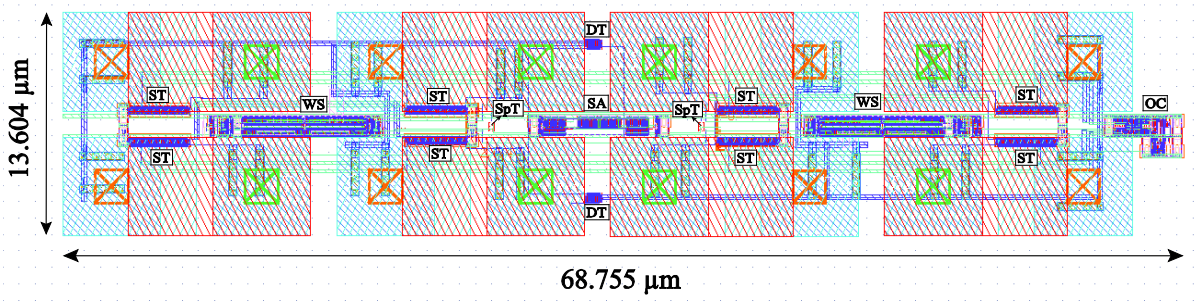


Figure 2.33 Layout of the memory unit. It has an area of 68.755 μm^2×13.604 μm^2. ST, DT, SpT represent the selection transistors P_4-P_{11}, discharge transistors N_2-N_3, separating transistors N_4-N_5, respectively. WS, SA, OC represent the write circuits, sense amplifier and output circuit.

The following structures are implemented in regular CMOS technology.

2.4.2.2 Local decoder

The 16B page is composed of 32 memory units and a decoder. The local decoder allows generating word line selection signals. With SB, SC, SD are equal to '1', the 3-input AND logic gate allows selecting one specific 16B array. Only one of the selection signals $WL[0:3]$ will be '0' depending on the configuration of $SA[0:3]$. For example, $WL[0]='0'$ and $WL[1:3]="111"$ when $SA[0]='1'$ and $SA[1:3]="000"$. The logic circuit includes four NAND logic gates and four buffers. The schematic of the local decoder and the layout are shown in Figure 2.34 and Figure 2.35, respectively.

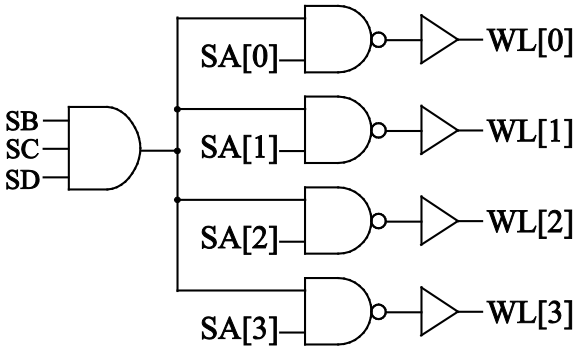


Figure 2.34 Schematic of the local decoder circuit

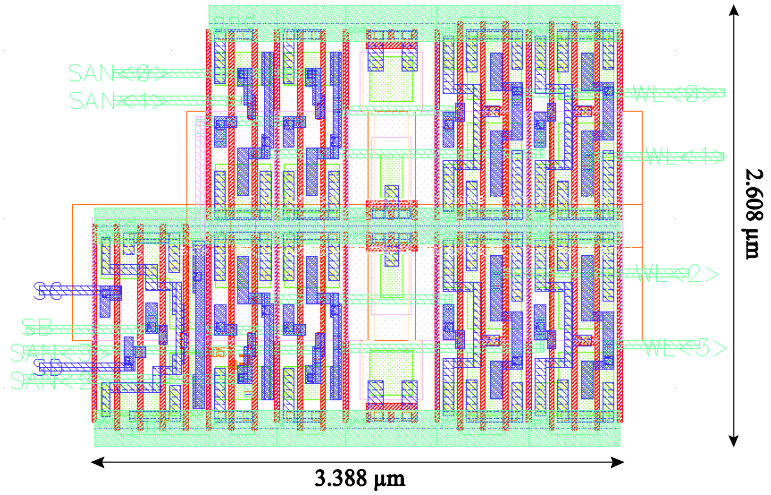


Figure 2.35 Layout of the local decoder and its area is 3.388 µm^2×2.608 µm^2.

2.4.2.3 Pre-decoder block

The "Pre-decoder" block decodes 8 inputs $Addr[0:7]$ into 16 address buses by using four 2-4 CMOS based decoder (see Figure 2.36 and layout in Figure 2.37). As mentioned above, a 16B array is activated by 12 address buses $SD[0:3]$, $SC[0:3]$ and $SB[0:3]$. Then the output signals $SA[0:3]$ of the "Pre-decoder" block select one word line to be read or written.

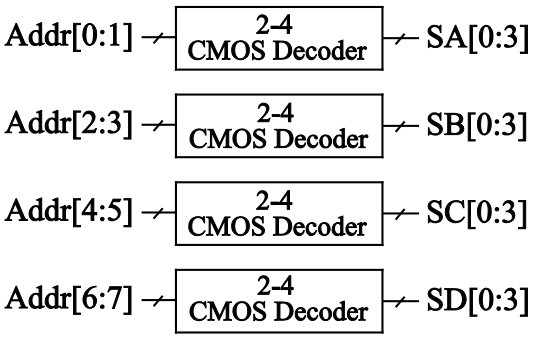


Figure 2.36 Schematic of the 8-16 pre-decoder circuit

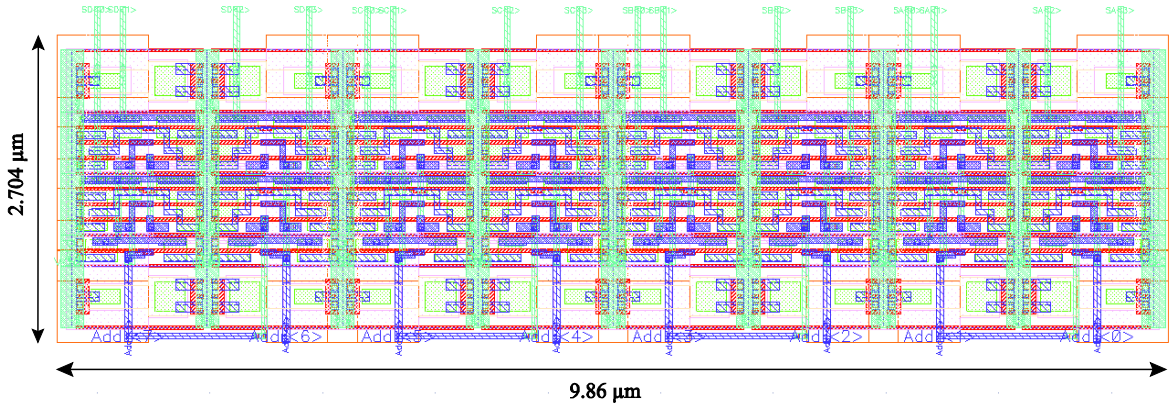


Figure 2.37 Layout of the 8-16 pre-decoder and its area is 9.86 $\mu m^2 \times 2.704\ \mu m^2$.

2.4.2.4 Byte selection block

Each word line is composed of 32 bits and can be divided in to four bytes by the byte selection signals $BL[0:31]$. It should be noted that $BL[i]='0'$ enables the output circuit and the write circuit of the i^{th} bit in a line (see Figure 2.31). The byte selection block (BL_select) and its layout are shown in Figure 2.38 and Figure 2.39.

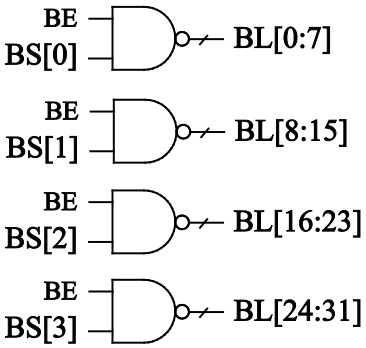


Figure 2.38 Schematic of the bit line selection block (BL_select)

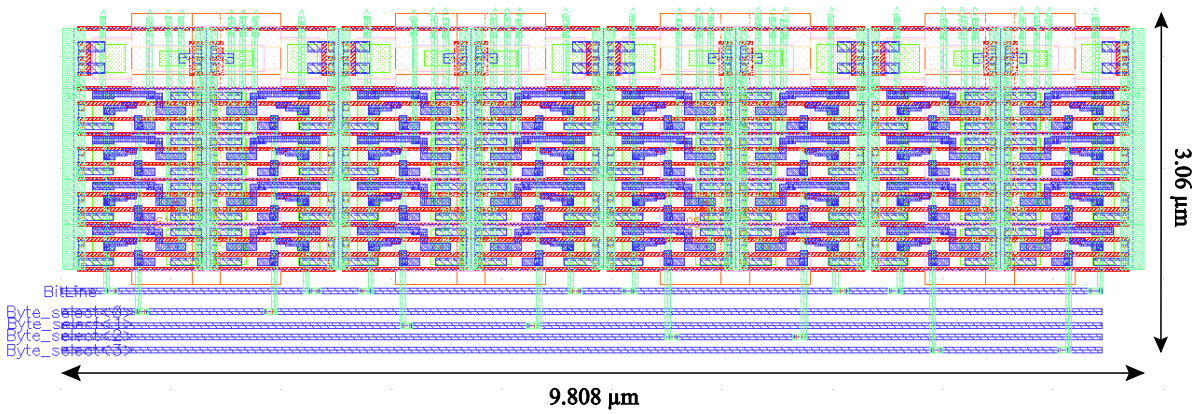


Figure 2.39 Layout of the byte selection block (BL_select). Its area is 9.808 µm^2×3.06 µm^2.

- If the enable input $BE = '0'$, all outputs $BL[0:31]$ are '1' regardless of the input combination $BS[0:3]$.

- If $BE = '1'$, this block performs NAND operation and allow a certain number of memory cells to be read or written. By controlling signals $BL[0:31]$, designers can read or program 8 bits (or 1 byte), 16 bits (or 2 bytes) or 32 bits (or 4 bytes), which increases the design flexibility.

2.4.3 Simulation of the basic blocks and the full 1KB MRAM

The functionality of the memory unit, peripheral blocks and full 1KB MRAM are validated by using the STMicroelectronics 28 nm FDSOI CMOS design kit and the PMA STT-MTJ compact model. The parameters of the STT-MTJ model are modified to reflect the reality of the fabrication process for magnetic devices: diameter $D = 200$ nm, tunnel magnetoresistance ratio $TMR = 0.4$, resistance area product $RA = 15\ \Omega\times\mu m^2$, Gilbert damping coefficient $\alpha = 0.01$, saturation field in the free layer $M_s = 19800$ Oe and the out of plane magnetic anisotropy $H_k = 1433$ Oe. 1 V supply voltage is applied to the whole circuit for searching, reading and writing. The simulated resistance of MTJ is about 477 Ω in parallel state and 668 Ω in anti-parallel state.

2.4.3.1 Simulation of the basic blocks

Each basic blocks forming the MRAM has been extensively simulated and evaluated before simulating the entire structure.

Figure 2.40 shows the transient simulation of the memory unit presented in Figure 2.31. $WL[0]='0'$ while other word line selection signals $WL[1:3]$ are set to '1', allowing data '1' to be written into a pair of MTJs (M_0 / M_1). The writing time and writing power consumption are 74.5 ns and 124.7 pJ with the optimized CMOS transistors size. $OUT='1'$ is obtained when RE is set to '1' and enables reading the data stored in MTJs. Simulation shows that this PCSA-based memory circuit has read speed of ~660 ps and read energy of ~371.7 nW (@5 MHz).

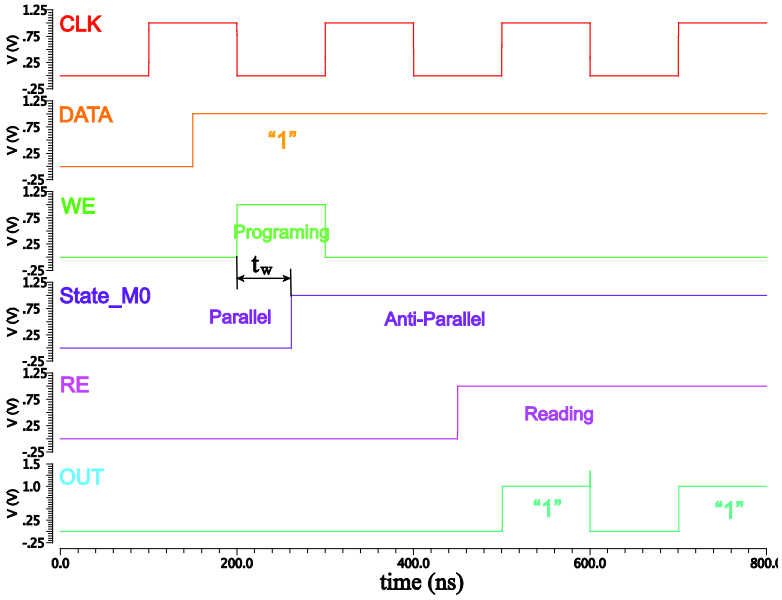


Figure 2.40 Transient simulation of the 4-bit memory unit

Figure 2.41 shows the simulation of the "Pre-decoder" block. Two-bit address inputs determine four-bit outputs, for example, $SA[0]='1'$ and $SA[1:3]="000"$ if $Addr[0:1]="00"$; the first word line is selected to be read or programmed. All the array selection signals $SB[0]$, $SC[0]$, $SD[0]$ are equal to '1' if $Addr[2:7]="000000"$, activating the first 16B array in the MRAM.

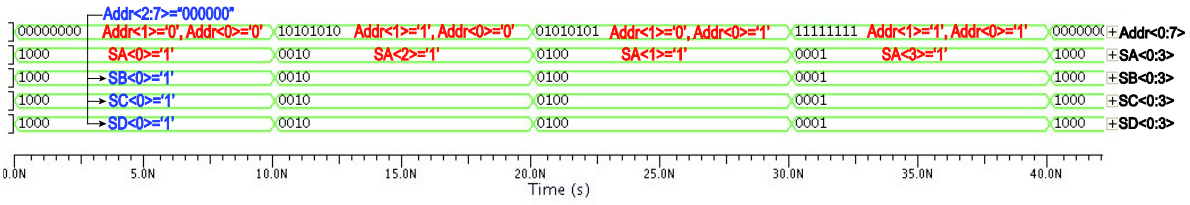


Figure 2.41 Simulation of the "Pre-decoder" block

Figure 2.42 shows the simulation of the byte selection block (BL_select). This block is enabled between M_0 and M_1. The first 8-bit outputs $BL[0:7]$ are obtained to be

"00000000" when $BS[0]='1'$, which enables read/write of the first octet in a line. $BS[1]$, $BS[2]$ and $BS[3]$ determine the read/write of the other three octets.

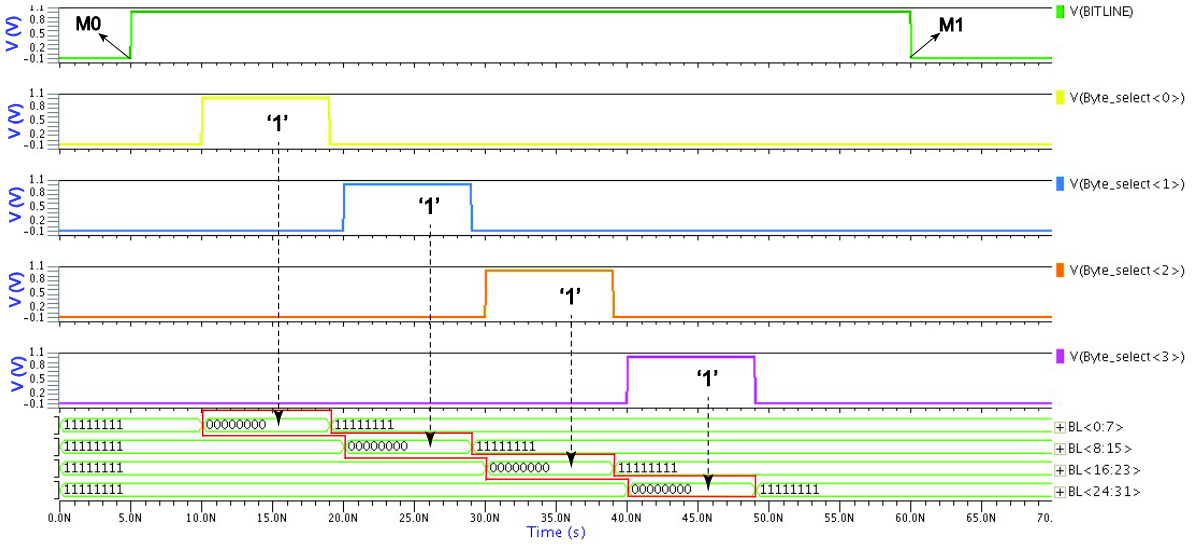


Figure 2.42 Simulation of the byte selection block (BL_select)

After demonstrating the full functionality of each block, we get further to the whole structure evaluation of the proposed MRAM.

2.4.3.2 Functional simulation of 1KB MRAM

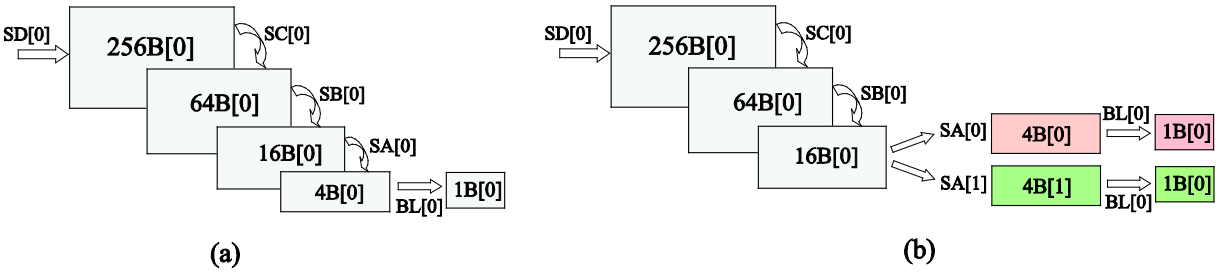


Figure 2.43 (a) Input address combination for bit/byte read and write validation (b) Input address combination for random read and write validation

In this sub-section, we validate the functionality of the 1KB MRAM.

1- *One byte reading and writing*: first, the combination of the row address $Addr[0:7]$ and column address $BS[0:3]$ are "00000000" and "1000", the first byte $1B[0]$ in the first word line $4B[0]$ of the subarray $16B[0]$ is activated (see Figure 2.43(a)).

2- *Random reading and writing*: second, we validate the random access of the 1KB MRAM. As can be seen in Figure 2.43(b), the lowest bits $Addr[0:1]$ are changed

from ‘”00” to “10”, switching between the context in two lines of one $16B[0]$ page.

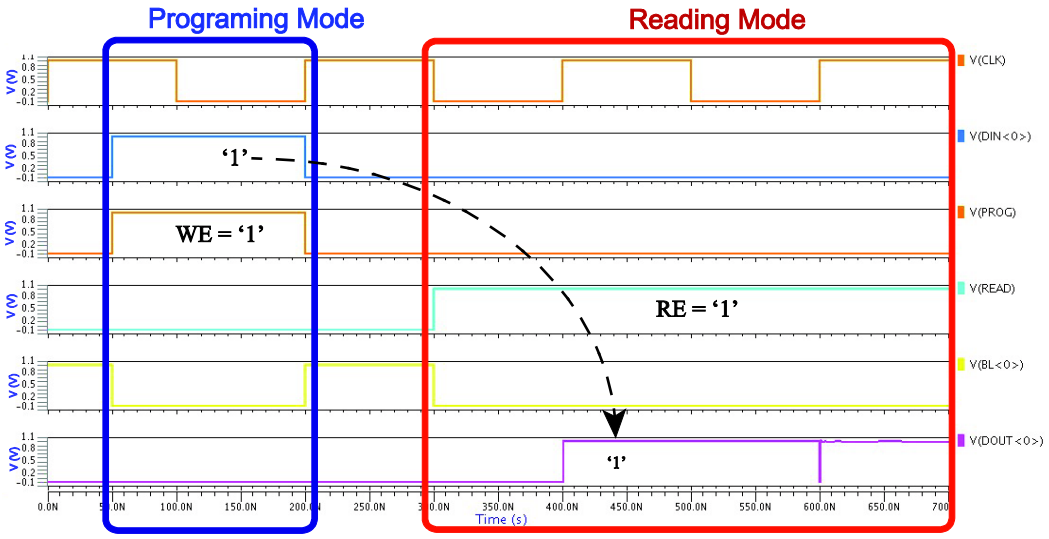


Figure 2.44 Simulation of the 1KB MRAM for single bit programing and reading

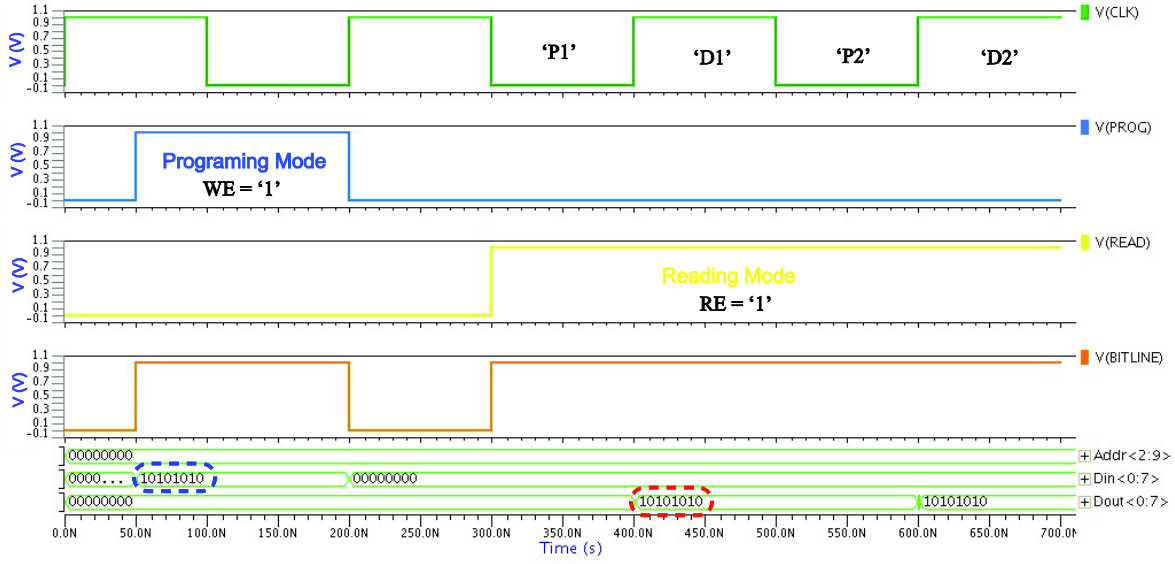


Figure 2.45 Simulation of the 1KB MRAM for one byte programing and reading

As shown in Figure 2.44, $Din[0]='1'$ is written into the first bit of $1B[0]$ when $WE='1'$, and then $Dout[0]='1'$ is obtained in the reading mode when $RE='1'$. Figure 2.45 shows the simulation of 1 byte reading and programing. 1) In the programing mode, for example, 8-bit input data “10101010” are written into eight pairs of MTJs by CMOS-based writing circuits. 2) In the reading mode, RE is set high while WE is set low. Data stored in these MTJs are read out ($Dout[0:7]="10101010"$) when meeting the rising edge of clock signal CLK. Then in the next pre-charge phase ($'P2'$), no switching current passing through the MJTs since WE keeps to be ‘0’. Therefore, the same output results “10101010” is obtained in the next

discharge phase ($'D2'$).

It is worth noting that output data stays the same even in the pre-charge phase of PCSA-based reading circuit. This is because the output driver is blocked if $CLK = '0'$ by signals $sense$ and its complement $\overline{sense}$ (see Figure 2.30). Simulation results show that the proposed peripheral circuits have row addressing time $t_{WL} \sim 1.74$ ns and column addressing time $t_{BL} \sim 880$ ps. 32 PCSA circuits perform sensing operation but only eight are output through output driver with a delay of 2.78 ns and dynamic energy of 0.033 W @100MHz.

In the second test, input data '1' and '0' are respectively written into the first bit in the first line $4B[0]$ and the second line $4B[1]$ of the subarray $16B[0]$ during the period (1) and (2) (see Figure 2.46). After the point M_0, RE se set high to enable reading the MRAM. WE is set low, hence, there will be no more writing. The stored data are read out in (3) and (4) with the output data $Dout[0] = '1'$ and $Dout[0] = '0'$, respectively. It confirms the switch between different storage memory cells.

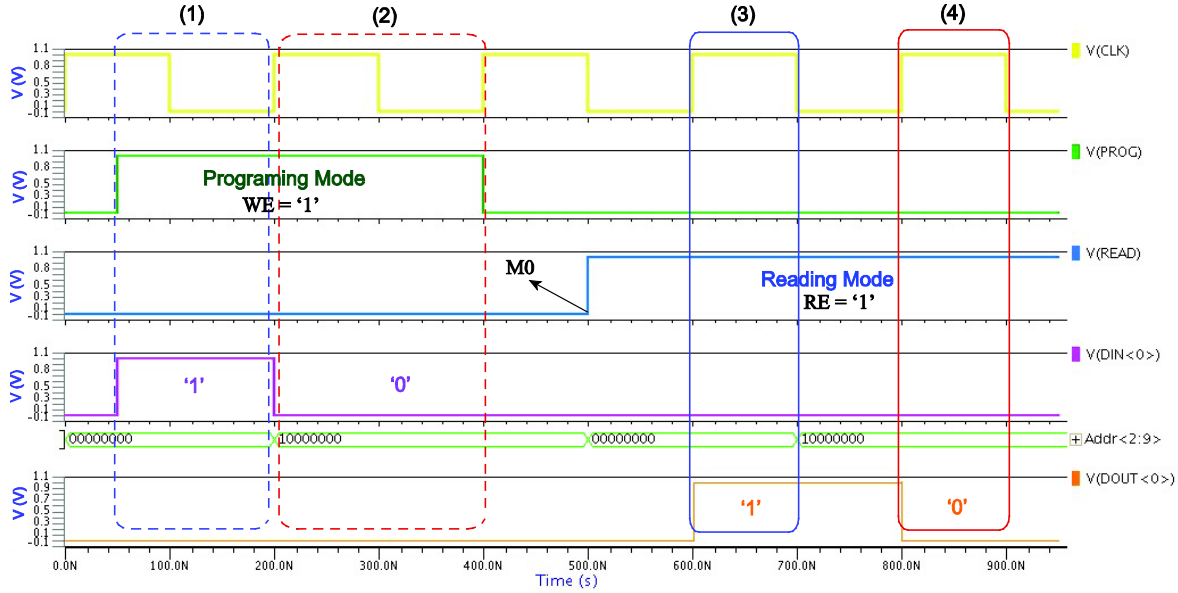


Figure 2.46 Simulation of the 1KB MRAM for random programing and reading

2.5 Conclusion

In this chapter, the compact model of PMA STT-MTJ was presented. It will be used in the following chapters for magnetic logic circuit design. PCSA is used to detect the magnetization configuration of MTJ, and 4T/6T writing circuits aim at switching the magnetization configuration of MTJ. By combining the reading and writing parts in the same circuit, the basic hybrid MTJ/CMOS circuit was designed and analyzed.

In order to ensure high reliable logic operations, structure-level and circuit-level optimizations of the multi-context hybrid MTJ/CMOS circuit were proposed. From the structure perspective, two multi-context hybrid MTJ/CMOS structures were proposed to integrate several MTJ nanopillars, i.e., sym-PCSA structure and sym-SPCSA structure. The former allows eliminating the limitation on the number of integrated MTJs of the conventional asymmetric structure, but with no advantages on reliability enhancement. The later one integrates a novel sensing circuit based on the symmetric structure to efficiently address the reliability issue caused by the scaled technology, which can hardly be achieved with the PCSA circuit. The evaluation results show that the proposed sym-SPCSA structure exhibits the best BER, whereas it consumes more power to perform the sensing operation. Some circuit-level design strategies were proposed to further optimize their reliability performance. The sym-SPCSA structure can reach a BER of zero by incorporating the three reliability design optimization methods with less area overhead than the other two PCSA based structures.

Finally, we proposed the design of a 1KB MRAM, which is based on the multi-context hybrid MTJ/CMOS circuit. A “Pre-decoder” block generates 16 address buses to select one line (32 bits). After that, a byte selection block is used for addressing a corresponding word (e.g., 8 bits) to be read or written. By using the MTJ model with modified parameters and the STMicroelectronics 28 nm FDSOI CMOS design kit, we validated its performances of addressing time (~ 1.74 ns), read time (~ 2.78 ns), read energy (~ 0.033 W/8 bits @100 MHz), write time (~ 74.74 ns) and write power consumption (~ 5.08 nJ/8 bits). The write time is relatively high due to the MTJ fabrication limitation of SPINTEC.

Chapter 3 Design of non-volatile logic circuits

As technology node shrinks below 45 nm, high static and dynamic power have become the major miniaturization obstacles for today's computing systems due to the increasing leakage currents and long data traffic between memory chip and logic units [75], [131]. Emerging hybrid logic-in-memory (LIM) architecture, where spintronics nanodevices are distributed over the logic-circuit plane, has recently been investigated to ensure ultra-low power and ultra-short interconnection delay. In such architecture, logic and memory functions are merged into the same spintronics nanodevices.

In order to fully take advantage of this architecture, the implemented non-volatile memory elements should have the capabilities of short access time (<10 ns), quasi infinite endurance (e.g., $>10^{12}$), small dimension and compatible resistance value with CMOS transistors (several kilohms) [123], [132]. STT-MTJ is an available candidate that can satisfy all the requirements and allow one to design hybrid non-volatile LIM-based circuits with high performances and new functionalities. Easy 3-D back-end integration of MTJs on top of CMOS technology [133], [134], [135] greatly shortens the distance between the memory and logic chips from millimeters to micrometers [136]. Consequently, this significantly reduces not only area overhead but also dynamic transfer power and latency compared to conventional systems.

Arithmetic logic unit (ALU) is one of the most important core execution parts in a central processing unit (CPU). In this chapter, LIM-based non-volatile ALU, combining MTJs with CMOS transistors, will be presented for low-power processor. The body of this chapter is composed of three sections. The general LIM architecture is introduced in the first section. Then, in the followings section, design and theoretical analysis of non-volatile logic gates (NVLGs), including NOT, AND, OR and XOR logic gates, are detailed. In the third section, low-power single-bit non-volatile full-adders (NVFA), the basic block of ALU, is presented and compared with the conventional CMOS-only FA to confirm its low-power advantage. The effect of discharge transistor size in the reading circuit, MTJ resistance-area product ($R \cdot A$) and TMR ratio on the delay time and dynamic power performances have been analyzed. In order to extend the single-bit NVFA to multi-bit case and also to realize full non-volatility, 8-bit NVFA architecture is then presented, where all the input signals are stored in MTJs. Three possible structures are proposed with respect to different locations of non-volatile data. Finally, voltage-mode sensing circuit (VMSC) and NVFA based on MTJ with spin-Hall assistance are proposed as a potential alternative to optimize the performances of NVFA in terms of area overhead, reliability and power consumption.

3.1 General logic-in-memory (LIM) architecture

As shown in Figure 3.1(a), the general logic-in-memory (LIM) architecture is mainly composed of three parts:

1) A current-mode sense amplifier (SA) to detect the currents of two branches and to evaluate the logic result on outputs. PCSA, whose operation mechanism has been described in Section 2.2.1, is used as the SA part in the design of hybrid logic circuits.

2) A writing block to program the data stored in non-volatile memory cells. It generates a bi-directional writing current I_W large enough to write the MTJs.

3) A logic network (LN) that performs the computation [137], [138]. LN contains MTJs that keep the non-volatile inputs and a CMOS logic tree for volatile inputs in order to keep an area-power-efficient advantage. In this case, the volatile logic data can be driven by high processing frequency contrarily to the non-volatile data, which should be changed with a relatively low frequency, i.e., they are quasi-constant for computing.

NMOS transistors and MTJs are the main components of LN (see Figure 3.1(b)).

- NMOS transistor is used as variable resistor, whose resistance is controlled by external volatile input voltage (X) applied to the gate (G) terminal. If $X = '1'$, NMOS transistor is conducted with a low resistance ($R_{ON} \sim \text{k}\Omega$). Otherwise, NMOS transistor is blocked and has a high resistance ($R_{OFF} \sim \text{G}\Omega$).

- MTJ cell is used not only as a storage element but also as an operand. The MTJ has a low resistance (R_P or R_L) and stores logic data '1' ($Y = '1'$) when it is in parallel state. If MTJ is in anti-parallel state, its resistance becomes high (R_{AP} or R_H) and it stores logic data '0' ($Y = '0'$). The values of R_P and R_{AP} can be controlled by changing the size of MTJ, and the resistance difference between two resistances depends on the TMR ratio.

By configuring the LN, different logic functions can be realized such as AND gate, XOR gate, etc. Two complementary outputs (z and z') correspond to two opposite logic values are produced, providing differential logic operations. The reading current (I_L or I_R) is inversely proportional to the total resistance (R_L or R_R) of the left or right branch in the LN. Thus, the

outputs are determined by the reading currents. If the current of the left branch is larger than that of the right branch ($I_L > I_R$), output results on nodes z and z' are '1' and '0', respectively. In contrast, $z = '0'$ and $z' = '1'$ if $I_L < I_R$.

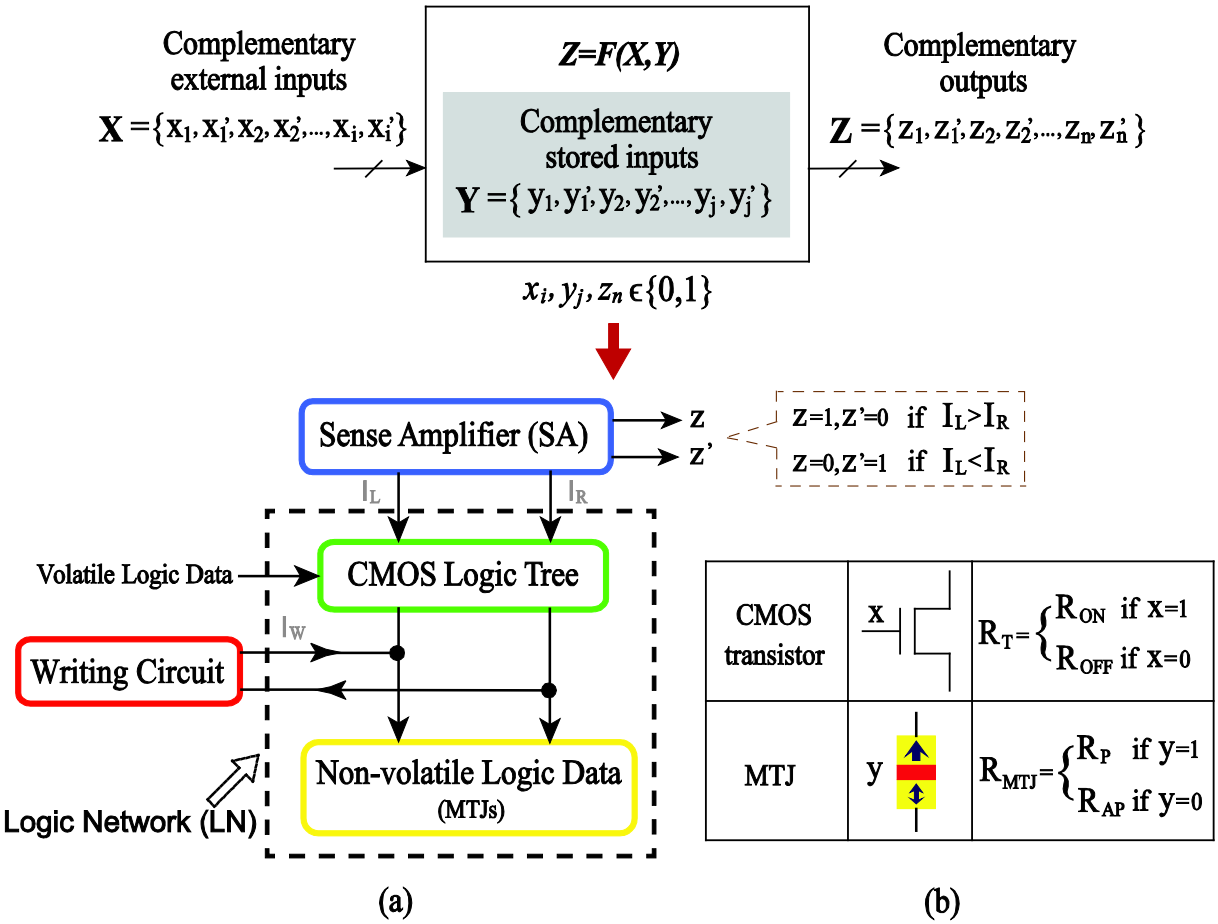


Figure 3.1 (a) Schematic of the logic-in-memory (LIM) architecture (b) Components in the logic network (LN)

Even though the reading (I_L and I_R) and writing currents (I_W) are produced in different paths for independent read/write operations, they flow through the same MTJs. Therefore, the reading currents should be design to be much smaller than the writing current (hundreds of micro-amperes) to avoid accidently writing during the read operation. Besides, there should be no steady current between the supply voltage (V_{dd}) and the ground during the "idle" state. PCSA is a promising candidate to satisfy all the requirements thanks to its low sensing current (tens of micro-amperes) and dynamic current-mode sensing, as it has been shown in Chapter 2 [23], [139].

3.2 Design and theoretical analysis of non-volatile logic gates

As discussed above, different logic operations can be realized by designing the CMOS logic tree shown in Figure 3.1(a). The output results of LIM-based logic circuits depend on the resistance configuration of both types of data in the logic network. Therefore, in this section, we design the basic non-volatile logic gates shown in Figure 3.2 and then analyze the impact of resistance configuration on logical operations.

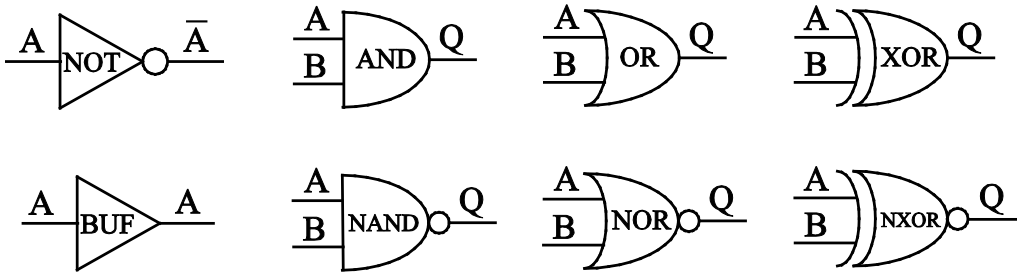


Figure 3.2 Symbols of logic gates

For the non-volatile NOT gate (NV-NOT) and the logic operation BUF (NV-BUF), there is only one input data that is stored in a pair of MTJs. For this reason, there is no need for the CMOS logic tree. We assume that the logic data '1' is stored in non-volatile state when the MTJ on the left side is parallel while the MTJ on the right side is anti-parallel. Other 2-input logic gates shown in Figure 3.2 need more complex design consideration to realize the corresponding logic function while keeping simpler structure. In the following sub-section, non-volatile AND/NAND gate will firstly be introduced and analyzed as an example.

3.2.1 Non-volatile AND/NAND gate (NV-AND/NV-NAND)

3.2.1.1 General NV-AND/NV-NAND structure and optimized structure-1

According to the truth table shown in Table 3.1, Eq. 3.1 and Eq. 3.2 can illustrate the logic function of AND/NAND logic. Figure 3.3(a) shows the LN structure designed directly from these equations. Nodes Qm, $\overline{Qm}$ and M are connected to the PCSA part, which will not be shown in all the following circuit schematics for simplifying the view. The left branch (LB) is constituted of an MTJ cell and a NMOS transistor connected in series. The right branch (RB) is composed of three sub-branches ($RB_0 - RB_2$) that are connected in parallel. The bottoms of all the left branch and right sub-branches are connected to the common node M. Any resistive level of MTJ allows correct non-volatile AND/NAND function. However, the

numerous NMOS transistors and MTJ cells lead to large die area. In particular, it needs complex writing circuits for programing the non-volatile data.

Table 3.1 Truth table of AND/NAND logic gate

A	B	Qm (AND)	$\overline{Qm}$ (NAND)
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

$$Qm = AB \qquad \text{Eq. 3.1}$$

$$\overline{Qm} = \overline{AB} = \overline{A} + \overline{B} = \overline{A}B + \overline{A}\overline{B} + A\overline{B} \qquad \text{Eq. 3.2}$$

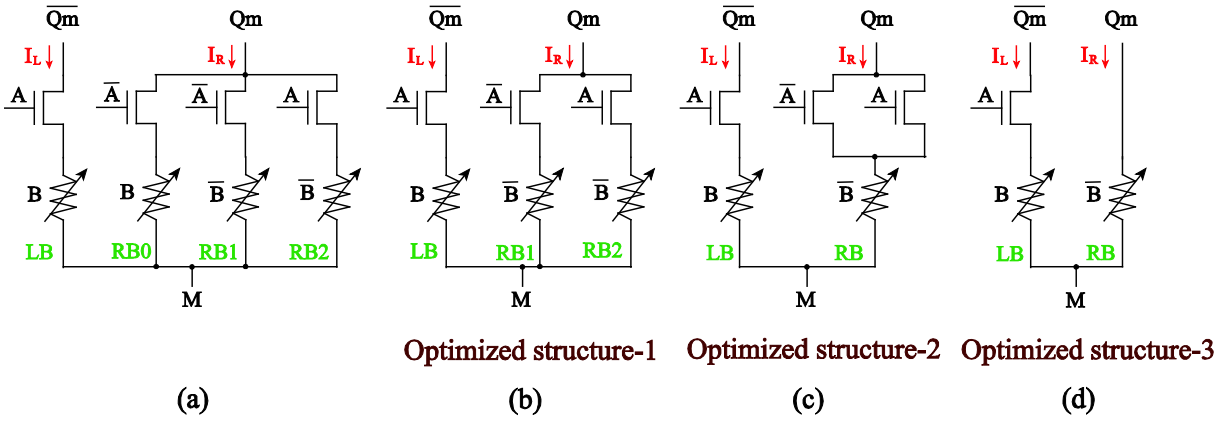


Figure 3.3 (a) General structure of the logic network for NV-AND/NV-NAND logic circuit (b) Optimized NV-AND/ NV-NAND structure-1 (c) Optimized NV-AND/ NV-NAND structure-2 (d) Optimized NV-AND/ NV-NAND structure-3

For the advanced CMOS technology [140], $R_{ON} \sim \text{k}\Omega$ and $R_{OFF} \sim \text{G}\Omega$. Thereby, R_{OFF} is much larger than R_{AP} (~kΩ) of MTJ and it determines directly the whole resistance of series connection. If $A = '0'$, RB_0 or RB_1 dominates the resistance of the right branch and the impact of RB_2 can be neglected. If $A = '1'$, RB_2 dominates the resistance of the right branch. Thereby RB_2 is critical, but one of the two sub-branches RB_0 or RB_1 can be deleted from the structure. In order to simplify the structure, we keep RB_1 and obtain the optimized structure-1 (see Figure 3.3(b)).

3.2.1.2 Optimized NV-AND/NV-NAND structure-2

The structure can be further optimized (structure-2) to be one MTJ for the right branch (see Figure 3.3(c)) with the condition that R_{OFF} is much larger than R_{AP} in the current technology. The equivalent resistance of the left and right branches can be expressed as $R_L = R_A + R_B$ and $R_R = \frac{R_A R_{\bar{A}}}{R_A + R_{\bar{A}}} + R_{\bar{B}}$.

Table B.1.a in Appendix B exhibits the truth table of structure-2 and the resistance configuration to allow correct AND logic function. When $R_L > R_R$, the output $Qm = '0'$; on the contrary, $Qm = '1'$. We can find that there are two uncertain cases. Depending on the resistance values of NMOS transistors and MTJs, the relationship between R_L and R_R is different, driving the result to the correct values or not. Table B.1.b illustrates resistance conditions to identify the uncertain cases for the optimized structure-2. It demonstrates that the difference between R_{AP} and R_P should be in the range $\{\sim \text{m}\Omega, \sim \text{G}\Omega\}$, which is large enough for MTJ devices.

Transient simulations have been performed to confirm this conclusion by using STT-MTJ compact model and CMOS 40 nm design-kit [140]. Figure 3.4 shows that the circuit of optimized structure-2 performs correctly the AND logic whatever the input configurations “00”, “01”, “10” or “11” are applied on A and B.

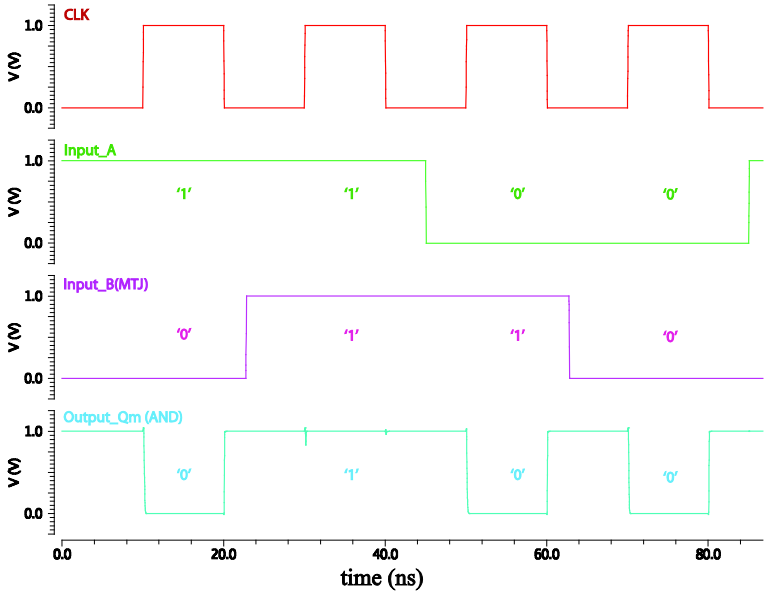


Figure 3.4 Transient simulation for optimized AND logic structure-2

3.2.1.3 Optimized NV-AND/NV-NAND structure-3

The parallel connected NMOS transistors respectively driven by data A and $\overline{A}$ (see Figure 3.3(c)) can be deleted considering the logic design strategy according to Eq. 3.3. As can be seen from Table B.1.a in Appendix B, the resistance of the right branch will always be the same whatever the input values of A and $\overline{A}$. Therefore, structure-2 can be further optimized and we can obtain simpler structure-3, where there is only one transistor in the left branch of the CMOS logic tree (see Figure 3.3(d)). The total resistance of the left and right branches can be expressed as $R_L = R_A + R_B$ and $R_R = R_{\overline{B}}$.

$$Qm = (A + \overline{A})\overline{B} = \overline{B} \qquad \text{Eq. 3.3}$$

By comparing the total resistance of the logic network shown in Table B.1.c and Table B.1.d, we can conclude that the resistance difference between R_{AP} and R_P should be in the range $\{R_{ON}, R_{OFF}\}$ to ensure the AND logic, which is much more rigorous than the range of optimized structure-2.

We performed transient simulations to validate the optimized structure-3. In Figure 3.5, an error appears when A and B are both '1'. This is caused by low resistance value of MTJ and limited TMR ratio, which is out of the acceptable range of the optimized structure-3.

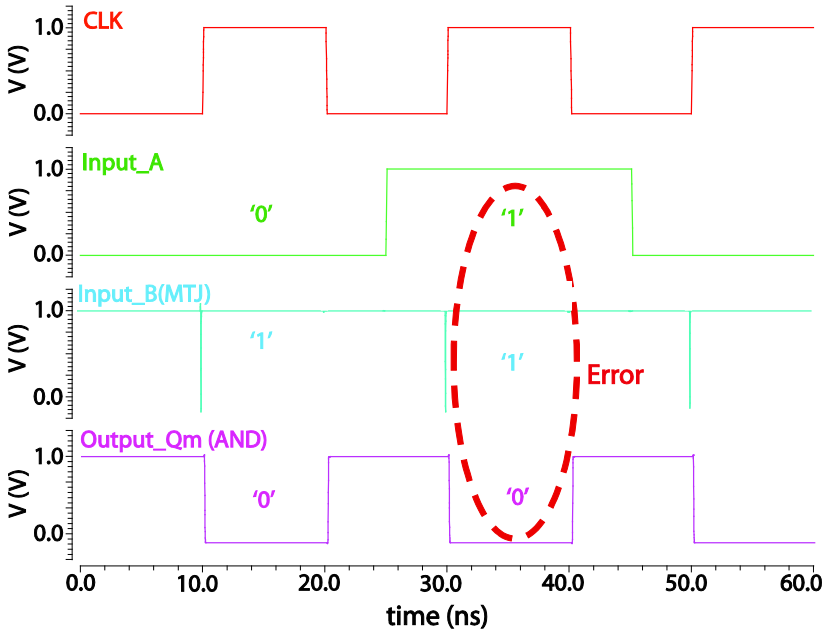


Figure 3.5 Transient simulation for the optimized NV-AND/NV-NAND structure-3. An error appears as input data $AB = "11"$.

When comparing all the optimized structures for AND/NAND logic operation, the optimized structure-2 and structure-3 saves one MTJ cell and thus consuming less writing power. Although the optimized structure-3 performs further area-efficiency, it is adapted to the MTJ cell in condition that the MTJ resistance meets the range criteria.

We use the same way to design and analyze the non-volatile OR/NOR logic gate and the non-volatile XOR/NXOR logic gate in the following two sub-sections.

3.2.2 Non-volatile OR/NOR gate (NV-OR/NV-NOR)

The truth table of OR/NOR logic is shown in Table 3.2. The general structure shown in Figure 3.6(a) of this logic gate is designed directly from Eq. 3.4 and Eq. 3.5. The optimized method (structure-1) shown in Figure 3.6(b) is similar to that of the NV-AND/NAND gate, which eliminates the sub-branch LB_0.

Table 3.2 Truth table of OR/NOR logic gate

A	B	Qm (OR)	$\overline{Qm}$ (NOR)
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

$$Qm = \overline{A}B + A\overline{B} + AB \qquad \text{Eq. 3.4}$$

$$\overline{Qm} = \overline{A}\,\overline{B} \qquad \text{Eq. 3.5}$$

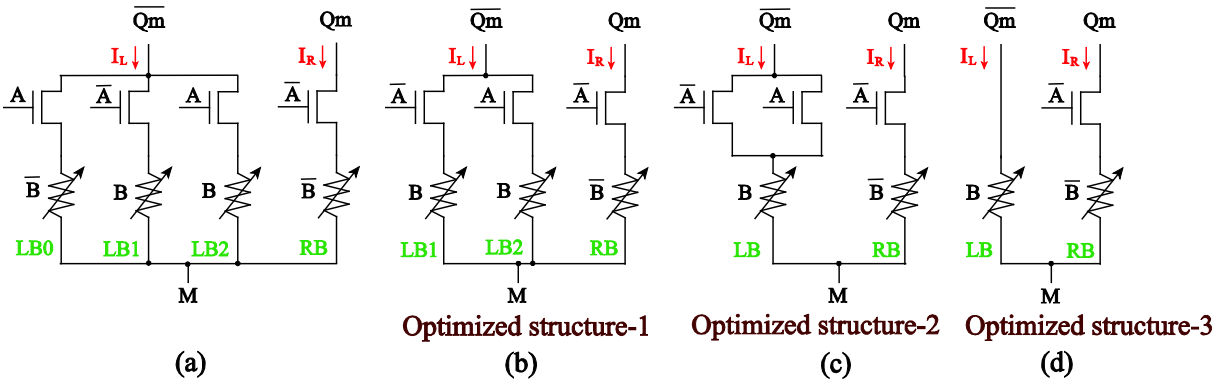


Figure 3.6 (a) General structure of the logic network for NV-OR/NV-NOR logic circuit (b) optimized NV-OR/ NV-NOR structure-1 (c) optimized NV-OR/ NV-NOR structure-2 (d) optimized NV-OR/ NV-NOR structure-3

The second optimized structure (structure-2) is shown in Figure 3.6(c) with one less MTJ in the left branch. Table B.2.a in Appendix B exhibits the truth table of the optimized NV-OR/NV-NOR structure-2 and the resistance configuration to allow correct OR logic function, and Table B.2.b illustrates resistance conditions to identify the two uncertain cases. It has the same range of $R_{AP} - R_P$ as the optimized NV-AND/NV-NAND structure-2, i.e., $\{\sim \text{m}\Omega, \sim \text{G}\Omega\}$.

For the optimized structure-3 (see Figure 3.6(d)), the comparison of total resistances in the logic network and the resistance condition of the uncertain cases are listed in Table B.2.c and Table B.2.d. In order to ensure correct OR logic operation, the resistance difference between R_{AP} and R_P should be in the range $\{R_{ON,}\ R_{OFF}\}$.

3.2.3 Non-volatile XOR/NXOR gate (NV-XOR/NV-NXOR)

The general structure of non-volatile XOR/NXOR circuit shown in Figure 3.7(a) is designed based on the truth table Table 3.3 and the arithmetic equations Eq. 3.6 and Eq. 3.7. This structure is suitable for all input configurations, however, needs large energy and writing circuit area to change the state of four MTJs.

Table 3.3 Truth table of XOR/NXOR logic gate

A	B	Qm (XOR)	$\overline{\text{Qm}}$ (NXOR)
0	0	0	1
0	1	1	0
1	0	1	0
1	1	0	1

$$Qm = \overline{A}B + A\overline{B} \qquad \text{Eq. 3.6}$$

$$\overline{Qm} = \overline{A}\overline{B} + AB \qquad \text{Eq. 3.7}$$

We can find that during the reading operation, Qm and $\overline{Qm}$ will never access the same MTJ cell (B or $\overline{B}$) whatever the value of A and $\overline{A}$. We can then obtain the optimized structure integrating only two MTJs, as shown in Figure 3.7(b). The optimized NV-XOR/NV-NXOR structure has a complex CMOS logic tree, where four transistors are cross-connected 2 by 2. In order to obtain the equivalent resistance of two branches, we did a

series of circuit transformations. And the resistance difference between the two branches can be described by Eq. 3.8, where the factor R_{ref} is a positive function of R_A, $R_{\bar{A}}$, R_B and $R_{\bar{B}}$. It is obtained by the Kirchhoff's current law [141] and Y-Δ transform described by Arthur Edwin Kennelly in 1899 [142]. The resistance difference depends on the values of A and B. Therefore, this optimized structure is suitable for all MTJs even with small TMR ratio.

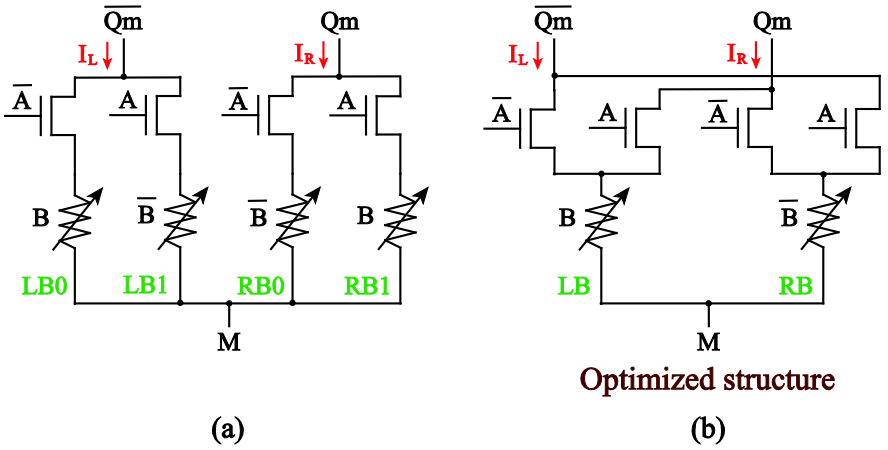


Figure 3.7 (a) General structure of the logic network for NV-XOR/NV-NXOR logic circuit (b) optimized NV-XOR/ NV-NXOR structure

$$\Delta R = R_L - R_R = R_{ref}(R_A - R_{\bar{A}})(R_B - R_{\bar{B}}), \quad R_{ref} > 0 \qquad \text{Eq. 3.8}$$

3.3 Design and optimization of low-power non-volatile full-adder (NVFA)

A single-bit full-adder (FA) is a three-input (A, B, C_i) two-output (SUM, C_o) circuit (see Figure 3.8). It is the basic building unit to perform arithmetic operation in a central processing unit (CPU). Therefore, the investigation of non-volatile full-adder (NVFA) is important for the purpose of building low-power high-density processors. This block can be connected to others to reform a more complex function.

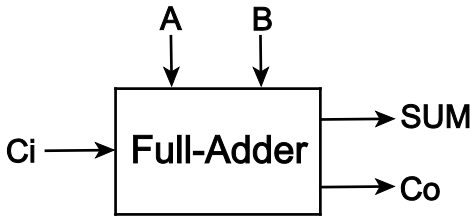


Figure 3.8 Symbol of single-bit full-adder (FA)

3.3.1 1-bit NVFA

Several single-bit NVFAs based on non-volatile memory have been proposed and exhibit satisfying properties [90], [137], [143]. However, the use of capacitance for data sensing and magnetic field for data programming limits further miniaturization. The PCSA-based NVFA proposed in [137] could lead to ultra-low power and high density ICs. But the inherent heating of this structure is contrary to power saving objective. In order to overcome these issues, we propose a novel 1-bit NVFA.

3.3.1.1 Structure and theoretical analysis of 1-bit NVFA

The CMOS logic tree of the designed NVFA is designed according to Eq. 3.9-Eq. 3.12.

$$SUM = A \oplus B \oplus C_i = ABC_i + A\overline{B}\,\overline{C_i} + \overline{A}B\overline{C_i} + \overline{A}\,\overline{B}C_i \quad \text{Eq. 3.9}$$

$$\overline{SUM} = AB\overline{C_i} + \overline{A}\,\overline{B}\,\overline{C_i} + \overline{A}BC_i + A\overline{B}C_i \quad \text{Eq. 3.10}$$

$$C_o = AB + AC_i + BC_i \quad \text{Eq. 3.11}$$

$$\overline{C_o} = \overline{A}\,\overline{B} + \overline{A}\,\overline{C_i} + \overline{B}\,\overline{C_i} \quad \text{Eq. 3.12}$$

For SUM logic, the CMOS tree corresponds directly to the logic relationship among the inputs A, B and C_i, we can simply adapt it to the general structure with a couple of complementary MTJs (see Figure 3.9). Based on the same strategy that we used in the case of simple non-volatile XOR logic gates, ΔR of SUM sub-circuit can be calculated by Eq. 3.13 and the R_{ref} is a positive function of R_A, $R_{\overline{A}}$, R_B, $R_{\overline{B}}$, R_{C_i} and $R_{\overline{C_i}}$.

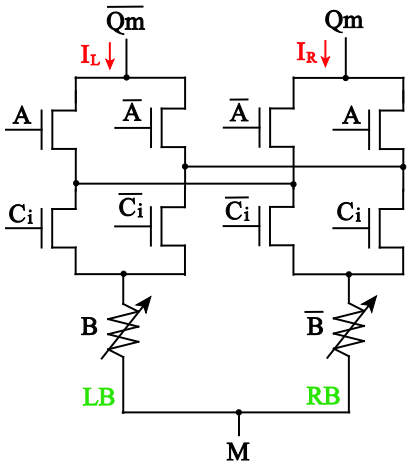


Figure 3.9 Structure of the logic network for SUM sub-circuit

$$\Delta R = R_L - R_R = R_{ref}(R_A - R_{\overline{A}})(R_B - R_{\overline{B}})(R_c - R_{\overline{c}}), \quad R_{ref} > 0 \qquad \text{Eq. 3.13}$$

It is a little more difficult for CARRY/C_o logic as there is the term AC_i in the logic function Eq. 3.11 that cannot be adapted to the general LIM structure. It can be inferred that the impact of the term AC_i on the resistance is equivalent to a sub-branch connecting PCSA and the discharging transistor, Table B.3.a in Appendix B shows the truth table and the resistance configuration of the CARRY logic as well as the AC_i and $\overline{AC_i}$ tails. We can find that whatever the value of A and C_i, the sub-branches AC_i have no impact on the output. If A and C_i are different, the resistance of the two sub-branches is the same; if they are the same, their comparison corresponds to that of R_L and R_R, which is always true for MTJs. This allows the term AC_i to be deleted from Eq. 3.11 and we can obtain the CARRY logic circuit shown in Figure 3.10(a). In this structure, two NMOS transistors are parallel connected and then serially connected with an MTJ. Based on the analysis of difference input cases, ΔR should be in the range of $\{0, \sim G\Omega\}$, which is large enough for the current MTJ technology.

Another CARRY sub-circuit structure is shown in Figure 3.10(b), where two NMOS

transistors are connected in series on both sides. After analyzing the resistance condition of this structure for CARRY logic (see Table B.3.d and Table B.3.e), we find that structure-2 also has a large resistance range $\{0, \sim \mathrm{G}\Omega\}$.

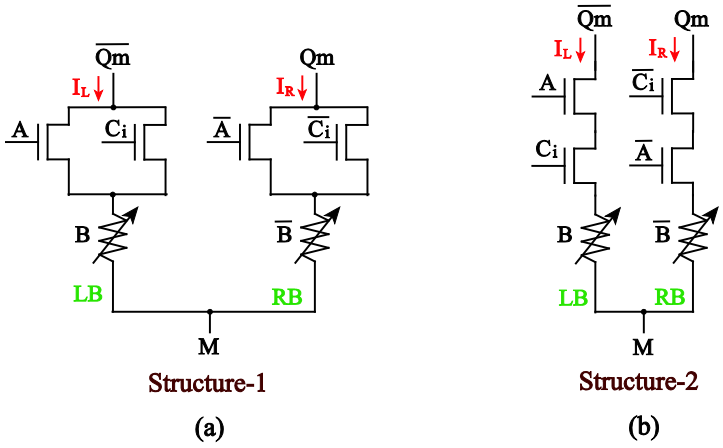
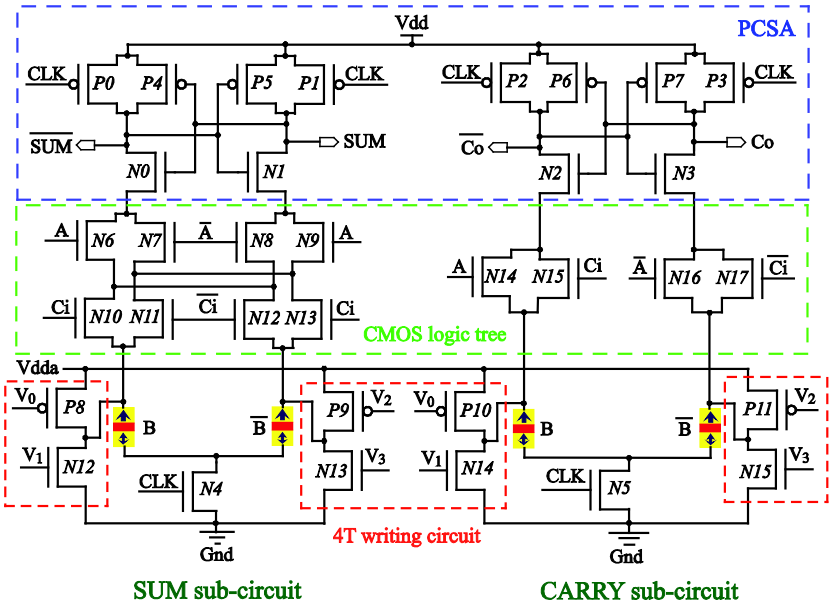


Figure 3.10 Logic network for CARRY sub-circuit (a) structure-1 (b) structure-2



Truth table of SUM and CARRY logic

A	B	Ci	SUM	Co
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

Figure 3.11 Full schematic of the 1-bit non-volatile full-adder (NVFA)

Figure 3.11 shows the full circuit of 1-bit NVFA by combining the SUM sub-circuit and the structure-1 of the CARRY sub-circuit. A ($\overline{A}$: the complement of A) and C_i ($\overline{C_i}$: the complement of C_i) are volatile inputs and B ($\overline{B}$) is non-volatile input. CLK synchronizes the results of this computing unit as the clock. The MTJs in both SUM sub-circuit and CARRY sub-circuit are always in opposite states to ensure the necessary high sensing speed and they are serially connected with a common central point. In order to program the MTJ cells, we use two 4T writing circuit described in Section 2.2.2.1.

Figure 3.12 illustrates the simulation of the 1-bit NVFA by using the MTJ compact model introduced above and the CMOS 40nm design kit [140]. The time-dependent behaviors of outputs (SUM and C_o) confirm the logic functionality of a full addition. For instance, for the operation $A='1', B='0', C_i='0'$, the result SUM is '1' and no carry propagates; for the operation $A='1', B='0', C_i='1'$, the result is '0' for SUM and '1' for CARRY.

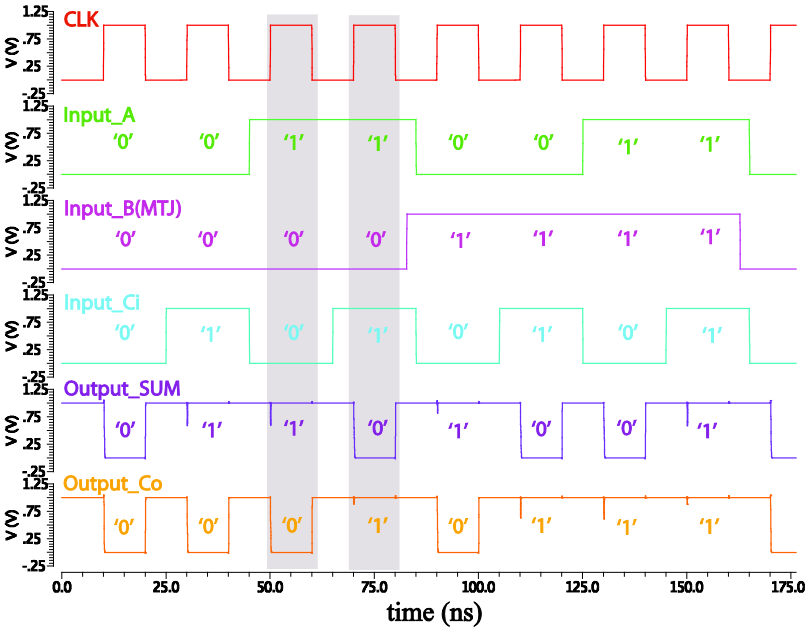


Figure 3.12 Functional simulation of 1-bit NVFA at 40 nm technology node

3.3.1.2 Performance analysis and comparison

The delay time and dynamic energy are generally two crucial parameters to evaluate the performance of computation system. We have studied the effects of three possible factors: the size of discharge transistor (N_4 and N_5 in Figure 3.11), MTJ resistance-area product (RA) and TMR ratio. Figure 3.13(a) demonstrates the performance dependence of this NVFA in terms of delay and dynamic power on the size of discharge transistor. We can find a tradeoff between the speed and power performance by varying the die area. A larger discharge transistor can drive a higher sensing current and faster amplification of PCSA circuit, but it generates more energy. Figure 3.13(b) shows the RA dependence for this NVFA. By decreasing RA, the delay time becomes shorter while keeping a relatively steady dynamic power performance since the current is larger. This confirms that using a low RA gives better speed. We also investigate the dependence between TMR ratio and NVFA performance. Figure 3.13(c) shows that higher speed is possible by increasing the TMR ratio while the dynamic energy changes slightly. According to the above analyses, a MTJ with lower RA

and higher TMR ratio is expected to perform fast computation while keeping nearly the same dynamic energy.

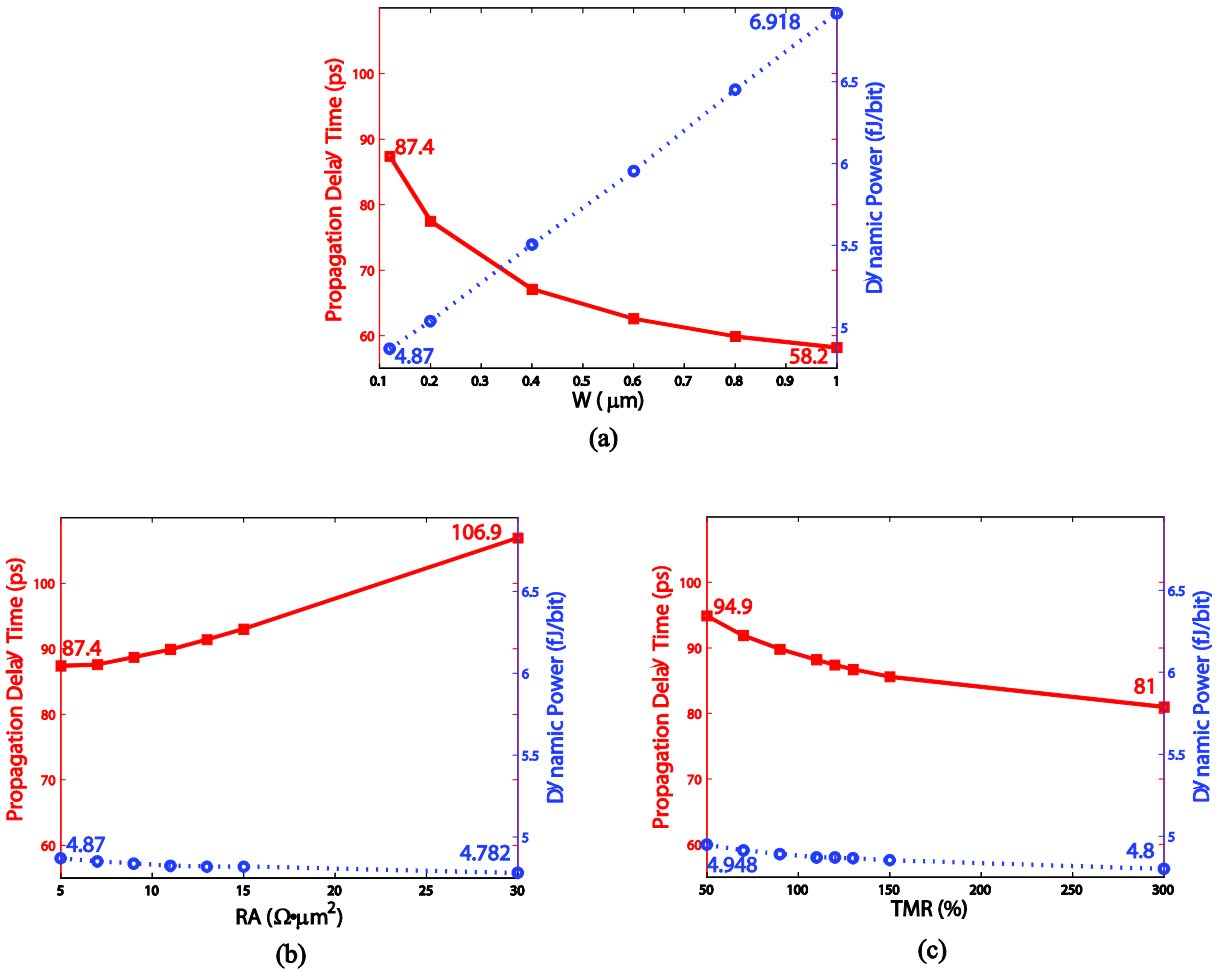


Figure 3.13 The dependence of propagation delay time (red solid line) and dynamic energy (blue dotted line) on the (a) width of discharge transistor (W) (b) MTJ resistance-area product (*RA*) (c) TMR ratio

We compare the 1-bit NVFA with conventional CMOS-only FA taken from the standard cell library in terms of sensing time, dynamic power, standby power, data transfer energy and die area (see Table 3.4). Thanks to the 3-D integration of MTJ, the die area of this design is advantageous compared to those of the CMOS full-adder. The data transfer energy becomes much lower thanks to the shorter distance between memory and computing unit. However, its energy-delay product (EDP) exceeds that of a CMOS full-adder by approximately 10% since it takes more time for PCSA amplification process. Thanks to the non-volatility of MTJ, the new chip can be powered off completely and this allows the standby power to be reduced significantly down to 0.75 nW. There is neither capacitance for the data sensing and nor magnetic field for data programming in this new structure beyond the previous structures [90], [137], [143]. Therefore, this design allows efficient area minimization and is suitable for

advanced fabrication nodes below 65 nm.

Table 3.4 Comparison of the 1-bit NVFA with CMOS-only FA

Performance	**CMOS FA**	**NVFA**
Sensing time	75 ps	87.4 ps
Dynamic power (@500MHz)	2.17 μW	1.98 μW
Standby power	~ nW	~ 0 [144]
Data transfer energy	> pJ/bit	< fJ/bit
Die area	46T	38T + 4 MTJ

3.3.2 Multi-bit NVFA

Single-bit NVFA based on the LIM architecture has been investigated in the previous sub-section. However, this FA is partial non-volatile. For the purpose of extending single-bit NVFA to multi-bit structure and realizing full non-volatility, 8-bit NVFA architecture is presented in this sub-section, where all the input signals are stored in MTJs. Three possible structures are proposed with respect to different locations of non-volatile data.

3.3.2.1 Structure of 8-bit NVFA

Three 8-bit NV-FAs are proposed where different the locations of non-volatile data are analyzed. Full structural schematics as well as the locational distributions of non-volatile data are illustrated in Figure 3.14. The architecture of 8-bit NVFA is composed of one half-adder (HA) and seven FAs serially connected, performing addition operation of two 8-bit words A ($A_7 - A_0$) and B ($B_7 - B_0$). A and B are all stored in non-volatile states while carry-in $C_i(i+1)$ is connected to the previous carry-out $C_o(i)$. The final 9-bit output includes eight SUM bits ($SUM_7 - SUM_0$) and one CARRY bit (C_{out}). It should be noted that the first structure (Structure-1) is designed based on traditional CMOS-only HA and FA, while the other structures, i.e., Structure-2 in Figure 3.14(b) and Structure-3 in Figure 3.14(c), use the aforementioned non-volatile FA and HA to perform addition operation.

In Figure 3.14(a), two 1-bit non-volatile flip-flops (NVFF) and a register are added to input and output nodes of each CMOS-only addition cell, resulting in large area overhead and energy consumption. NVFFs are used to generate and store non-volatile input A and B. In Figure 3.14(b), Structure-2 stores input A and B in MTJs that are embedded in NVFFs and non-volatile adders, respectively. The number of NVFFs is thus reduced from sixteen to

eight. The use of registers dedicates to CARRY transition between two additions. There is no need to add latch circuits at the outputs (SUM and C_{out}) since the non-volatile adders are naturally synchronized. In Figure 3.14(c), Strucutre-3 is able to store 8-bit input data A with sixteen MTJs and read by a multi-bit NVFF, which saves more area than Structure-2. The disadvantage is that only 1-bit data (e.g., A_0) can be read or written during one operation. The CMOS switches are controlled by three external signals $S_2S_1S_0$ through a CMOS-based 3-8 decoder.

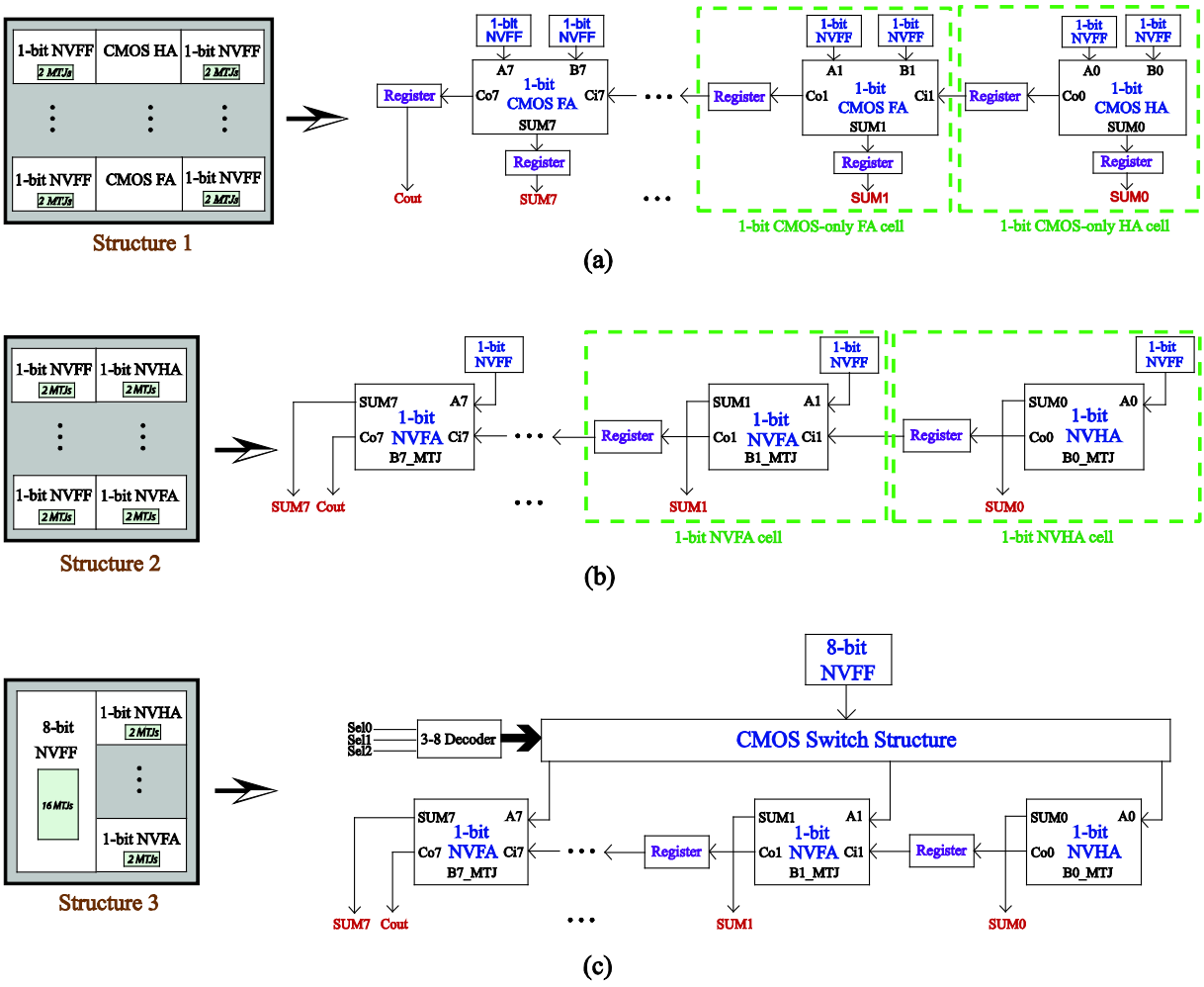


Figure 3.14 Locational distributions of non-volatile data and full schematics of the proposed 8-bit NVFA structures (a) Structure-1: A and B are stored in non-volatile flip-flops (NVFFs) (b) Structure-2: 8-bit data B are stored in MTJs embedded in non-volatile adders while data A are stores in 8 NVFFs (c) Structure-3: 8-bit data A are all stored in an 8-bit NVFF circuit for area cost reduction

Input A of the Structure-2 is stored in 1-bit NVFFs, while that of the Strucutre-3 is stored in a 8-bit NVFF. The 1-bit NVFF can be designed from the hybrid MTJ/CMOS circuit shown in Figure 2.16, by adding a CMOS latch at the output buffer for data transition. The 8-bit NVFF,

however, is a little different from the multi-context hybrid MTJ/CMOS circuit shown in Figure 2.20. As can be seen from Figure 3.15, the MTJs on two sides are completely separated and one more discharge transistor MN_5 is added. In such configuration, two MTJs are programmed in individual current-flow path. The raison is that the 4T writing circuit is not practical anymore for the 8-bit NVFF due to the extra NMOS transistors for MTJ selection. The size of writing transistors will be largely increased in order to generate a big enough current for MTJ switching, which leads to significant area overhead and limited writing current increase.

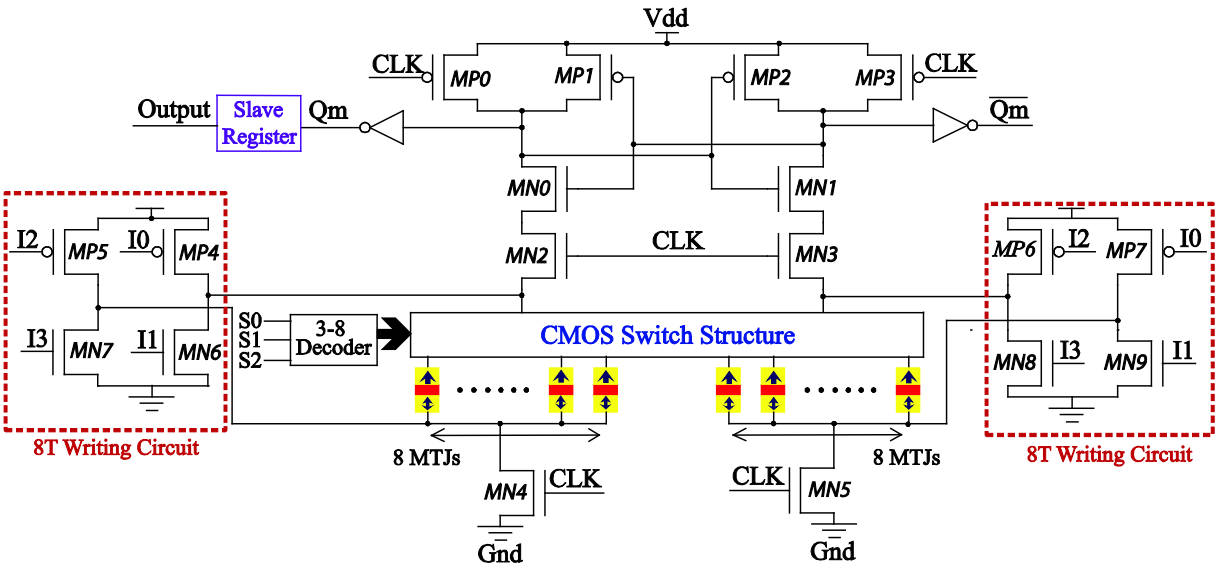


Figure 3.15 Full schematic of 8-bit NVFF. During a sensing operation, only one out of eight NMOS transistors in the left sub-branch and another in the right sub-branch are turned ON to connect the upper PCSA part with the addressed MTJs.

Traditional 1-bit CMOS-only HA and FA, the basic addition cells of Structure-1, are taken from the standard cell library of STMicroelectronics 28 nm design kit (see Appendix C). We use the NVFA shown in Figure 3.11 to perform addition operation of Structure-2 and Structure-3. According to the equations Eq. 3.14 and Eq. 3.15, the SUM sub-circuit and the C_o sub-circuit of NVHA can be designed from the NV-XOR logic gate and the NV-AND logic gate, respectively (see Figure 3.16).

$$SUM\_NVHA = \overline{A}B + A\overline{B} \qquad \text{Eq. 3.14}$$

$$C_o\_NVHA = AB \qquad \text{Eq. 3.15}$$

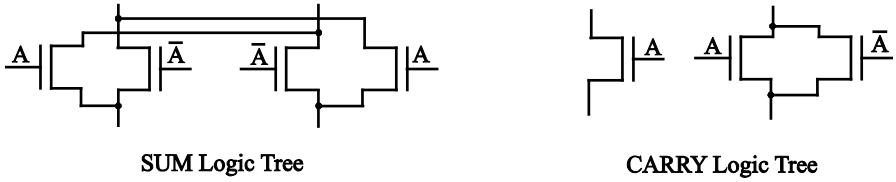


Figure 3.16 CMOS logic tree diagrams of 1-bit NVHA

3.3.2.2 Simulation of 8-bit NVFA

Transient simulations are performed at CMOS 28 nm technology node to validate the functionalities of the proposed 8-bit NVFAs. $TMR(0)$ is set to 200%. The diameter of the MTJs is 32 nm and the MTJ resistances are $R_P \approx 6.2\ k\Omega$ and $R_{AP} \approx 18.6\ k\Omega$.

Figure 3.17 shows the logic behavior simulation of the 1-bit NVFF. In this simulation, two MTJs (MTJ_0 and MTJ_1) are initialized at logic '1' and '0', respectively. During the first cycle, write enable signal $WE = '0'$ and no writing current passes through the MTJs. Non-volatile data stored in MTJs is read and propagated to Qm node during the evaluation phase (E) with a delay as low as 132 ps. This data is then propagated to $Output$ node of the slave latch when CLK meets a falling edge. During the pre-charge next phase, writing operation is activated ($WE = '1'$) and the configuration of MTJ_1 is switched from anti-parallel to parallel. The previously detected output data is retained during this phase.

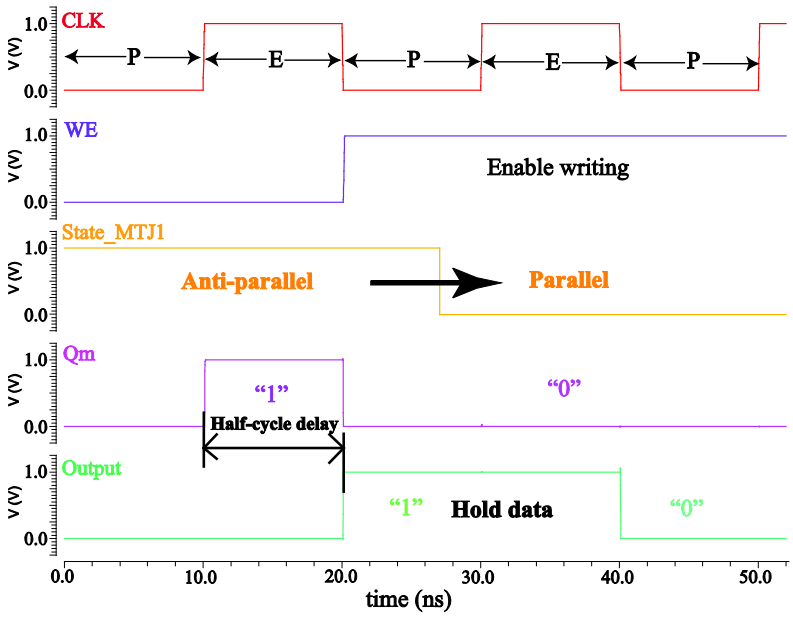


Figure 3.17 Transient simulation of the 1-bit NVFF. Qm and Output are signals before and after the slave latch.

Figure 3.18 shows the transient simulation of the 8-bit NVFF, where all the combination of signals $S_0 - S_2$ are used for sequential reading of the MTJ pairs. Only 1-bit data can be read or written during one cycle. This circuit has a higher sensing time (~170 ps) than the 1-bit NVFF since there is one more transistor for MTJ selection in each current path, which reduces the sensing current.

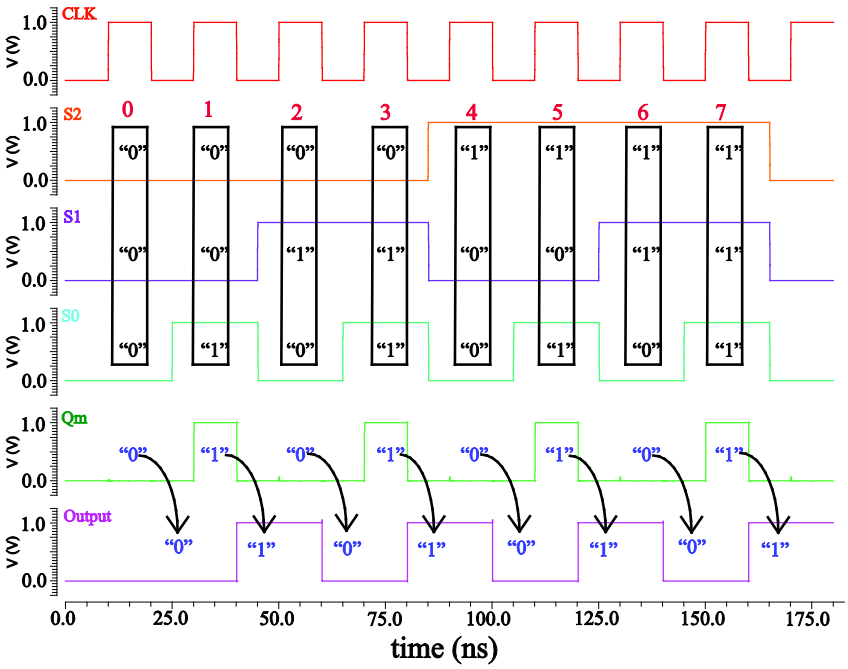


Figure 3.18 Transient simulation of the 8-bit NVFF ("01010101" are stored in the MTJs as an example)

Figure 3.19 shows the simulation waveforms of the proposed 8-bit NVFA (Structure-1). During period (1), inputs A and B integrated in NVFFs are programmed when $CLK = '0'$ and then evaluated when $CLK = '1'$. During period (2), data are first transferred to the inputs of adders through slave registers, and then performing the adding operation when meeting a rising edge of CLK. The final results, SUM_0 and C_{o0}, are transferred to outputs through two registers during period (3). Serial addition is then performed cycle-by-cycle. For example, two 8-bit words $A_7 - A_0 = "00000001"$ and $B_7 - B_0 = "11111111"$ are applied to the circuit, it is confirmed that the expected outputs are observed as $SUM_0 = '0'$, $SUM_1 = '0'$, $SUM_2 = '0'$, $SUM_3 = '0'$, $SUM_4 = '0'$, $SUM_5 = '0'$, $SUM_6 = '0'$, $SUM_7 = '0'$, $C_{out} = '1'$. Thus, Carry bit from the lowest bit propagates all the way through to the highest bit and the whole propagation chain is activated.

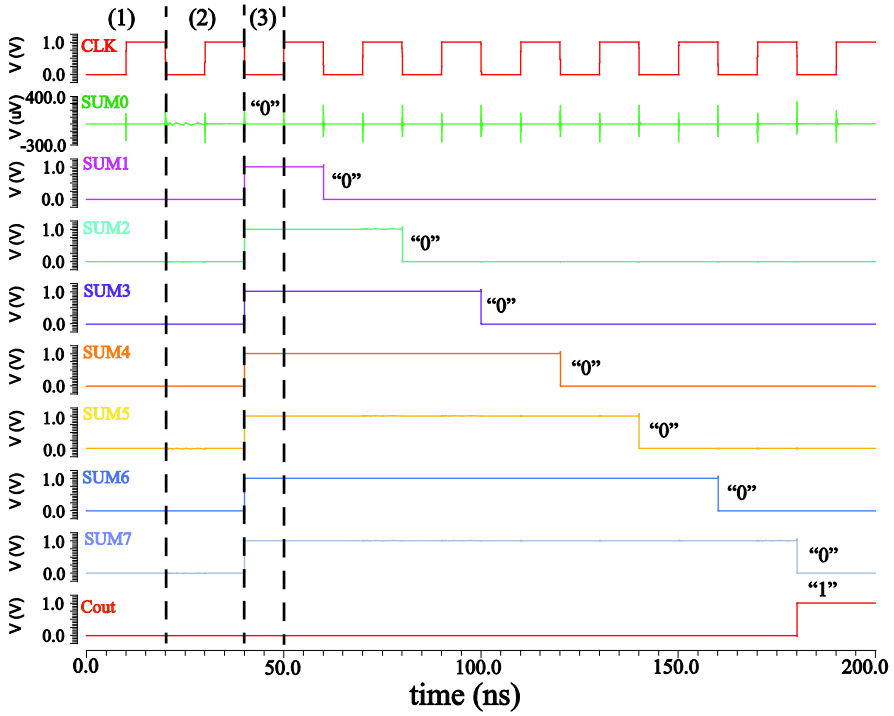


Figure 3.19 Functional simulation of the synchronous 8-bit NVFA (Structure-1)

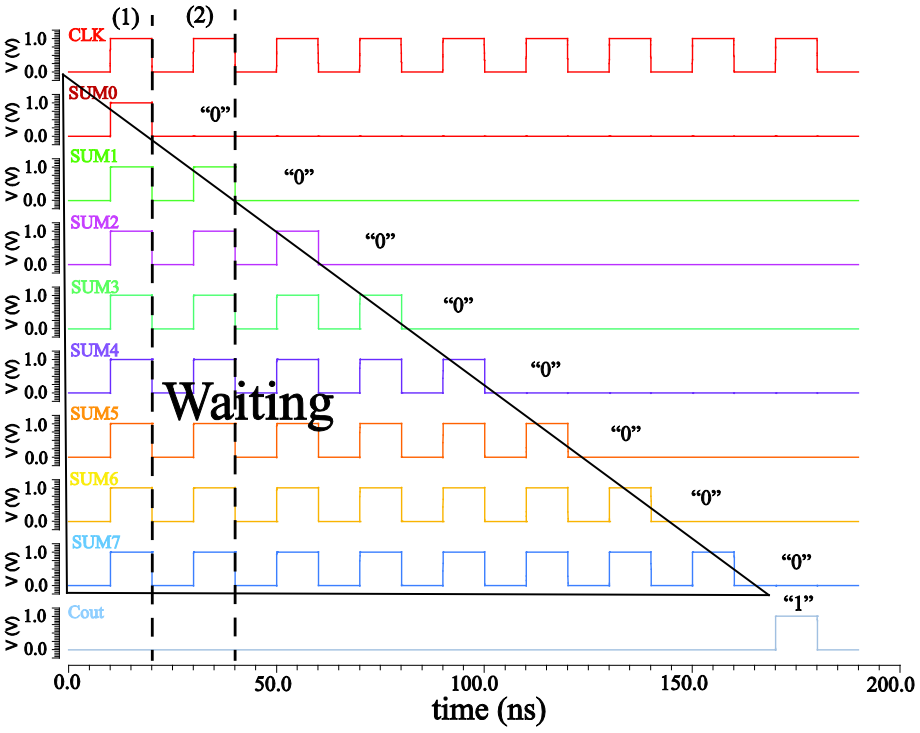


Figure 3.20 Functional simulation of the synchronous 8-bit NVFA (Structure-2 and Structure-3)

The proposed 8-bit NVFA (Structure-2 and Structure-3) have the same time-dependent behaviors of outputs, shown in Figure 3.20. The only difference is the time when the data stored in NVFFs are evaluated and then transferred to the inputs of adders. 8-bit data $A_7 - A_0$

for Structure-2 are all evaluated during period (1), while those for Structure-3 are read cycle-by-cycle. The carry bit from the lowest bit propagates all the way through to the highest bit after nine cycles. Transient simulation shows the addition of two 8-bit words “00000001” and “11111111” and the expected outputs are observed as $SUM_7 - SUM_0 = \text{"00000000"}$ and $C_{out} = \text{'1'}$ at the end of the calculation.

3.3.2.3 Layout Implementation and Performance Analysis

3.3.2.3.1 Layout of the proposed 8-bit NVFA

Hybrid MTJ/CMOS process can be used for 8-bit NVFA where MTJs can be embedded above the CMOS circuits. Figure 3.21 shows the layout of a 1-bit NVHA cell, which is composed of a 1-bit NVFF, a 1-bit NVHA and a slave register. Its effective area is about 24.81μm^2. The full layout of the three proposed 8-bit NVFA circuits are then carried out. The overall sizes of the proposed 8-bit NVFAs are about 218.74 μm^2, 219.46 μm^2 and 194.96 μm^2, respectively. The layout of CMOS-only HA and FA are taken from the standard cell library of STMicroelectronics 28nm design kit.

Structure-3 becomes more advantageous in size when increasing the number of bits because more adders can share the same multi-bit NVFF and the 3-8 decoder. This can be confirmed by Figure 3.22, which shows the sizes of the three NVFA structures versus the number of addition bits. For instance, the total area of the 32-bit NVFA based on Strucutre-3 is reduced by 23.37% and 24.46% of that based on Structure-1 and Structure-2, respectively.

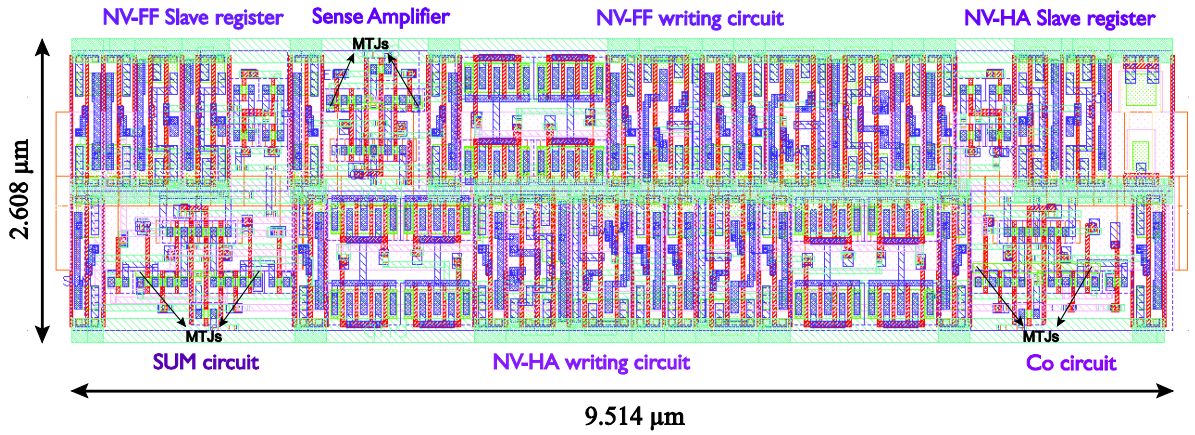


Figure 3.21 Layout of 1-bit NV-HA using CMOS 28 nm design kit

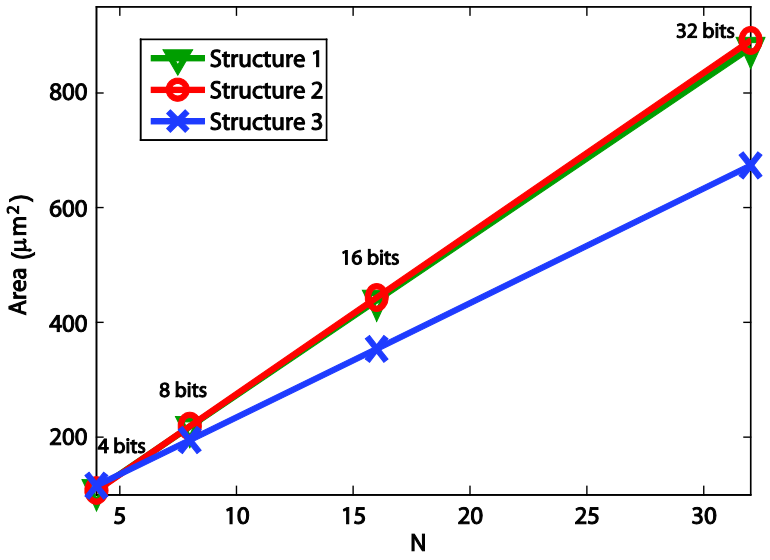


Figure 3.22 Size of the three proposed synchronous 8-bit NVFAs with respect to the number of addition bit (N)

3.3.2.3.2 Performance summary and comparison

Simulations have been carried out to understand the advantages and shortcomings of the proposed 8-bit NVFA structures. Table 3.5 summarizes the simulation results. When compared to Structure-1 and Structure-2, Structure-3 shows advantages in terms of die area because it has the less NVFFs than other structures. This advantage becomes more significant with the increase of bit since more non-volatile adders can share the same 8-bit NVFF and 3-8 decoder. In order to perform an 8-bit addition of two 8-bit words, Structure-2 and Structure-3 consume respectively 16.1% and 34.1% less dynamic energy than Structure-1. Non-volatile adders (with simple PCSA-based circuit) consume less energy during the read operation by reducing the number of current paths from V_{dd} to Gnd when compared to the CMOS-only adders.

We then compare their performances with those of the 8-bit NVFA based on domain wall (DW) racetrack memory (RM) presented in [98]. The proposed NVFAs need larger area overhead owing to the combination of NVFFs with NVFAs. However, they show advantage in terms of latency and power consumption. RM based NVFA consumes 50 times dynamic power more than the proposed NVFAs (Stucuture-1, 2 and 3) since energy needed for nucleation and propagation is too large with current technology. It also has a large delay of one operation of about 2.1 ns due to DW nucleation (～1.2 ns) and motion (～0.7 ns). Since the proposed NVFAs and previously proposed RM based NVFA are fully non-volatile, they can be powered off during the “idle” state to reduce static power consumption (or standby

energy) and power them on instantly without data loss.

Table 3.5 Comparison of different 8-bit full-adders

Parameter	Area (μm^2)	Latency (ns)	Dynamic energy (pJ/8 bits)
Structure-1[a)]	218.74	0.14	1.039
Structure-2[b)]	219.46	0.15	0.8718
Structure-3[c)]	194.96	0.18	0.6845
RM based NVFA[d)]	34	2.1	50.39

a) Structure-1 uses CMOS-only adders to perform adding operation. Input data A_7-A_0 and B_7-B_0 are stored in non-volatile states and generated by sixteen NVFFs.
b) Structure-2 integrates input data B_7-B_0 in non-volatile adders, and generates input data A_7-A_0 by eight NVFFs.
c) Structure-3 performs further area efficiency by using the 8-bit NVFF to store and sense the 8-bit input data A_7-A_0.
d) 8-bit NV-FA based on domain wall (DW) racetrack memory (RM) proposed in [98] (@65 nm).

Writing power consumption and writing delay are also two critical factors to determine the performances of integrated circuits. For the purpose of updating the stored A and B, Structure-1 and Structure-2 use the 4T writing structure. Different from these structures, Structure-3 has more complicated writing methods: 1) 4T writing structure is used for writing data B. Therefore, there is only one current path for switching a couple of MTJs, which store the 1-bit data in non-volatile state. 2) 8T writing structure is employed for writing inputs data A. The current path is separated into two to create a higher writing current and reduce the writing latency, and each has three transistors and one MTJ. Moreover, only two MTJs in the 8-bit NVFF of Structure-3 are selected during one sensing phase, thus the programing of stored inputs $A_7 - A_0$ must be bit-by-bit. On the contrary, the other two structures, Structure-1 and Structure-2, can read or write $A_7 - A_0$ by eight NVFFs at the same time.

A study of the tradeoff among the width (W) of transistors, switching speed and power dissipation have been made to find out optimal operation point (see Figure 3.23). It can be seen that both latency and power decreases quickly as W increases when $W < 1\,\mu\text{m}$ and then they slightly go down. Since the resistance of CMOS transistor in the open state is inversely proportional to its width (W), the increase of width leads to lower resistance, providing higher writing speed.

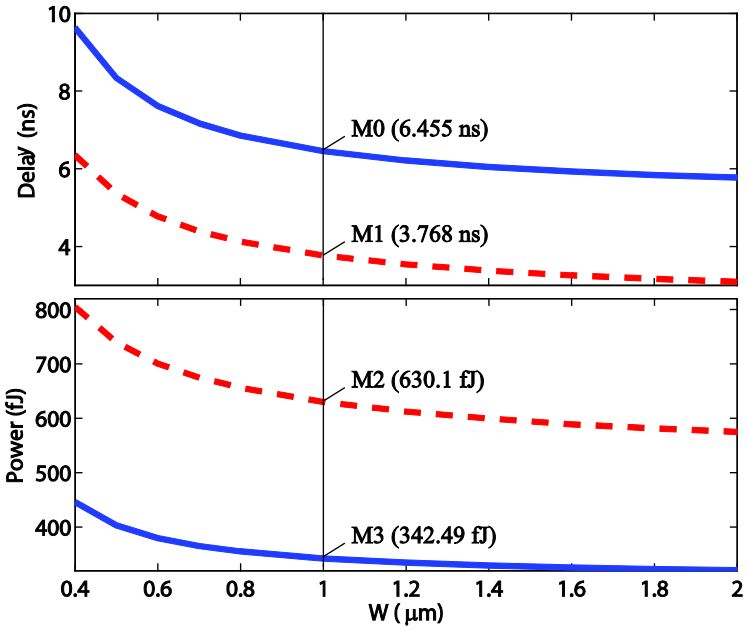


Figure 3.23 Delay and power consumption for writing a pair of MTJs. Blue solid and red dotted lines present the simulation results of 4T writing circuit and 8T writing circuit.

3.3.2.3.3 Reliability analysis

In order to evaluate the reliability of the proposed 8-bit NVFAs against the process variations, we did full MC simulations (1000 times) of all the basic addition cells. We consider the CMOS process variations and 3% MTJ process variations (TMR ratio, free layer thickness and oxide barrier thickness). As can be seen in Figure 3.14, each cell of Structure-1 is composed of a CMOS-only adder, two 1-bit NVFF and two registers. Each cell of Structure-2 can be divided into three parts: a NV adder, a 1-bit NVFF and a register. For Structure-3, a NV adder, a register and an 8-bit NVFF are used to perform each adding operation.

Figure 3.24 shows the dependence of reading bit error rate (BER) on size of transistors (W) of adders and flip-flops. BER is the error percentage when performing the read/calculation operation. It can be seen that BER can be significantly reduced by increasing the circuit area. For instance, by doubling the circuit size, BER becomes lower than 2.5% for calculating SUM (BER_SUM) and 0.3% for calculating CARRY (BER_C_o) of the 1-bit NVHA cell in Structure-2. Therefore, the proposed 8-bit NVFAs can reach the ultra-high reliability requirement at the expense of die area. From the device-level, higher TMR value can result in fewer errors. For example, for the 1-bit NVHA cell of Structure-2, BER_SUM (BER_C_o) decreases greatly from 34% (20.7%) to 24.1% (10.5%) as TMR increases from 100% to 200%.

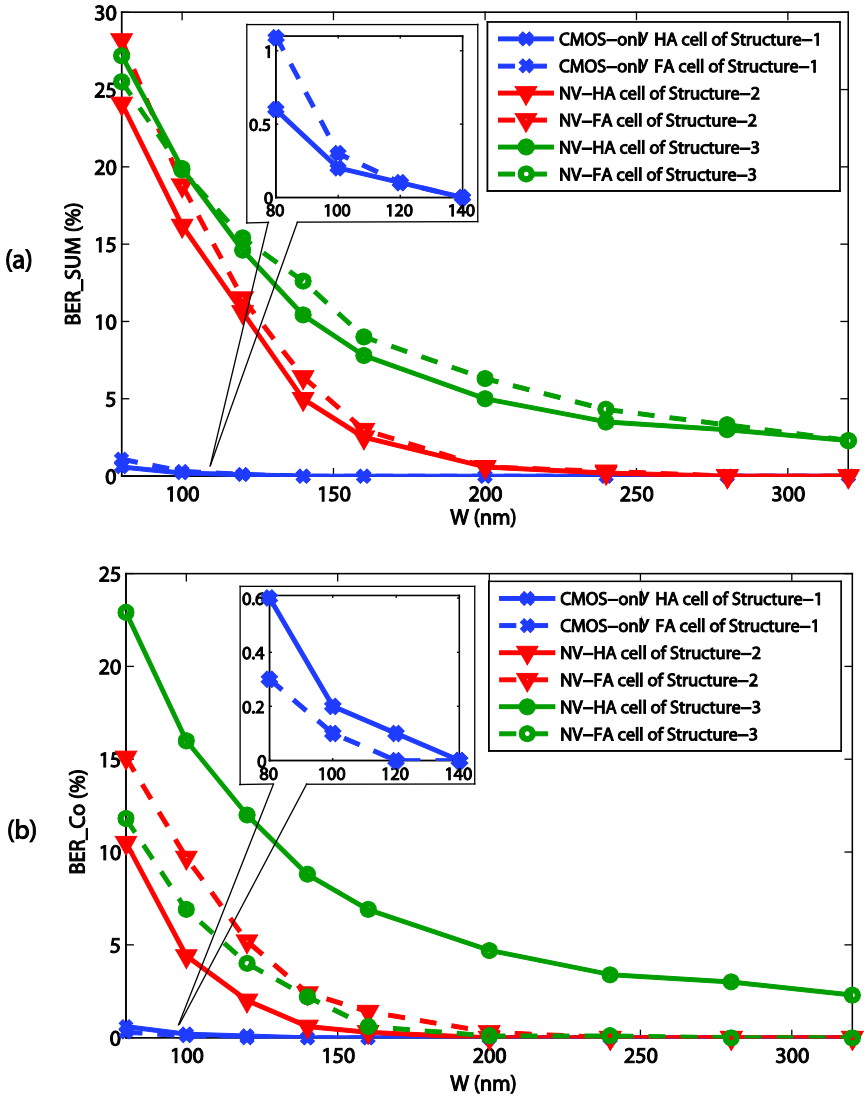


Figure 3.24 (a) Bit error rate (BER) of the SUM circuit part with respect to the width of transistors (W) in each adder cell (b) BER of the CARRY circuit part with respect to the width of transistors in each adder cell

We then investigate the effect of supply voltage (V_{dd}) on *BER* of each basic addition cells. Figure 3.25 shows the MC simulation results by varying the supply voltage from 0.7 V to 1.1 V. Simulations show that low supply voltage V_{dd} causes low sensing currents, improving energy efficiency of logic circuits at the expense of speed, which is acceptable for applications. Nevertheless, the sensing margin becomes smaller with the reduction of V_{dd}, leading to higher sensing *BER*.

As mentioned above, reliability is a key factor for the logic circuits because error correction blocks are not easy to be embedded. In order to realize full non-volatility for the LIM-based NVFA, local storage cells (i.e., NVFFs shown in Figure 3.14) are necessary. Even though we can reduce the BER by increasing the size of CMOS transistors, the overall area will be

significantly increased. Therefore, new structure/circuit that can replace the NVFF need to be investigated. Moreover, long switching delay of MTJ (~ ns compared with the sensing delay ~ ps) greatly limit the computing frequency if the non-volatile data is changed very often. Spintronics devices that can reduce the writing time will be more advantageous for high frequency logic applications. In order to solve these issues and improve the performances of the NVFA, we then study the optimization approaches from the device and circuit levels.

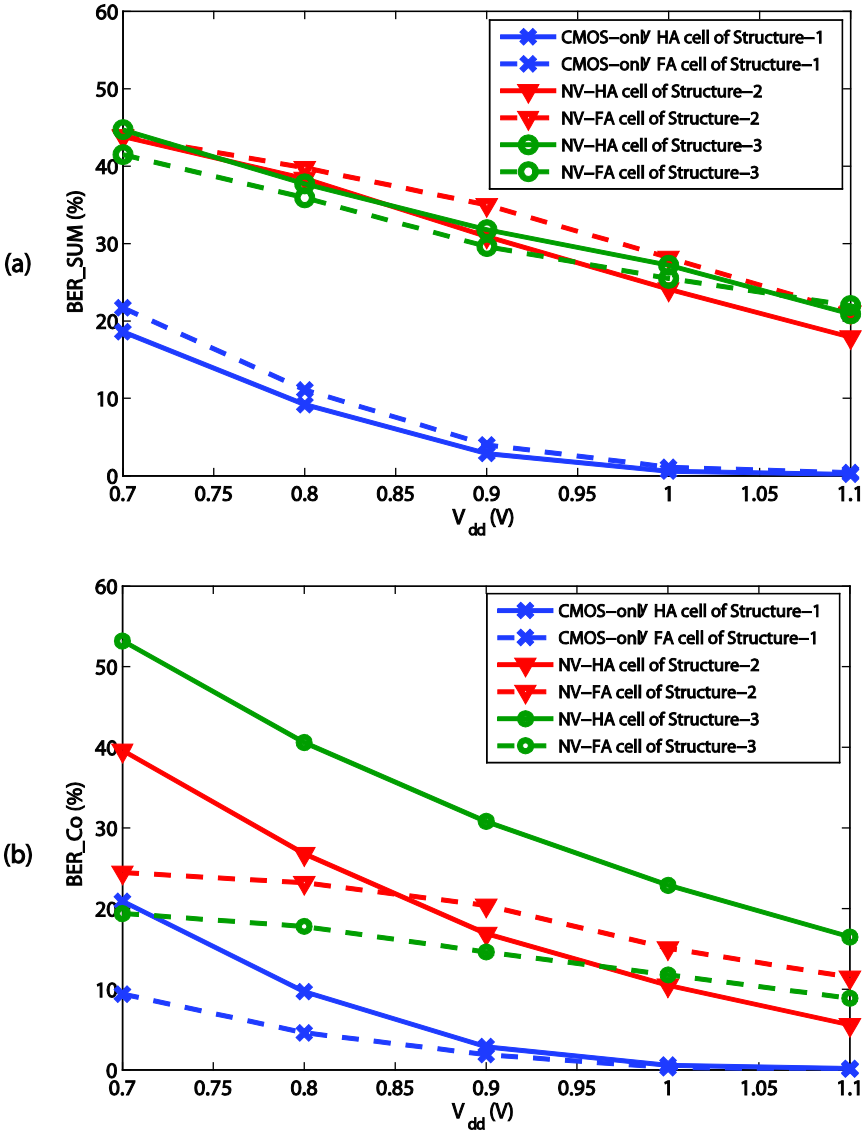


Figure 3.25 (a) BER of SUM circuit part with respect to supply voltage (V_{dd}) (b) BER of CARRY circuit part with respect to V_{dd}

3.3.3 Optimizations of NVFA

3.3.3.1 Circuit-level optimization

Several voltage-mode memory cells are proposed in [77], [145], [146] for content addressable

memory (CAM) or filed-programmable gate array (FPGA). However, these cells need complicated writing circuit. In this section, we propose a simple and reliable voltage-mode sensing circuit and integrate it into the NVFA to replace the NVFF shown in Figure 3.14.

3.3.3.1.1 Voltage-mode sensing circuit (VMSC)

As shown in Figure 3.26, the proposed 2T/2MTJ memory cell is composed of two MTJs in differential modes, one NMOS transistor and one PMOS transistor connected in series. The connected node M is joined with a CMOS latch which converts V_M to V_{dd} or Gnd. M_0 and M_1 have the same configuration except for that they are in complementary states, i.e., one MTJ has high resistance while another one has low resistance. They form a voltage divider. V_M depends on the characteristics of the series connected two MTJs.

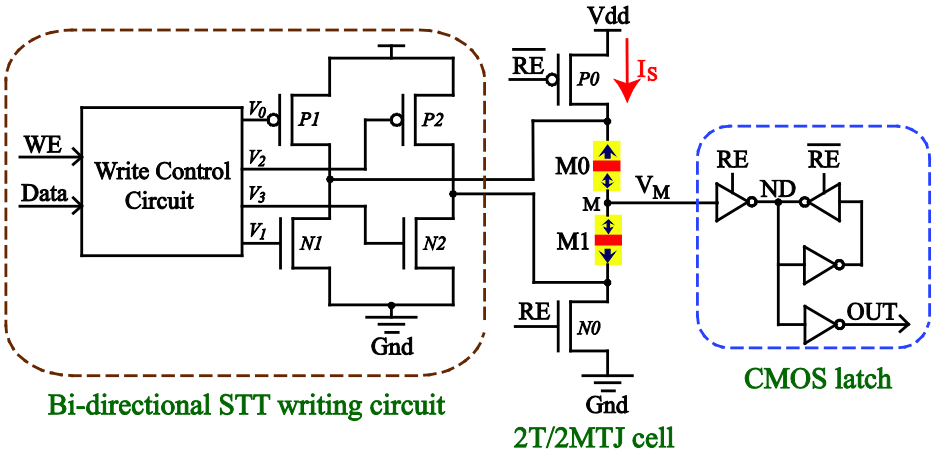


Figure 3.26 Proposed voltage-mode sensing circuit (VMSC) integrating 2T/2MTJ cell

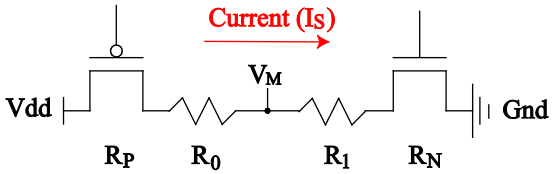


Figure 3.27 Equivalent resistance of the VMSC

To read the 1-bit storage data, a supply voltage V_{dd} is applied to the cell, generating a static reading current I_S passing through the cell (see Figure 3.27). V_M is either high when the resistance of M_0 (R_0) is less than that of M_1 (R_1), or low when R_0 is more than R_1. CMOS latch amplifies the voltage V_M at the junction M. V_M can be calculated with Eq. 3.16 and Eq. 3.17 in these two cases.

- $V_{OUT} = '1'$ when $R_0 = R_L < R_1 = R_H$:

$$V_M = V_H = \frac{V_{dd} \times (R_N + R_H)}{R_N + R_P + R_H + R_L} \quad \text{Eq. 3.16}$$

- $V_{OUT} = '0'$ when $R_0 = R_H > R_1 = R_L$:

$$V_M = V_L = \frac{V_{dd} \times (R_N + R_L)}{R_N + R_P + R_H + R_L} \quad \text{Eq. 3.17}$$

where V_{OUT} is the output voltage at node OUT, R_N and R_P are respectively the ON resistances of NMOS transistor and PMOS transistor, R_L and R_H are respectively the resistances of MTJ in parallel and anti-parallel states.

To write data into the MTJs, 4T writing circuit with two NMOS (N_1 and N_2) and two PMOS (P_1 and P_2) transistors is employed. In order to write data '0' into the corresponding MTJs, write enable signal $WE = '1'$ and $Data = '0'$. P_1 and N_2 are turned ON while P_2 and N_1 are turned OFF, forming a current loop. To write logic '1', a reversed writing current is generated by setting $Data = '1'$ and thus transistors P_2 and N_1 will be open.

This architecture realizes a simple read/write system. Note that the transistors, i.e., N_0 and P_0, play different roles during the read ans write modes. When reading data, the additional resistances R_N and R_P reduce the reading current, thus alleviating the unintentional write issues. When performing write operation, RE is kept to be low to close N_0 and P_0, separating the MTJs from V_{dd} and Gnd.

Timing diagram of the voltage-mode sensing circuit is shown in Figure 3.28. The width of P_0 (W_P_0) is 200 nm while that of N_0 is in minimum size (80nm). M_0 and M_1 in Figure 3.26 are initialized at anti-parallel state and parallel state, respectively. When $RE = '1'$, it can read from the figure that V_M (385.81 mV) is smaller than the threshold voltage of the latch and $OUT = '0'$. WE is then set to 1 V, switching the configurations of the two MTJs. During the second read period, V_M (592.04 mV) becomes larger than the threshold voltage of the latch. After amplification with the latch, we can obtain an output signal of 1 V

corresponding to the input data (*Data*).

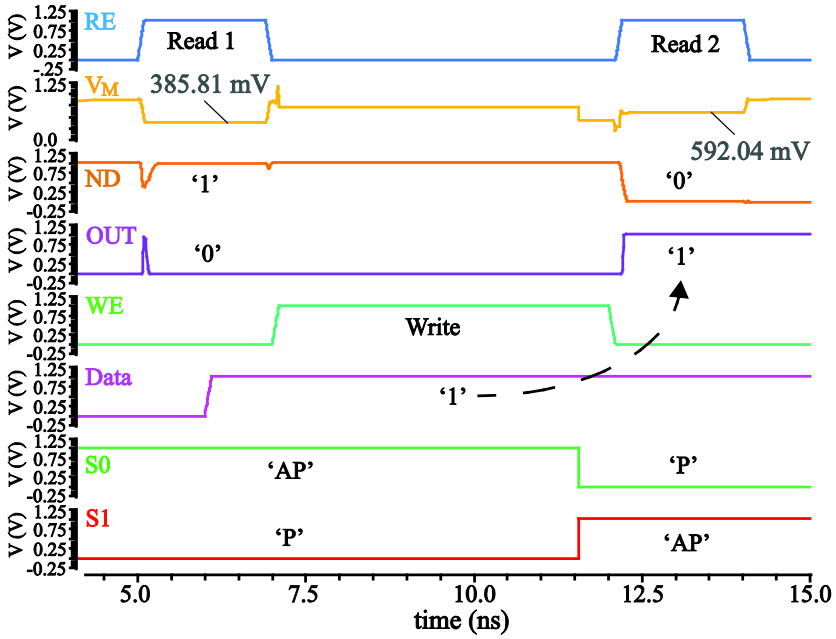


Figure 3.28 Simulation of the VMSC. S_0 and S_1 represent the state of M_0 and M_1, respectively. Data is sensed if RE='1' or written if WE='1'.

Figure 3.29 illustrates the structure of the proposed NVFA using 2T/2MTJ cells (2T/2MTJ-NVFA). Different from the NVFA presented in Figure 3.14, 2T/2MTJ-NVFA has two voltage-mode sensing circuits (instead of MFFs) to store and generate input data A and C_i (and their complements $\overline{A}$ and $\overline{C_i}$). The carry-in input C_i can also be directly connected with the carry-out output of other arithmetic unit to form more complex functions.

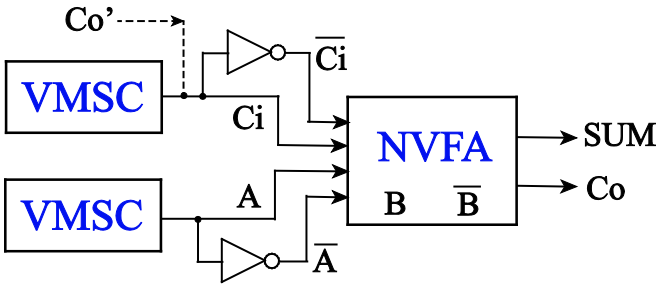


Figure 3.29 Full schematic of fully non-volatile NVFA using VMSCs

3.3.3.1.2 Performance analysis

We first analyze the sensing of 2T/2MTJ cell. Figure 3.30 demonstrates the influence of width of PMOS transistor P_0 (W_P_0) on the static sensing current (I_S) and the sensing margin (ΔV_M). The supply voltage here is 1 V. It shows that bigger W_P_0 leads to larger ΔV_M,

which is advantageous for reliable sensing. However, the resistance of transistor becomes smaller and hence I_S gets closer to the critical writing current of MTJ (~50 μA). For instance, $I_S > 40\,\mu A$ when W exceeds 200 nm. An unintentional writing may occur during read operation due to the increasing process variations at ultra-deep submicron node. Besides, ΔV_M increases with higher TMR ratio because it is easier to distinguish the states of serially connected MTJs.

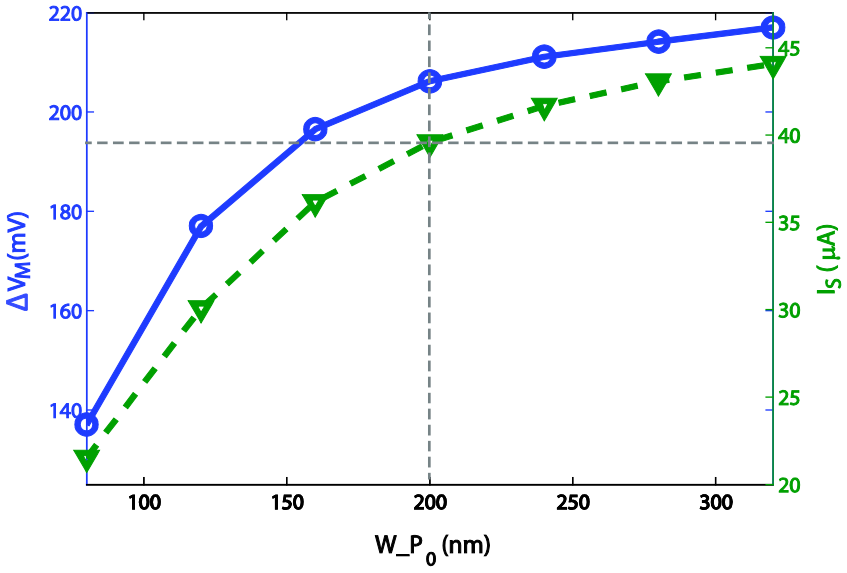


Figure 3.30 Sensing margin and sensing current of the 2T/2MTJ cell versus the width of P_0

Table 3.6 presents simulated results of the 2T/2MTJ-NVFA when ABC_i are initialized to "101". Simulations have been conducted under supply voltages (V_{dd}) varying from 1 V to 0.75V. It can be seen that lower V_{dd} leads to larger sensing latency. Both static sensing currents for reading input A or B and total dynamic current for performing adding operation are smaller with the decrease of V_{dd}, thus less energy is required.

MC simulations show that the BER (error percentage for reading the data stored in the MTJs) of the proposed VMSC is nearly zero, which can hardly be reached by the PCSA-based MFF without much area overhead. From Table 3.6, we can find that the 2T/2MTJ-NVFA becomes less reliable when the supply voltage decreases from 1 V to 0.75 V. This can be explained as: 1) Lower V_{dd} results in smaller ΔV_M; 2) The dynamic sensing currents for calculating SUM and C_o becomes smaller when V_{dd} is lower, leading to smaller current difference between two branches of both SUM sub-circuit and CARRY sub-circuit. When V_{dd} is lower

than 0.7 V, *BER* is simulated to be bigger than 40% for SUM sub-circuit and 30% for CARRY sub-circuit. By tripling the circuit area, the 2T/2MTJ-NVFA can reach a low BER smaller than 1%.

Table 3.6 Simulation results of the 2T/2MTJ-NVFA with V_{dd} varying from 1 V to 0.75 V

V_{dd} / V	**1**	**0.95**	**0.9**	**0.85**	**0.8**	**0.75**
T_*SUM* (ps)	224.1	238.3	254.7	277.1	304.9	379.2
T_C_o (ps)	130.6	139	149.6	162	181.6	209.3
Static current I_S (μA)[a]	39.58	35.85	32.02	28.07	23.99	19.68
Operation energy (fJ/bit)	26.13	22.93	20.13	15.49	15.43	14.79
BER _ *SUM*	4.9%	6.7%	9.1%	13%	17.4%	29.3%
BER _ C_o	11.1%	12.6%	15.3%	18.7%	22.4%	25.6%

[a] Static sensing current for reading 2T/2MTJ cell

3.3.3.1.3 Optimized VMSC

The static sensing approach of the proposed VMSC has a constant current passing through the MTJs during the reading operation, resulting in high sensing energy consumption. In order to solve the high power issue, we propose the optimized circuit with a self-enable control circuit (see Figure 3.31). During the reading operation, once outputs *SUM* (or C_o) and $\overline{SUM}$ (or $\overline{C_o}$) are different, the transistors in the 2T/2MTJ cell will be closed, and then the sensing operation is disabled.

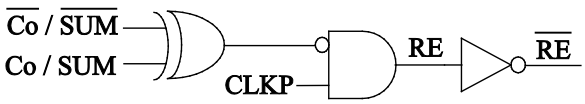


Figure 3.31 Self-enable control circuit for the optimized VMSC

The simulation result of the NVFA with the optimized VMSC is illustrated in Figure 3.32. It can be seen that once the *SUM* is different from $\overline{SUM}$ (detected at the point T_0), the static current is cut-off without disturbing the outputs *SUM* and $\overline{SUM}$. In this way, this circuit can greatly save the energy for low-power computing system.

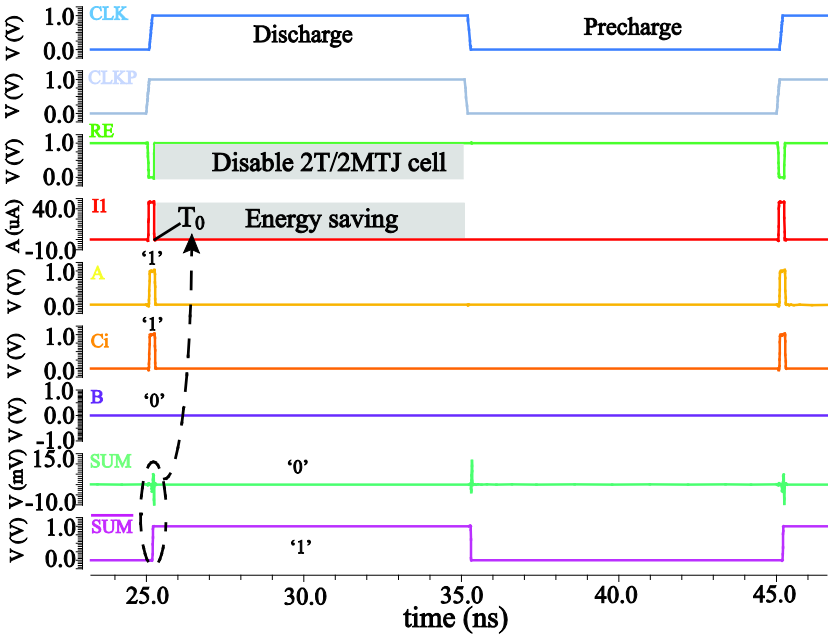


Figure 3.32 Simulation of the NVFA using the optimized voltage-mode sense amplifier

3.3.3.2 Device-level optimization

Previously proposed NVFAs are mainly based on MTJs switched by spin transfer torque (STT). Even though they show advantages in reading speed and reading energy, they suffer from low writing speed and high writing power dissipation because STT switching requires a large incubation delay at the initial process. In order to achieve high-speed operation, transistors of the writing circuit should be enlarged, resulting in not only large area overhead but also high risk of MTJ barrier breakdown. Another solution is to reduce the critical write current I_{C0}, which, however, decreases the thermal stability barrier.

Recently, spin-Hall effect (SHE) and Rashba effect were proposed to solve this issue [63], [64], [147], [148]. Among them, spin-Hall-assisted STT switching was proposed to achieve high-speed write operation in the perpendicular-anisotropy MTJ [66], [148]. As can be seen in Figure 3.33(a), an MTJ is fabricated at the top of a heavy metal strip (β-W) with its free layer in contact to the metal strip. Two currents, STT writing current I_{STT} and SHE writing current I_{SHE}, are combined to switch the magnetization of free layer. I_{STT} is responsible for generating the conventional STT, while I_{SHE} can inject a spin current into the free layer due to the SHE in heavy metal [60]. The injected spin current exerts a so-called spin Hall torque which can assist the STT to ease the switching of MTJ. Therefore, the STT writing current can be limited to a relative small value while keeping high write speed. Moreover, the writing

voltage of MTJ can be greatly reduced to improve its endurance.

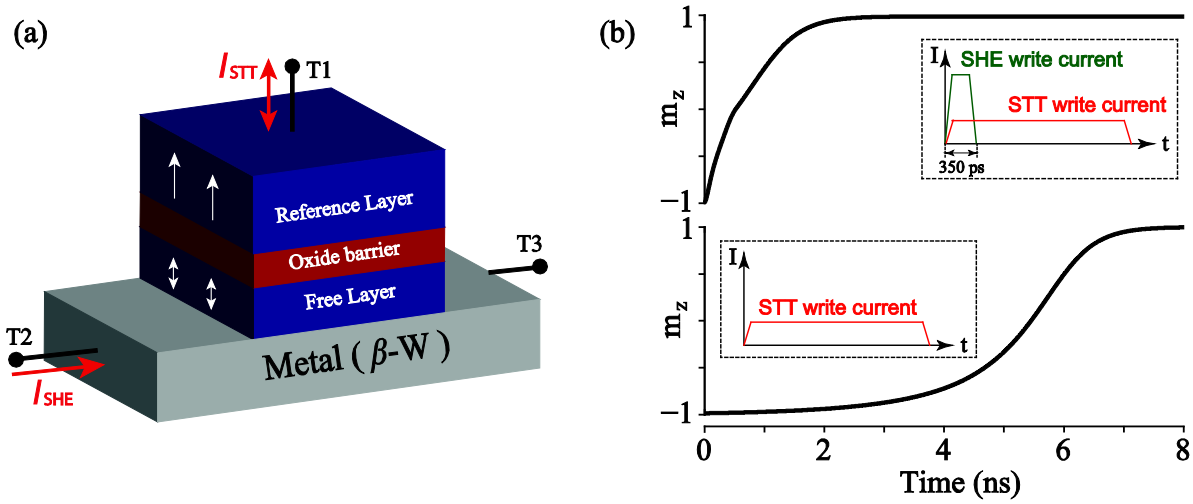


Figure 3.33 (a) Three-terminal MTJ device structure (b) Time evolution of perpendicular-component magnetization (m_z) driven by the combination of STT and SHE writing currents (upper), and the single STT writing current (lower)

3.3.3.2.1 Spin-Hall-assisted STT MTJ model

For the spin-Hall-assisted STT switching, the magnetization dynamics of free layer of MTJ is described by a modified Landau-Lifshitz-Gilbert (LLG) equation [148], as

$$\frac{\partial \mathbf{m}}{\partial t} = -\gamma\mu_0 \mathbf{m}\times\mathbf{H}_{\text{eff}} + \alpha\mathbf{m}\times\frac{\partial \mathbf{m}}{\partial t} - \xi P J_{\text{STT}}\mathbf{m}\times(\mathbf{m}\times\mathbf{m}_{\text{r}}) - \xi\eta J_{\text{SHE}}\mathbf{m}\times(\mathbf{m}\times\boldsymbol{\sigma}_{\text{SHE}}) \qquad \text{Eq. 3.18}$$

where $\mathbf{m}$ and $\mathbf{m}_r$ are unit vectors along the magnetization orientation of the free layer and reference layer, respectively. J_{STT} and J_{SHE} are STT and SHE write current densities, respectively. SHE is the polarization direction of spin current induced by SHE. $\mathbf{H}_{\text{eff}}$ is the effective field. More details about other coefficients can be found in [148].

Numerical simulation based on Eq. 3.18 indicates that two requirements are mandatory for a fast spin-Hall-assisted STT switching [148]: first, J_{SHE} must be large enough to produce sufficient spin-Hall torque to eliminate the incubation delay of the conventional STT; second, J_{SHE} must be removed at an appropriate time in order that STT continues to achieve deterministic switching. Figure 3.33(b) shows a comparison of the time evolution of m_z (perpendicular-component magnetization in the free layer of MTJ) between spin-Hall-assisted STT switching and the conventional STT switching. It can be seen that the former achieves

faster magnetization switching than the latter thanks to the elimination of incubation delay.

A spice-compatible model for the proposed spin-Hall-assisted MTJ has been developed in [148]. It integrates Brinkman model [106], Slonczewski model [149], and aforementioned LLG equation to describe the tunneling resistance and magnetization switching, respectively. It is programmed in Verilog-A language and provides a feasible interface between MTJ signals and CMOS circuits (see Appendix D). Table 3.7 shows the critical parameters used in the following simulations.

Table 3.7 Parameters of the spin-Hal-assisted STT MTJ model used in fitting functions

Description	**Default value**
Oxide barrier thickness	0.85 nm
Free layer thickness	0.7 nm
MTJ surface	40 nm × 40 nm
Heavy metal volume	50 nm × 40 nm × 3 nm
Resistance-area product	10 Ω·μm^2
TMR ratio with V_{bias}=0	150%
MTJ thermal stability factor	30
Spin Hall angle	0.3
MTJ resistances	~ 6 kΩ, ~ 15 kΩ
Heavy metal resistance	~ 833 Ω

3.3.3.2.2 NVFA based on MTJ with spin-Hall assistance

The NVFA based on MTJ switched by spin-Hall-assisted STT (STT+SHE NVFA) is illustrated in Figure 3.34. The reading circuit (Part 1 in Figure 3.34) of the STT-SHE NVFA is the same as the STT-based NVFA. But it has a more complex writing circuit, which is composed of STT PMOS transistor (P_1 or P_3), STT NMOS transistor (N_1 or N_3), SHE PMOS transistor (P_0 or P_2) and SHE NMOS transistor (N_0 or N_2). STT and SHE transistors are used to generate STT and SHE writing currents, respectively. V_{STT} and V_{SHE} control the direction of SHE and STT write currents. N_0 and N_2 are connected onto terminals T_3 while P_0 and P_2 are connected onto the terminals T_2.

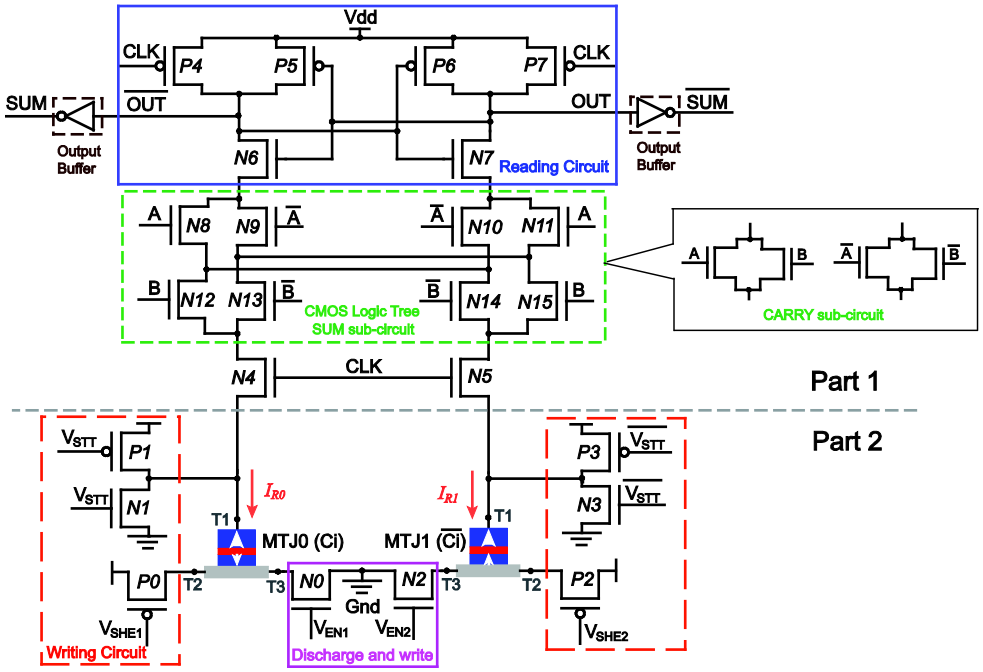


Figure 3.34 Schematic of the STT+SHE NVFA

To write data, both SHE and STT writing currents are firstly applied. SHE writing current should be removed after a short duration and finally the deterministic switching is achieved by STT [148]. Figure 3.35 presents the equivalent resistor networks as well as current directions during the two write phases (i.e., before and after SHE writing current is removed). Initially, we assume that MTJ_0 is in anti-parallel configuration and MTJ_1 is in parallel configuration. Their corresponding resistances are denoted as R_{MTJ0} and R_{MTJ1}. $R_{N0} - R_{N3}$ and $R_{P0} - R_{P3}$ are the resistances of the NMOS transistors $N_0 - N_3$ and PMOS transistors $P_0 - P_3$.

- During the first writing phase, P_1 and N_3 are closed while other transistors are open. For both MTJs, SHE writing current flows from terminal T_2 to T_3. STT writing current of MTJ_0 flows from bottom (free layer) to top (conference layer), while that of MTJ_1 flows from the top to the bottom (see Figure 3.35(a)). As can be seen in Table 3.7, the resistances of MTJs are much larger than that of metal strip. Therefore, I_{STT} should be smaller than I_{SHE} when all the transistors are in minimum size. For MTJ_0, by increasing the width (W) of STT NMOS transistor N_1 or decreasing that of SHE NMOS transistor N_0, I_{STT} will increase while I_{SHE} will decrease. For MTJ_1, bigger W of STT PMOS transistor P_3 and smaller W of SHE PMOS transistor P_2 result in larger I_{STT} and smaller I_{SHE}.

- During the second phase (SHE writing current is removed), only $N_1 - N_2$, P_0 and P_3 are open. Writing current flows from T_2 to T_1 for MTJ_0 and from T_1 to T_3 for MTJ_1, as shown in Figure 3.35(b).

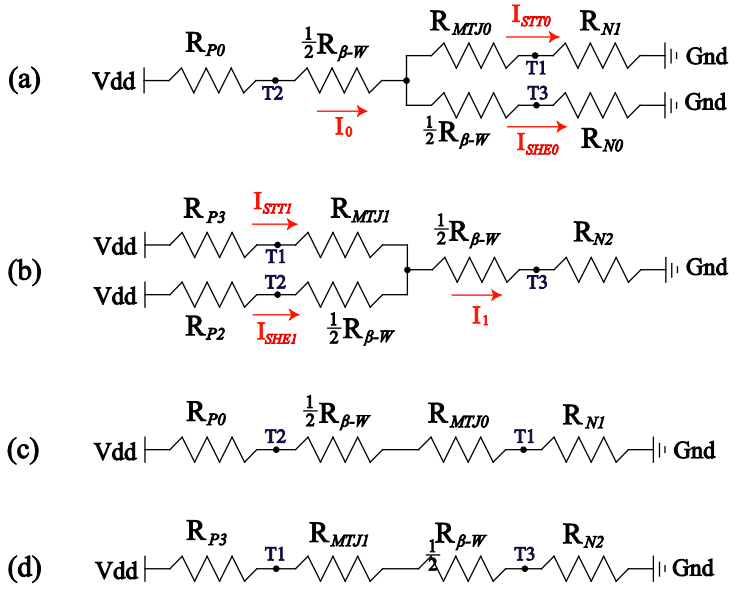


Figure 3.35 Equivalent resistor networks and write current directions (a) for switching MTJ_0 before I_{SHE0} is removed (b) for switching MTJ_1 before I_{SHE1} is removed (c) for switching MTJ_0 after I_{SHE0} is removed (d) for switching MTJ_1 after I_{SHE1} is removed.

3.3.3.2.3 Simulation and discussion

Figure 3.36 confirms the functionality of the STT+SHE NVFA at CMOS 28 nm technology node including reading and programing operations. m_{z_0} or m_{z_1} represents the perpendicular component magnetization in the free layer of MTJ_0 or MTJ_1. MTJ is in anti-parallel state if $m_z = -1$; or in parallel state if or $m_z = 1$. Data '1' is stored when MTJ_0 is in parallel state and MTJ_1 is in anti-parallel state. Otherwise, input data C_i is '0'. MTJs are programmed after four periods, switching the non-volatile input data C_i from logic '0' to logic '1'. All the input patterns ABC_i are applied to the SUM and CARRY sub-circuits. It is confirmed that the expected outputs, i.e., SUM and CARRY, are observed as "00", "10", "10", "01", "10", "01", "01" and "11", respectively.

Figure 3.37 presents the simulations curves of one programing operation, which can divided into four parts. In part (1), all the transistors of the writing circuit are closed and MTJ is initialized in parallel configuration. No I_{SHE} or I_{STT} passes through the MTJs. In part (2),

I_{SHE} and I_{STT} are generated and I_{SHE} plays a dominant role during this period. After a short duration of 350 ps, SHE current is removed and STT continues to complete the switching during part (3). In part (4), z-component magnetization (m_z) switched to '1' when the state of MTJ is changed from anti-parallel to parallel (AP→P); or '0' when the state of MTJ is changed from parallel to anti-parallel (P→AP), and then stays stable.

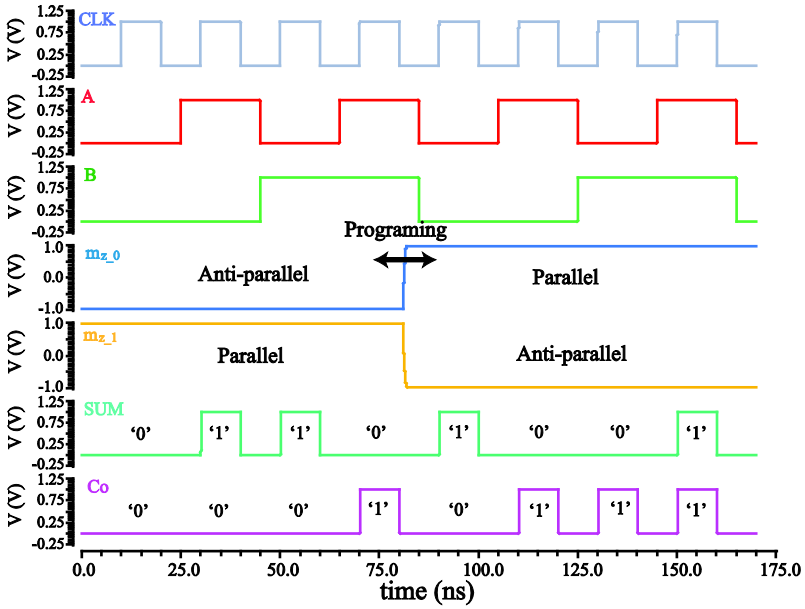


Figure 3.36 Simulation of the STT+SHE NVFA

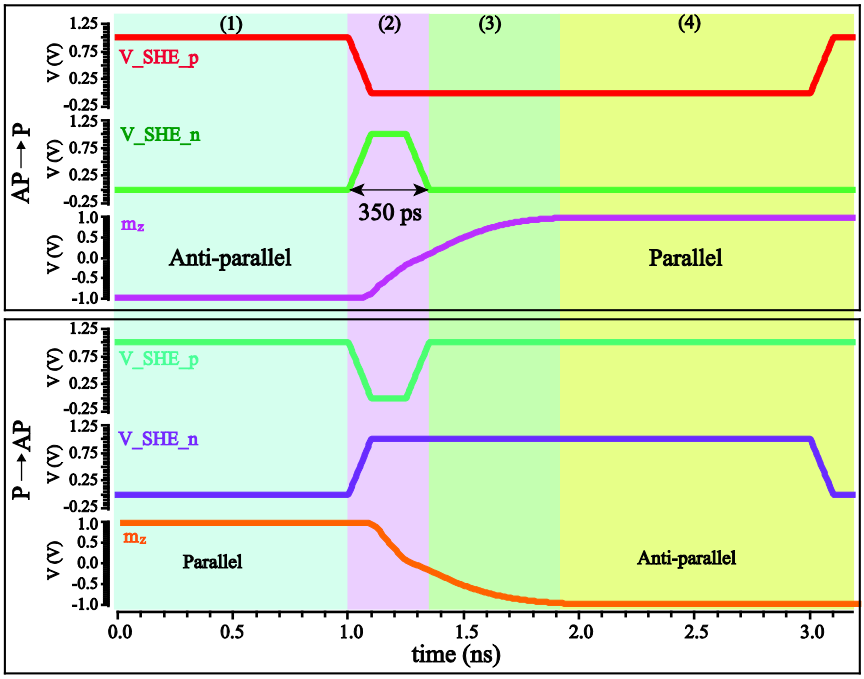


Figure 3.37 Simulation of MTJ switching. m_z=1 represents that the relative magnetization orientations of two ferromagnetic layers are parallel, while m_z=0 represents that they are anti-parallel.

Simulation results show that both cases operate at ultra-high speed and low power consumption (0.948 ns and 77 fJ the case for AP→P, 0.981 ns and 83 fJ for the case P→AP). The differences in writing time and energy dissipation originate from the different resistances of MTJs in complementary states. Therefore, a 1ns STT pulse assisted by a 0.35 ns SHE pulse is expected to switch the magnetization of free layer of an MTJ. The width (W) of writing circuit transistors ($N_0 - N_3$ and $P_0 - P_3$ in Figure 3.34) are fixed at 4X, while other transistors are kept in minimum size (X = 80 nm).

We compare the STT+SHE NVFA with the conventional NVFA switched by STT (STT NVFA). In order to achieve the same write time as the proposed NVFA (<1 ns), STT NVFA need to increase the size (W) of write circuit transistors to 20X (～1.6 μm), resulting in large area overhead. In this case, voltage applied to MTJ becomes larger than 500 mV, which can easily damage the MTJ barrier.

Table 3.8 shows the performance comparison of STT+SHE NVFA and STT NVFA. The size (W) of writing circuit transistors in STT NVFA are set at 300 nm in order to keep the same circuit area as the STT+SHE NVFA. Simulation results show that STT+SHE NVFA has advantages in delay and energy when keeping the same circuit size. For performing an adding operation including write and read, the proposed NVFA needs 38% less operation time (read time + write time) and 30.8% less energy than STT NVFA.

Table 3.8 Comparison of STT+SHE NVFA with STT NVFA

		STT+SHE NVFA	**STT NVFA**
Read time (ps)		137.7	151
Read energy (fJ)		1.23	1.33
Write time (ns/bit)	*AP→P*	0.948	1.654
	P→AP	0.981	1.62
Write energy (fJ/bit)	*AP→P*	77	116.4
	P→AP	83	115.2

3.4 Conclusion

In this chapter, we proposed logic/arithmetic circuits based on the LIM architecture. Storage and logic functions are merged into MTJs, which largely reduces the transfer energy and delay. The basic NVLGs were first proposed by integrating the MTJs into the current-mode sense amplifier, i.e., PCSA. We then presented a novel design of NVFA architecture. The effect of discharge transistor size, MTJ resistance-area product (RA) and TMR ratio have been respectively studied. It was compared with the conventional CMOS-only FA, confirming its performance advantages of die area and power consumption.

In order to extend the single-bit NVFA to multi-bit case and realize full non-volatility to promise nearly zero standby power and instant ON/OFF, three possible synchronous 8-bit NVFA structures were proposed according to the location of non-volatile data and system requirements. All their input data are stored in non-volatile state. Even though the first structure of 8-bit NVFA (Structure-1) performs high reading frequency and small area, it consumes high reading energy. The second structure (Structure-2) addresses this problem by replacing the CMOS-only adders by non-volatile ones. This configuration reduces eight NVFFs. Structure-3 further reduces the power consumption as well as area by storing by using the 8-bit NVFF. One major shortcoming of this structure is that eight cycles are needed to read or write the 8-bit data.

After that, we improved the performances (reliability and writing power dissipation) of the NVFA. From the circuit level, a novel voltage-mode sensing circuit is investigated for reliable reading against the process variations. Non-volatile data is stored in a 2T/2MTJ memory cell, which can be read with low reading BER smaller than 1%. From the device level, NVFA integrating MTJ switched by spin-Hall-assistance STT was proposed. In such configuration, STT writing is assisted by a current passing through the heavy metal below the MTJ due to SHE. The STT+SHE NVFA can achieve ultra-fast switching (<1 ns) and low energy (<100 fJ). The endurance of oxide barrier is largely enhanced as the requirement of lower write voltage. When keeping the same area, STT+SHE NVFA saves 38% operation delay and 30.8% energy dissipation to perform an addition including writing and reading operations. It shows great potential in high-frequency and low-energy applications.

Chapter 4 Non-volatile content addressable memory (NVCAM)

Content addressable memory (CAM) is a computer memory that is widely used in many applications such as network routers, processors, etc. It compares the search data with a table of storage data and then output the match location. The mainstream SRAM-based CAM is presented in Figure 4.1, which consists of m words, and each word has n bits. There is a match line corresponding to each word ML_{0-m} connected with a sense amplifier. A pair of search lines (e.g., SL_0 and $\overline{SL_0}$) correspond to 1-bit search data. There are mainly two types of storage cell (e.g., NOR type and NAND type), where the storage data is remained in two cross-coupled inverters. During the search operation, the search word is loaded to the search data drivers and then onto the search lines. The match line will be discharged to the ground if there exists one or more "Mismatch" or remained at high level if all bits match the search data.

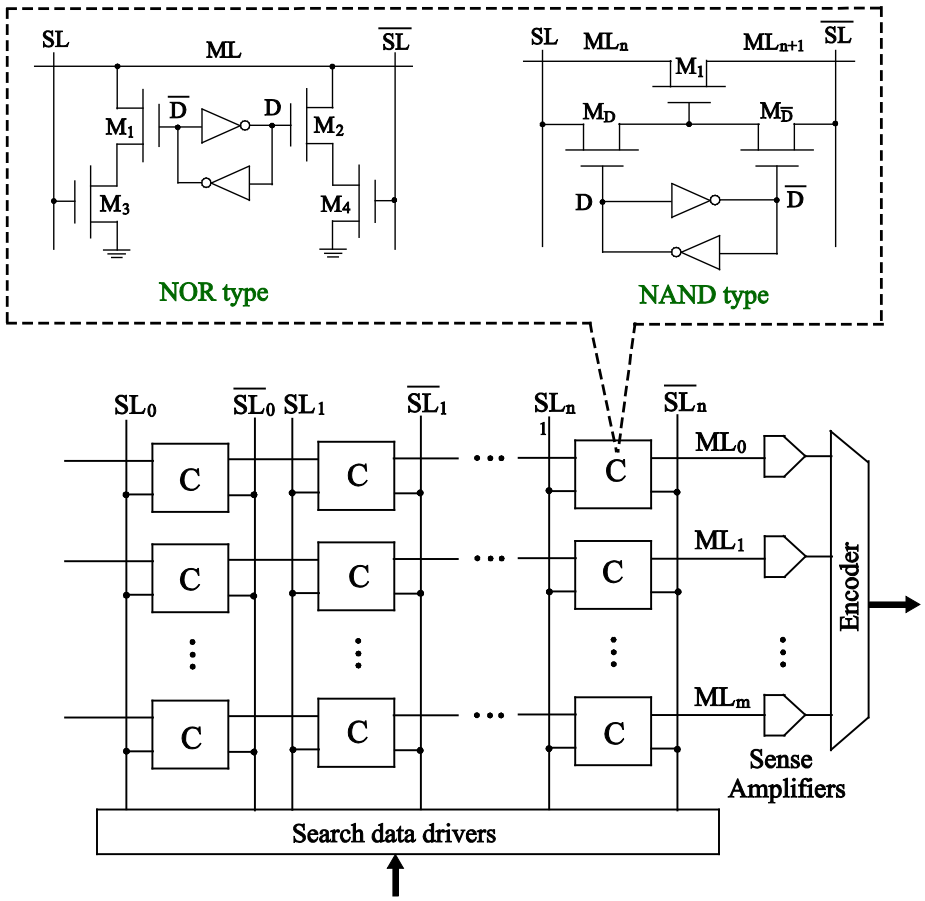


Figure 4.1 Conventional content addressable memory (CAM) and two types of core cells (NOR type and NAND type) [150]

Many research teams propose techniques to reduce the dynamic energy of CMOS-based CAM [151], [152], [153]. However, they still suffer from high standby power issue due to the leakage current, especially as technology node shrinks below 45 nm. Moreover, CMOS-based

CAM cannot be of high density because logic operation and data storage are performed in separate circuits [154]. Non-volatile CAM (NVCAM) based on spintronics devices such as MTJ is an efficient solution for the above problems.

In this chapter, we propose a NVCAM as one of the applications of LIM architecture. In this NVCAM, multiple MTJ cells used for storage and logic function share the same comparison circuit to provide area efficiency. Two types of magnetic decoders (MDs) are designed for word line selection. By using an industrial 28 nm FDSOI CMOS design kit and the PMA STT-MTJ compact model, we validate the functionality of the NVCAM and evaluate its performance merits.

4.1 Structure of NVCAM

Figure 4.2 illustrates the structure of four-word-width-four-word-depth NVCAM. The match line (ML) is pre-charged through the pre-charge PMOS transistor (T_p) when the signal PRE is activated. When a word (e.g., "0100") is searched, PRE is set high and the first word ($Word_0$) will be loaded. With the response of "Mismatch", the next word ($Word_1$) will be addressed and so on until a "Match" is detected.

The CMOS-based comparison circuit and writing circuit are shared by storage cells in the same column for area efficiency (see Figure 4.3). The CAM cell has five parts: PCSA for detecting the magnetization of MTJs and output the comparison result, 4-bit non-volatile memory part (M_{0-7}), a writing circuit to change the state of MTJ, a CMOS logic tree (N_{5-8}) for building up a XOR logic network along with the MTJs (presented in Section 3.2.3) and a pass transistor (N_0) to determine the critical path between ML and the ground. A couple of complementary MTJs (e.g., M_{0-1}) are used to present binary data and loaded by switch transistors (e.g., N_{11-12}). S_{0-3} are signals used for controlling the ON/OFF state of the switch transistors N_{11-18}.

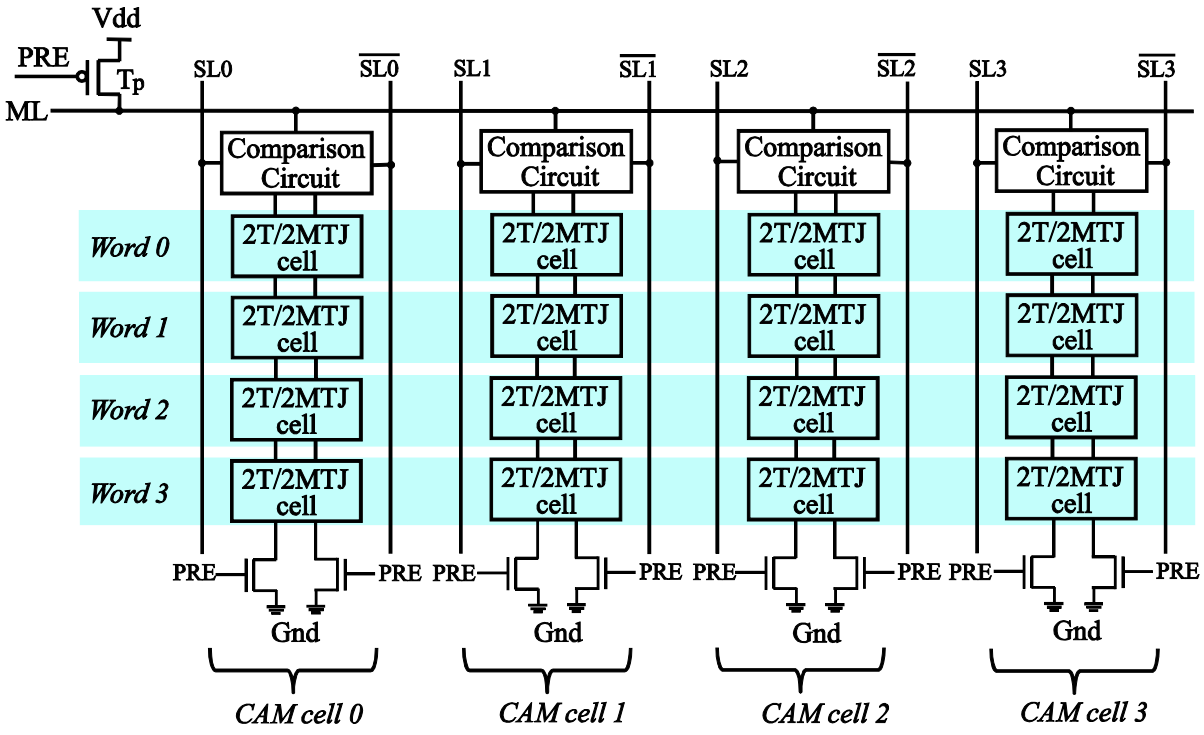


Figure 4.2 Structure of the proposed non-volatile content addressable memory (NVCAM) with 4×4 array

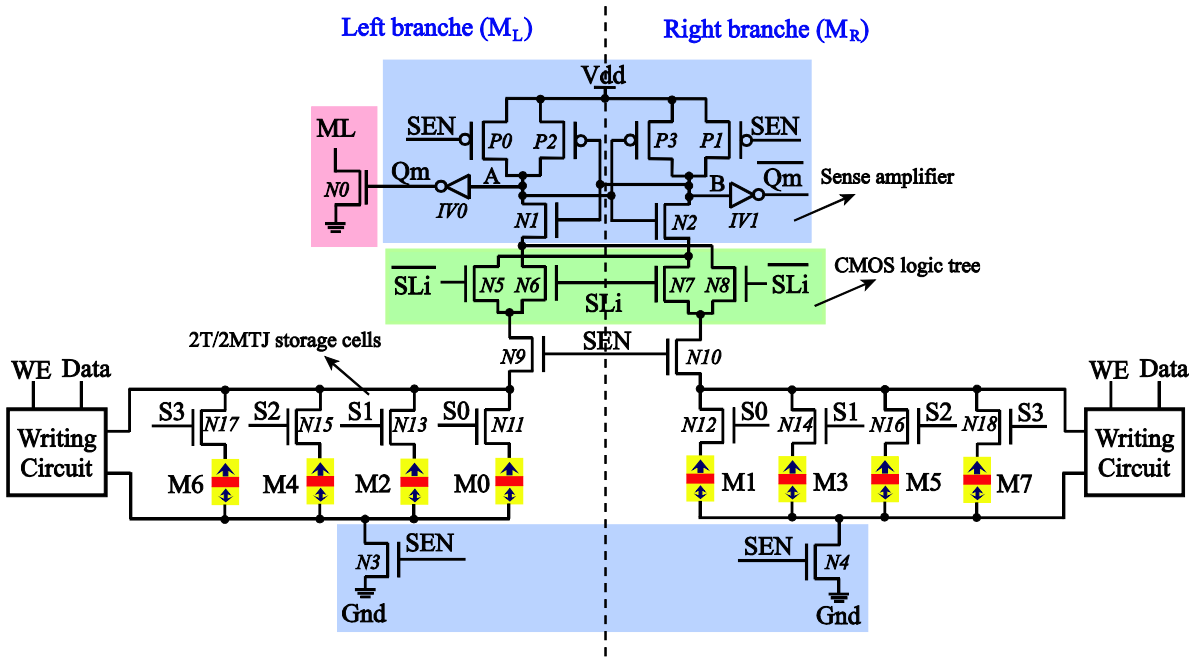


Figure 4.3 Schematic of the basic CAM cell. SL_i represents the search line, where i is the number of word line.

Search operation (or comparison operation) is performed by comparing the search data on the search lines (SL) with the data stored in MTJ cells. The NVCAM has two phases:

- **Pre-charge phase**: Signals SEN and PRE are at low voltage, pre-charging match line (ML), nodes A and B to V_{dd}. Node Qm (and its complement $\overline{Qm}$) is then pulled down to the ground through the output inverter IV_0 (and IV_1), closing the pass transistor N_0. Thus, there is no path between ML and the ground. Discharge transistors (N_3 and N_4) are turned OFF and the comparison operation is disabled.

- **Comparison phase**: SEN and PRE are turned high, closing the pre-charge transistors P_0, P_1 and T_p. Nodes A and B begin to discharge at different speeds. According to the resistance difference between two branches, one output (e.g., Qm) will be pulled up to V_{dd}, while the other one (e.g., $\overline{Qm}$) will continue to discharge to 0 V. When the stored data equals to the search data, Qm will be 0 V and close the pass transistor N_0. ML will hold the charge when all the bits in a word match the search lines $SL_3 - SL_0$. Otherwise, ML will be discharged to the ground, denoting a mismatch. The corresponding truth table (see Table 4.1) summarizes the relationship among the stored data, search data and match result.

Table 4.1 Operation mechanism of the CAM cell

Stored Data		**Search Data**	**Qm**	**N_0**	**Match Result**
(M_L, M_R)	NV data				
(P, AP)	0	0	Gnd	Closed	*Match*
		1	V_{dd}	Open	*Mismatch*
(AP, P)	1	0	V_{dd}	Open	*Mismatch*
		1	Gnd	Closed	*Match*

M_L and M_R are MTJ placed in the left branch and right branch, respectively. SD represents the corresponding stored data.

For writing a pair of MTJs, transistors N_9 and N_{10} separate the memory part from the PCSA circuit and the CMOS logic tree to minimize their influence on the writing current. WE is the activation signal and $Data$ controls the direction of the writing current. In the case of switching the MTJs in the left branches (M_L) to parallel state (P) and the MTJs in the right branches (M_R) to anti-parallel state (AP), data is set to ‘0’ and vice versa.

4.2 Simulation and performance analysis of NVCAM

Functional simulation of CAM cell integrating four contexts (see Figure 4.3) is carried out by using the Cadence Spectre simulator (@28 nm technology node). Figure 4.4 shows the timing diagram of reading and writing operations. The first context stored in $M_0 - M_1$ (M_0 is at P state and M_1 is at AP state to store data ‘0’) has been loaded with S_0 turns high while the other switch signals stay low.

- During the first read phase “Read 1”, the output value is ‘0’ ($Qm = '0'$) since the stored data and the search data on search line (SL) are both ‘0’.

- During the second read phase “Read 2”, Qm turns ‘1’ because the stored data (‘0’) is different from the search data on SL (‘1’). Pass transistor N_0 is then open and discharges ML to the ground.

- When $SEN = '0'$, $WE = '1'$ and $Data = '1'$, of M_0 and M_1 are switched to be in AP state and P state, respectively. The corresponding storage data is now ‘1’, which matches again the search data ‘1’. The expected output value $Qm = '0'$ is obtained during the third read phase “Read 3”.

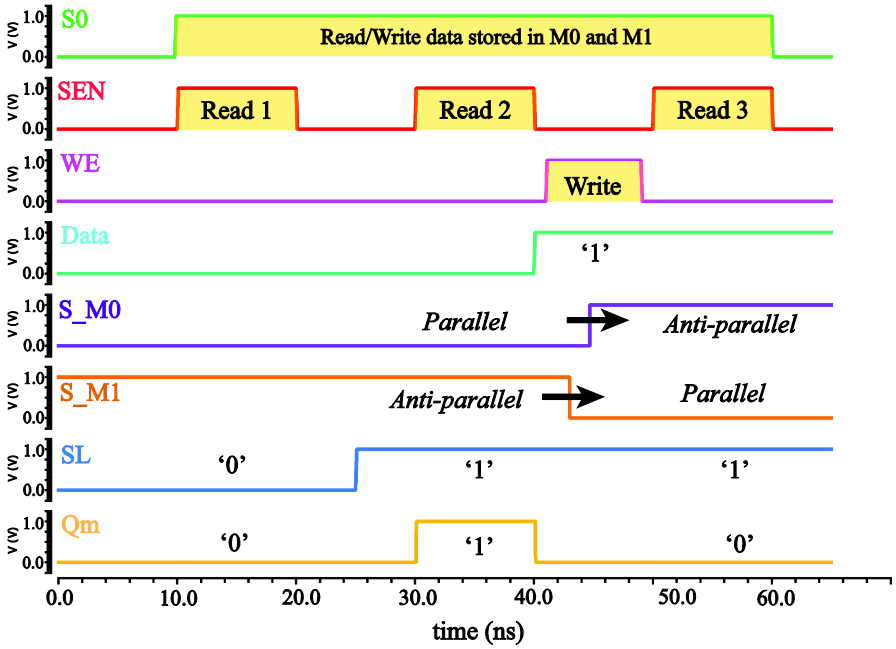


Figure 4.4 Transient simulation of the basic CAM cell. S_M_0 and S_M_1 represent the states of MTJs (M_0 and M_1).

Table 4.2 Performance comparison of different CAMs

	Proposed CAM	RM-CAM [155]	DW-CAM [156]	MTJ-CAM [157]	SRAM-CAM [158]
Technology	28 nm	65 nm	90 nm	0.18 μm	45 nm
Search delay	0.11 ns	~0.45 ns	5 ns	5 ns	0.306 ns
Energy (fj/bit/search)	~3.2	~12	~30	7.1	0.533
Static power	No	No	No	No	Yes
Cell area	19/N+2	11/N+2	12T	8T	8T

N represents the number of word line

Table 4.2 summarizes the performance comparison among different magnetic CAMs and the optimized CMOS-based CAM. The search operation of the NVCAM needs only 110 ps, thanks to the fast sensing of PCSA circuit. The energy consumption in case of "Mismatch" (which is larger than the case of "Match") is as low as 3.2 fJ/bit/search. The attractive feature of non-volatile magnetic CAMs is the non-volatility, which can eliminate the standby power in power-off state. Even though SRAM-based CAM proposed in [158] shows low search energy, it still suffers from high static power issue because power must be supplied to maintain the storage data. MTJ-based CAM proposed in [157] uses voltage mode sensing. That is, a continuous static current is applied when comparing the storage data with the search one. The proposed NVCAM, on the contrary, provides dynamic sensing and gives better

energy save. Same as RM-CAM in [155], comparison circuit of context NVCAM is shared by several MTJs to optimize the cell area. Figure 4.5 shows that area efficiency becomes more significant when increasing the word number. Multi-context NVCAM promises fast context switching because all the storage elements are directly connected with the comparison circuit.

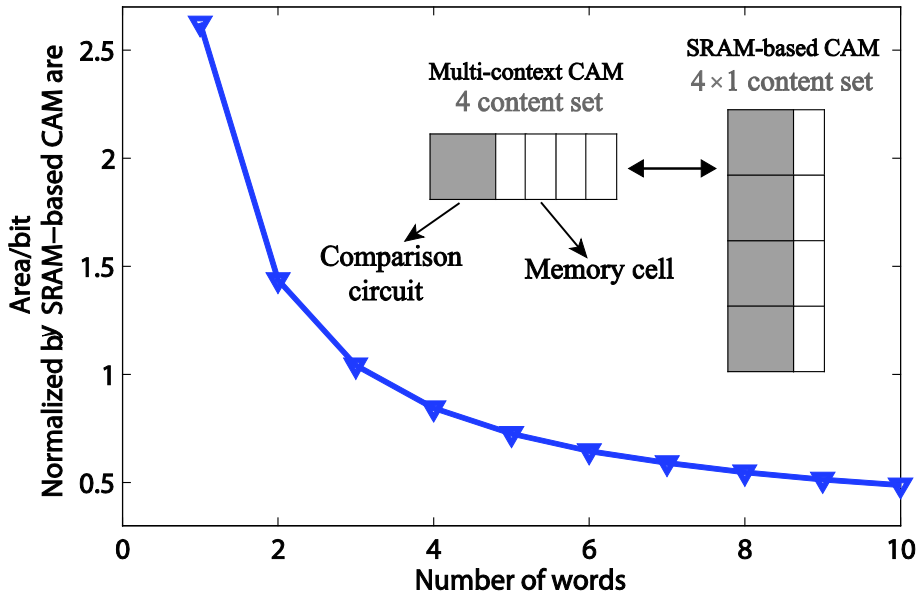


Figure 4.5 Bit-cell cost versus the number of words

In the following, we will focus on the reliability performance of the NVCAM. We consider the CMOS process variations and 3% process variations of MTJ, including TMR ratio, free layer thickness and oxide barrier thickness. As the basic element of NVCAM, we performed MC simulation of the PCSA-based CAM cell shown in Figure 4.3. Figure 4.6(a) shows that BER decreases greatly from 39% to 0.6% with the TMR ratio increases from 50% to 350%. This can be explained by larger resistance difference between two correspondent MTJs leads to larger sense margin between two branches. Therefore, larger TMR should be provided for reliable sensing. Figure 4.6(b) shows the BER values with respect to the size of different transistors in the CAM cell. It is confirmed that larger transistor size leads to lower BER. Each curve is realized by configuring the size of corresponding transistors (e.g., discharge transistors T_{dis}) while all the other transistors are always kept to the minimum size. Here, T_{sep} is the separating transistor (N_9 and N_{10}) shown in Figure 4.3. T_{dis} is the discharge transistor (N_3 and N_4). $T_{\log}$ is the transistor constituting the CMOS logic tree ($N_5 - N_8$). T_{inv} is the NMOS transistor in two inverters (N_1 and N_2). By increasing the size of transistors in the comparison circuit, resistances in two branches can be decreased and in turn sense margin $\Delta I = |I_{read0} - I_{read1}|$ is increased. Further MC simulations show that this 4×4

NVCAM can reach nearly zero *BER* by tripling its overall area.

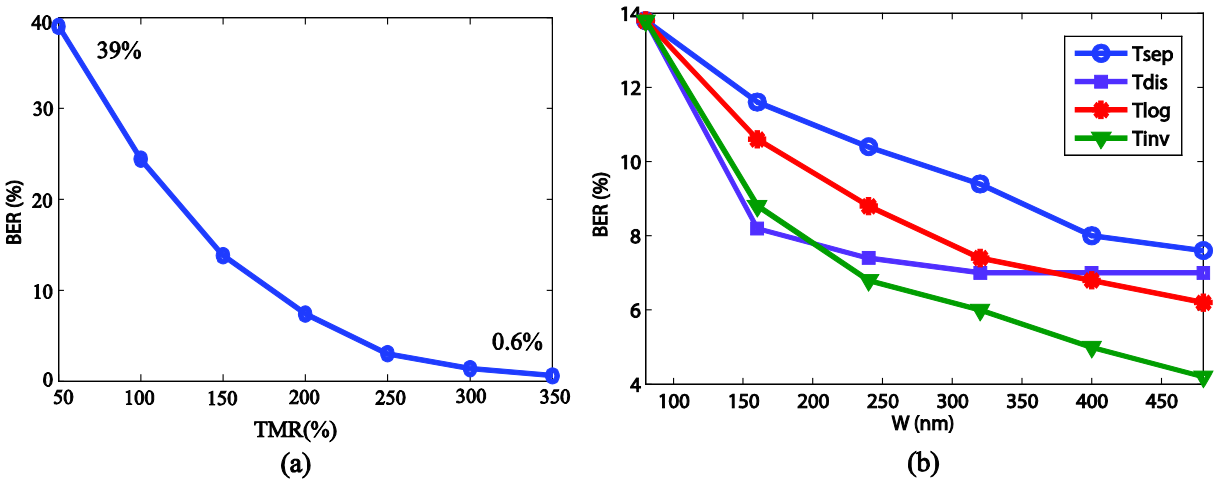


Figure 4.6 Sensing bit-error-rate (BER) of the CAM cell with respect to (a) the TMR value with all the transistors kept in minimum size (b) the size (W) of different transistors in the comparison cell with TMR(0)=150%

We need decoder to control the signals $S_0 - S_3$ (see Figure 4.3) and switch between the contexts. The mainstream CMOS-based decoders need power to keep data. If an unexpected power-off occurs, new logic operation has to be re-executed after the input data is retrieved from the memory block. However, as the previous data is lost, everything has to be started again with additional transfer energy and latency loss. The search operation of the NVCAM needs to start from the very beginning, which greatly reduces the performances of the circuits. Non-volatile decoder that can store data by using MTJs is a solution for this issue. And we propose for the first time magnetic decoder (MD) to address the word to be compared or written.

4.3 Magnetic decoder (MD) for word line selection

Two MDs, i.e., decoder based on shift register (SRMD) and decoder based on counter (CMD), serve as switching circuit for the NVCAM. The former has four MFFs and four 2-1 multiplexers connected in series. The latter is composed of a CMOS-based counter and a 2-4 decoder cell. Both MDs allow retaining the selected word location even in power-off state. Moreover, designers are able to choose one certain line to compare with the search data or to rewrite the storage data according to specific requirements.

4.3.1 MD based on shift register (SRMD)

4.3.1.1 SRMD circuit design

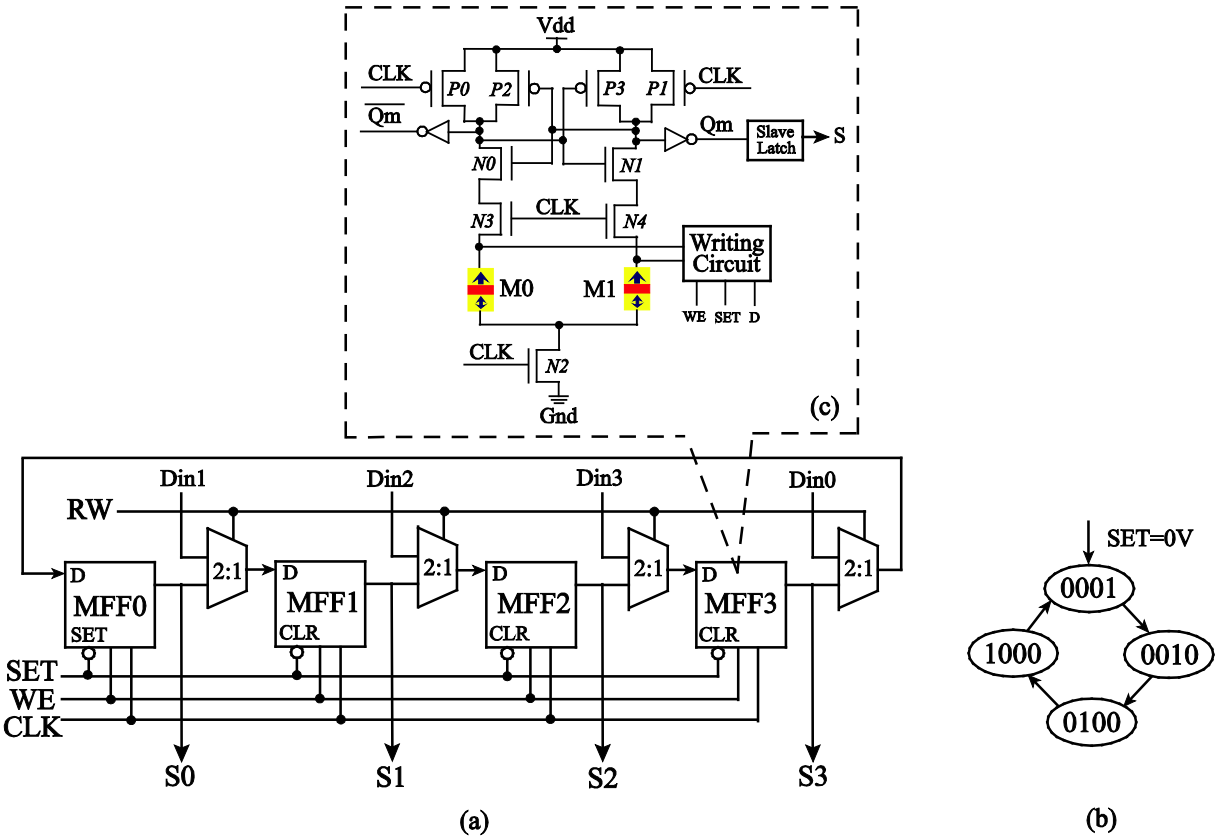


Figure 4.7 (a) Schematic of the magnetic decoder based on shift register (SRMD) for word line selection (b) State diagram of SRMD ($S_3S_2S_1S_0$) (c) Magnetic flip-flop (MFF) using a couple of MTJs that are always in complementary states

Figure 4.7(a) shows the structure of SRMD, whose basic element is MFF. S (Selection signal for NVCAM) is activated ($S = '1'$) when M_0 is in anti-parallel state and M_1 is in parallel state. Table 4.3 summarizes its two function modes.

- In Mode-1 ($RW = '1'$), all the outputs $S_0 - S_3$ equal to the inputs $D_{in0} - D_{in3}$, selecting a certain word line in the memory table of the NVCAM to be loaded or to be written.

- In Mode-2 ($RW = '0'$), it works as a circular shift register, searching word line-by-line. Data stored in MTJs are firstly initialized by setting the control signal SET to low voltage. All the outputs $S_1 - S_3$ equal to '0' except S_0, which equals to '1' instead. When SET is turned to be '1', the output of each stage (e.g., S_0) is connected to the input of MFF of the next stage (e.g. MFF_1) on the low voltage of a clock pulse. It should be noted that S_3 is feed back to the MFF of the least significant stage (MFF_0).

Table 4.3 State table of the SRMD

	RW	**SET**	**S_{3next}**	**S_{2next}**	**S_{1next}**	**S_{0next}**
Mode-1	1	X	Din_3	Din_2	Din_1	Din_0
Mode-2	0	0	0	0	0	1
	0	1	S_2	S_1	S_0	S_3

X means don't' care, that is, either '0' or '1' is a valid value

4.3.1.2 Simulation and analysis of SRMD

Figure 4.8 shows the transient simulation waveforms of the SRMD including three parts:

1) In the initialization phase when SET is low, "0001" are written into MTJs in $MFF_3 - MFF_0$ during the first pre-charge phase (P1) and detected during the first evaluation phase (E1).

2) In shift register phase, SET is forced high and RW stays low. The previous output $S_0 = '1'$ is written into the next MFF_1 during the second pre-charge phase (P2), and the next output S_1 is detected to be '1' during the second evaluation phase (E2) and so on. In this way, data '1' propagates from the least significant bit (S_0) all the way to the most significant bit (S_3) on each rising edge of a clock pulse. Data '1' will be written back into MFF_0 when $S_3 = '1'$. This MD allows the proposed NVCAM to be searched line-by-line until a "Match" is detected.

3) In the third phase, RW turns high, writing input data $Din_3 - Din_0$ into the MFFs. When a word "1000" is applied to $Din_{3\text{-}0}$, it is conformed that the expected outputs are observed as $S_3 = '1'$, $S_2 = '0'$, $S_1 = '0'$, $S_0 = '0'$.

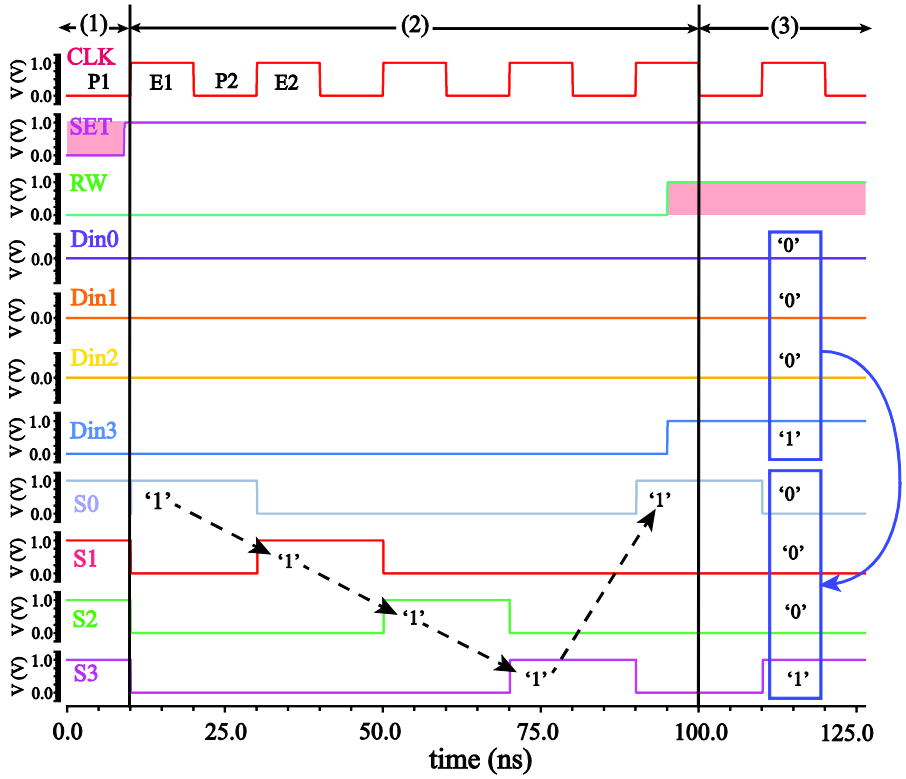


Figure 4.8 Transient simulation of the SRMD

4.3.2 MD based on counter (CMD)

4.3.2.1 CMD circuit design

As shown in Figure 4.9(a), the second MD based on counter uses a 2-4 decoder cell, whose inputs are connected with the outputs of a CMOS-based counter. Same as the SRMD, the MD based on counter (CMD) works in two modes:

- In Mode-1 ($RW = '1'$), outputs $Q_0 - Q_1$ equal to inputs $D_{in0} - D_{in1}$.
- In Mode-2 ($RW = '0'$), we switch from one selected word to the next one (or context switching), which is realized by the CMOS-based counter (see Figure 4.9(b)), following the state diagram shown in Figure 4.9(c).

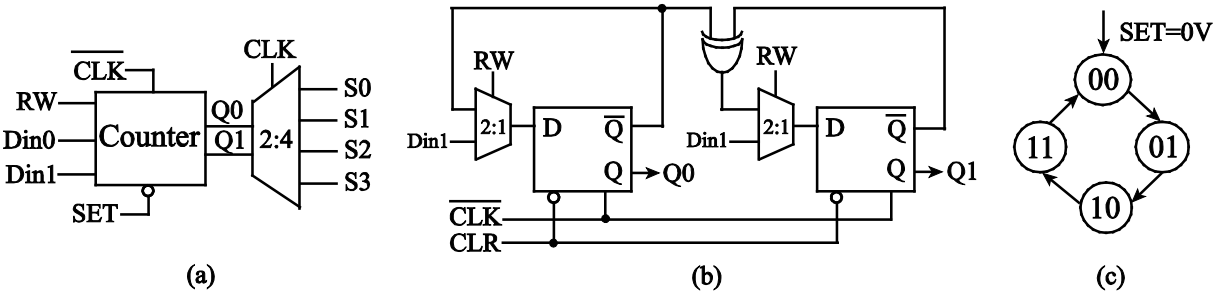
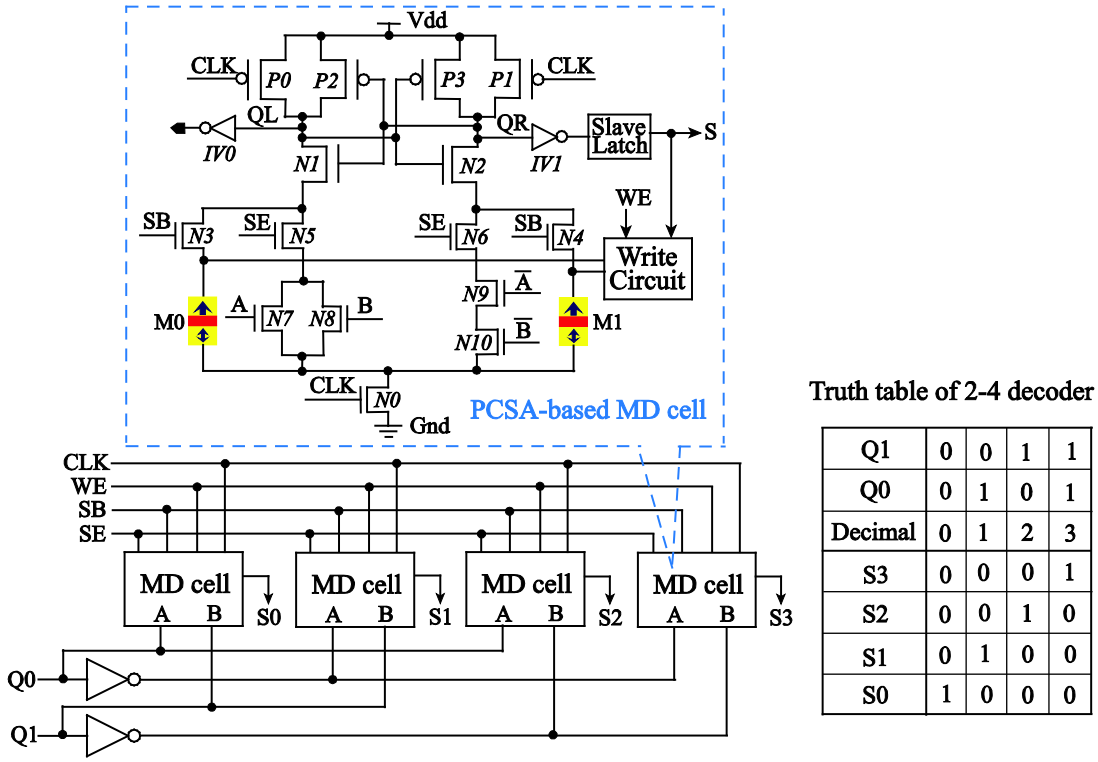


Figure 4.9 (a) Schematic of the magnetic decoder based on counter (CMD) (b) Structure of the CMOS-based counter (c) State diagram of the CMOS-based counter (Q_1Q_0)

We propose the first 2-4 decoder cell, where the output data is restored in non-volatile devices (e.g., MTJs). Figure 4.10 illustrates the structure of the 2-4 decoder cell, containing two inputs (Q_0 and Q_1), four outputs ($S_0 - S_3$), two read/write control signals (CLK and WE), two selection signals (SB and SE) and four PCSA-based MD cells. Four NMOS transistors ($N_7 - N_{10}$) compose the dynamic decoder logic part, and two MTJs always in complementary states store the output data S. $N_3 - N_6$ are mode selection transistors.



Truth table of 2-4 decoder

Q1	0	0	1	1
Q0	0	1	0	1
Decimal	0	1	2	3
S3	0	0	0	1
S2	0	0	1	0
S1	0	1	0	0
S0	1	0	0	0

Figure 4.10 Schematic of the non-volatile 2-4 decoder cell

By activating or deactivating the selection signals SE (for non-volatile data sensing mode, Mode 1) and SB (for CMOS-based dynamic decoder, Mode 2), the proposed MD cell performs two discharge modes during the evaluation phase (CLK = '1').

- In Mode 1 ($SE = '1'$ and $SB = '0'$), the sensing currents pass the sub-branches composed of $N_7 - N_{10}$. S will be '1' only if both inputs A and B are '0'. In this case, both N_9 and N_{10} are open, enabling a sensing current passing through the right branch while the left branch is blocked. QR is then discharged to the ground.

- In Mode 2 ($SE = '0'$ and $SB = '1'$), two MTJs are loaded. S will be '1' when M_0 is in anti-parallel state and M_1 is in parallel state. Output signal S is directly connected to the writing circuit and controls the direction of the writing current.

Writing operation takes place when CLK is low and the activation signal WE is high. In the case of $S = '0'$, M_0 is written into parallel state and M_1 into anti-parallel state, and vice versa. The 2-4 decoder cell promises area saving by sharing the same sense amplifier for normal CMOS-based dynamic decoder mode and non-volatile data sensing mode. Moreover, due to the symmetric structure, the impact of the sneak current is reduced and the sensing operation is more suitable. Based on these achievements, general K to 2^K (or K-2^K) decoders can also be built (K is the number of inputs).

4.3.2.2 Simulation and analysis of SRMD

Figure 4.11 shows the simulation of the PCSA-based MD cell with 1 V supply voltage for sensing, 1.2 V for MTJ writing. Inputs A and B are both initialized to '0'. SE is first activated to perform dynamic decoder mode (Mode 1). $S = '1'$ can be obtained on the output node, which is then backed up into M_0 and M_1 during the following pre-charge phase. Output data '1' can be recovered by sensing the states of MTJs after a sudden power-off during "Read Mode 2". Finally, inputs A and B are switched to '1' at the point M, and the output S is now '0' during the third discharge phase.

The proposed PCSA-based MD cell can achieve high sensing speed (～122.9 ps for "Mode 1" and ～137.5 ps for "Mode 2" when keeping the sensing energy low (～5.92 fJ for Mode 1 and ～4.76 fJ for Mode 2). The dynamic sensing currents passing through M_0 and M_1 are ～7.33 μA and ～11.37 μA, respectively, which are much smaller than the critical writing current I_{C0} (～50 μA).

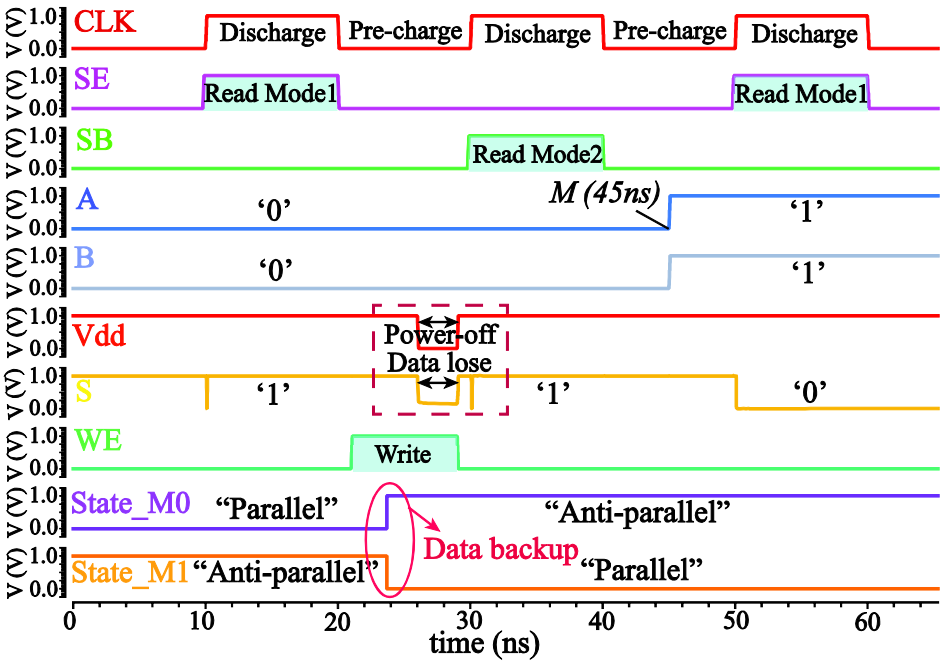


Figure 4.11 Transient simulation of the basic MD cell

The simulation of the 2-4 decoder cell (see Figure 4.10) including all possible input cases ($Q_1Q_0 = "00","01","10","11"$) is shown in Figure 4.12. Each case includes two read operations ("Read Mode1" and "Read Mode2") and one writing operation. For instance, the expected outputs $S_3 - S_0 = "0001"$ are obtained when two inputs are both '0' ("Case 1").

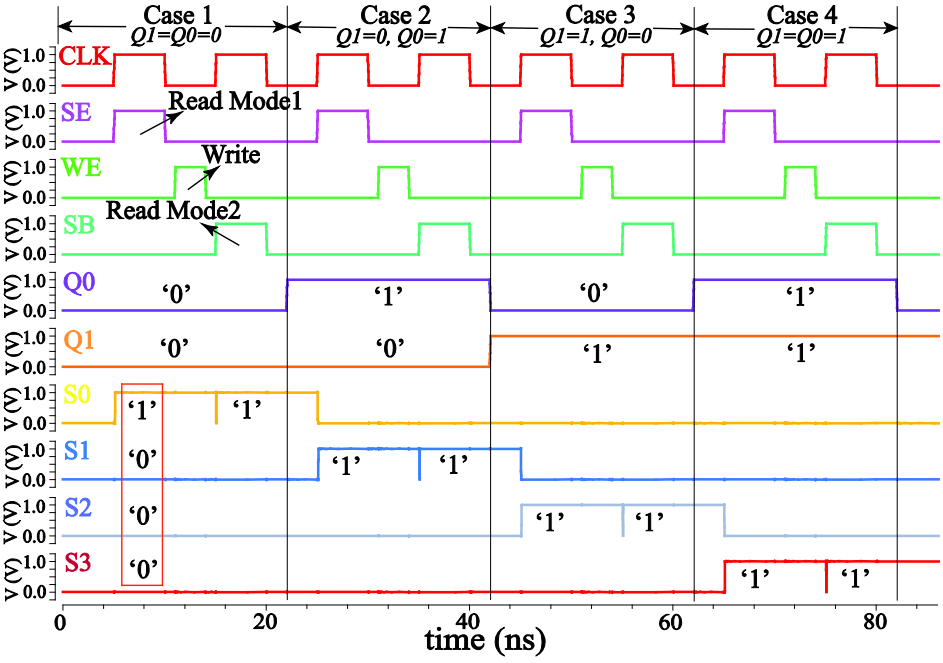


Figure 4.12 Simulation of the 2-4 MD (see Figure 4.10)

Full simulation of the CMD is illustrated in Figure 4.13. When signal $CLR = '0'$, Q_1Q_0 is

initialized to “00” through the CMOS-based counter. S_0 will be ‘1’ while other outputs are ‘0’ when meeting a rising edge of the clock signal CLK. During period (2), data ‘1’ propagates from S_0 all the way to S_3, and then back to the least significant bit after four cycles. RW turns high in (3) to directly choose the fourth line with $Din_1Din_0 = "11"$.

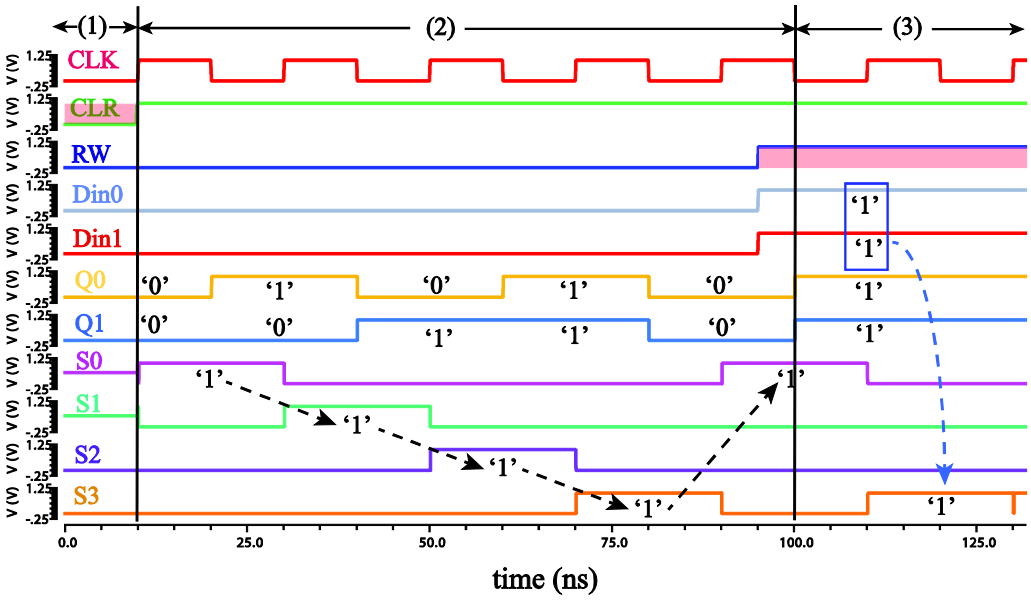


Figure 4.13 Transient simulation of the CMD (see Figure 4.9(a))

As discussed above, the PCSA-based circuit promises ultra-low read currents to avoid erroneous writing during the sensing phase. However, this also leads to low sensing margin (ΔI) and relatively high sensing BER. Besides, the proposed 2-4 decoder cell cannot meet the requirement of large resistance difference between two branches for reliable sensing, due to limited TMR value. Simulation results show that $\Delta I \approx$ 9.73 µA (or 9.07 µA) and BER = 15.2% (or 17.8%) with M_0 = AP (or P), M_1 = P (or AP) and all the transistors ($N_0 - N_{10}$ and $P_0 - P_3$) in minimum size.

In the following, we study the impacts of two main factors, i.e., the TMR ratio and the size of different transistors, on the sensing performance and find ways to improve the reliability of the proposed 2-4 MD through MC simulations. As it can be seen in Figure 4.14(a), the sensing BER decreases greatly from 36.4% (37.4%) to 1.2% (2.4%) when the TMR ratio increases from 50% to 350% with M_0 = AP (P), M_1 = P (AP). Figure 4.14(b) shows the BER values with respect to the size of different transistors, where T_{sel_MTJ} and T_{dis} represent the MTJ selection transistors (N_3 and N_4) and discharge transistor (N_0). It confirms that larger

transistor size leads to lower BER. Further MC simulations show that nearly zero BER can be reached with $T_{sel\_MTJ} = T_{dis} = 560\ nm$, while all the other transistors are kept in minimum size.

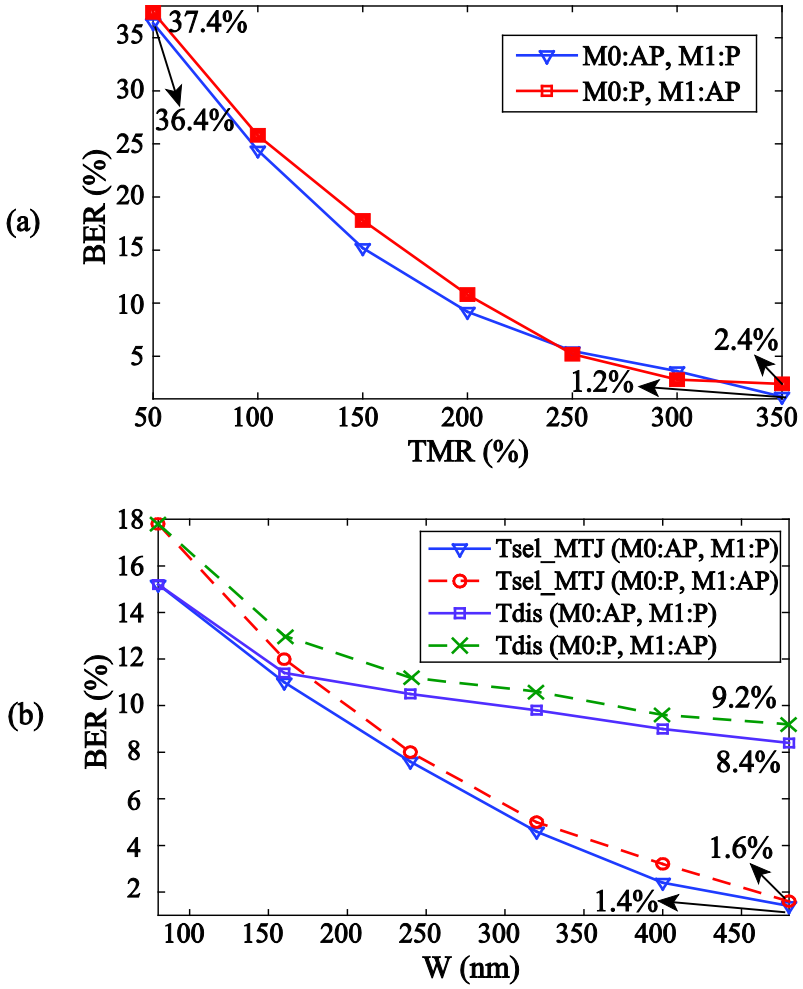


Figure 4.14 Sensing bit-error-rate (BER) of the MD cell with respect to (a) the TMR value (b) the size (W) of different transistors in the comparison cell

When comparing two proposed MDs, SRMD has less area overhead due to its simpler circuit. However, it may consume much more power and long data transfer delay (~ns) because 4-bit outputs $S_3 - S_0$ are written to the next stage (MFF) during each cycle. That is, there are four parts of writing energy for switching the data stored in MTJs. CMD can significantly reduce the writing energy thanks to its two-mode function mechanism. Data is backed up to MTJs or not according to specific applications.

4.4 Full implementation of NVCAM with switching circuit

The full function of the proposed multi-context NVCAM and its access circuit (SRMD or CMD) is demonstrated in Figure 4.15. In our simulation, data stored in four words are initialized to be "0000" , "1111", "1010" and "0101". When $PRE = '0'$ ("P1"), the match line ML is pre-charged to V_{dd}. Search word ("0101" as an example) is loaded on the search lines $SL_3 - SL_0$. The first storage word ($Word_0$) is addressed by SRMD or CMD. When $PRE = '1'$ ("C1"), $Word_0$ is compared against the search word with a response of "Miss". Afterwards, S_1 is enabled to load $Word_1$, and so on. Simulation waves show that only the last word ($Word_3$) matches the search data because ML maintaining unchanged during the fourth comparison phase ("C4").

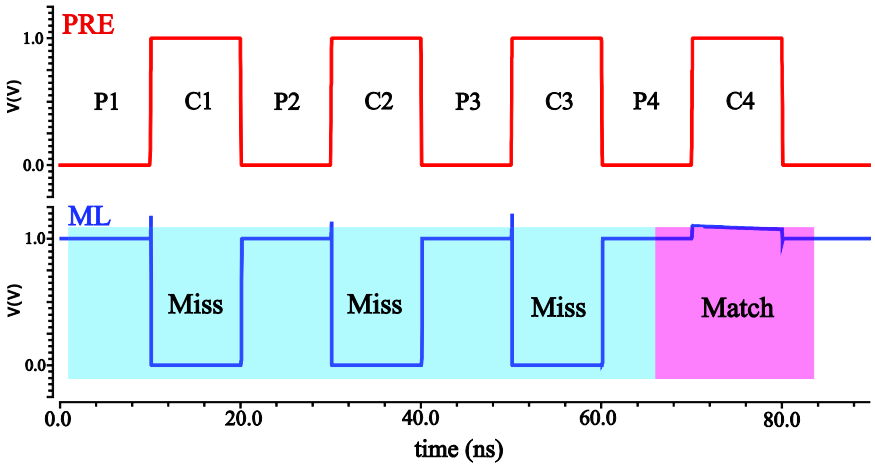


Figure 4.15 Full simulation of the proposed multi-context NVCAM. "P" and "C" represent the pre-charge phase and the comparison phase, respectively.

As discussed above, the proposed NVCAM loads word line-by-line for writing or comparing with the search word. Since four bits share the same comparison and writing circuits, MDs are investigated to switch the addressed word. This search approach promises to save energy because the non-addressed memory cells, that are waiting to be searched, do not need to be sensed once a "Match" is detected. However, it needs longer comparison time than the parallel search approach, which limits its use in some applications that require high speed. As shown in Figure 4.16, we propose a multi-context NVCAM structure, where each CAM cell is composed of a comparison circuit integrating several non-volatile memory cells (4 in Figure 4.16). When performing a search operation, all the storage words are compared with the search word and output "Match" or "Miss" on match lines $ML_0 - ML_m$.

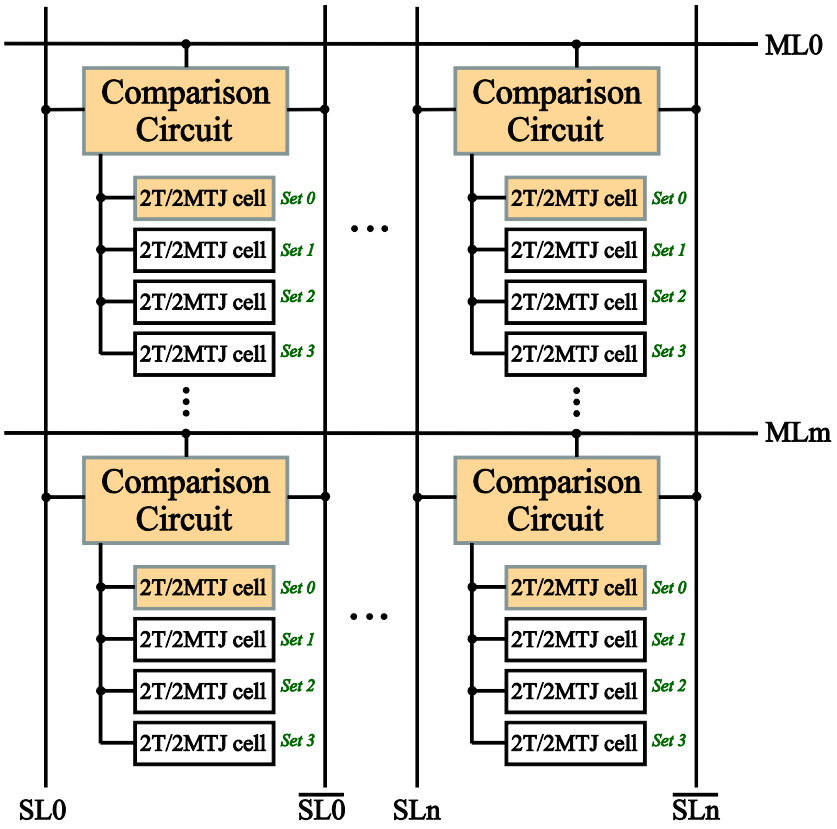


Figure 4.16 Four-set multi-context NVCAM structure. Set_0 is activated as an example.

The multi-context NVCAM can be used in translation lookaside buffers (TLB) [156]. Search operation is performed after one context set is loaded. For instance, Set_0 is selected in Figure 4.16. Other content sets ($Set_1 - Set_3$) store data of inactivate processes. When process is switched, data correspond to the next process is loaded. If the number of set is increased, the search power consumption keeps low because only the selected set for the active process will be compared.

4.5 Conclusion

In this chapter, we designed an NVCAM with the capabilities of low bit-cell cost, high search speed (110 ps), low dynamic power (～ 3.2 fJ/bit/search) and nearly zero standby energy. It is an application combining the multi-context hybrid MTJ/CMOS and LIM architecture. For the 2-input non-volatile logic circuits investigated in Chapter 3, the non-volatile input is stored in MTJs that do not change very often, while the volatile input is able to be changed with a high frequency by CMOS-based circuit. The proposed NVCAM is exactly the case. It takes the advantages of non-volatility of MTJs and changeable search data. Multiple contexts share the same comparison circuit, leading to higher density and fast switching.

In order to store the search location of the NVCAM in non-volatile state, two magnetic decoders, i.e., SRMD and CMD, were developed for word line selection. If there is an occurrence of unexpected powered off, search operation starts again from the stored line instead of the beginning of the comparison sequence. Hence, the overall energy is reduced.

Multi-context NVCAM was finally proposed to realize parallel search for high speed requirement. However, it also suffers from high search energy issue since all the words are simultaneously compared against the search word. By integrating several-bit non-volatile data in the same comparison circuit, the multi-context NVCAM can search the selected content set while remaining the unselected sets for inactive processes for further energy saving.

General conclusion

This thesis aims at designing and simulating the non-volatile logic circuits integrating MTJs not only as storage cells but also as operands. A 1KB MRAM and NVCAM were proposed based on the multi-context hybrid MTJ/CMOS circuit as well.

The background and theory of spintronics were first presented. We studied the current research and development of MTJ (especially the increasing of TMR ratio), its writing approaches and applications in memories and logic circuits. High TMR ratio (> 600% at room temperature) makes it easier to detect the state of MTJ (parallel or anti-parallel) by CMOS-based circuits (e.g., PCSA). Among various switching approaches, STT writing approach shows good performances on power dissipation, scalability and speed. Spin-Hall-assisted STT writing approach features higher writing speed and lower writing energy, but it needs more complex circuit design when combined with the CMOS technology. On this basis, MTJ has been used in both memory (e.g., MRAM) and logic design. For the logic application, LIM architecture opens a way to bring the non-volatile memories directly into the logic circuit. This architecture greatly shortens the communication distance, and hence reducing the transfer delay and energy.

The model of PMA STT-MTJ, which was used as the key element in our circuit design, was presented. We learned the physical models integrated in this model, the way to combine it with the CMOS circuits and different simulation methods. And then we validated this model by DC, transient and MC simulations on Cadence platform. Hybrid MTJ/CMOS structure was analyzed including reading (i.e., PCSA) and writing circuits. Multi-context hybrid MTJ/CMOS circuit was designed for further area efficiency, where several storage memory cells share the same reading and writing circuit. In order to solve the issues caused by the asymmetric structure, symmetric structure and SPCSA reading circuits were proposed. By using the multi-context hybrid MTJ/CMOS structure and the MTJ model, we designed and validated a novel 1KB MRAM, which was part of our work for the DIPMEM project.

Logic and arithmetic circuits based on the LIM architecture were designed by using the MTJ model. After the design and analysis of NVLGs, we focused on the conception and improvement of NVFA, which is the largest part in the thesis. First, the structure of 1-bit NVFA was detailed, followed by the study of the effect of three factors, i.e., discharge transistor size, MTJ resistance-area product (RA) and TMR ratio, on the operation delay and energy. It was shown that larger area, lower RA and higher TMR ratio were expected for

high-speed and low-power operation. When compared to the CMOS-based FA, NVFA showed advantages on static power consumption and die area due to 3-D integration technology. Second, we proposed and compared three structures of 8-bit NVFA which realizes full non-volatility and enables addition of two words. Finally, we optimized the NVFA in terms of reliability and writing speed/energy. From the circuit-level, voltage-mode sensing circuit using 2T/2MTJ memory cell can replace the magnetic flip-flop to store, read and write one binary data of NVFA. According to the MC simulation results, this circuit had nearly small sensing errors (<1%) thanks to larger sensing margin. From the device-level, spin-Hall-assisted STT writing was applied to a three-terminal MTJ device, replacing the two-terminal MTJ device in the FA. Simulation results showed that this writing approach had switching time smaller than 1 ns and switching power consumption lower than 100 fJ, which can hardly be reached with the STT writing mechanism. Besides, lower write voltage was required, thus the endurance of oxide barrier can be enhanced.

The LIM architecture was also applied to the design of NVCAM. In each CAM cell, there was 1-bit data stored in the memory cell and 1-bit data coming from the search lines. This NVCAM had advantages in search speed and power consumption when compared to other CAMs. Two magnetic decoders (MDs) were designed for word line selection. In such MDs, the search location can not only be chosen but also stored in MTJs for data security. Search operation could start from one certain line instead of the very beginning, saving large amount of search energy. Another structure of non-volatile CAM called multi-context NVCAM was finally proposed for parallel search.

We propose some points to improve our work and continue the research on non-volatile logic circuits.

First, we have only integrated the process variations of MTJ and stochastic switching in this model. Other phenomenon can be considered for the future work such as temperature fluctuation, dielectric breakdown phenomenon and sub-volume activation effect, etc. The temperature fluctuation may influence the robustness of the hybrid MTJ/CMOS circuits based on current-mode sense amplifier. In non-volatile logic circuits, a large writing current is necessary to meet the requirement of high-speed computing, which however, increases the risk of MTJ barrier breakdown. For this raison, breakdown phenomenon also needs to be integrated in this model and considered when evaluating the performances of the hybrid circuits.

Second, the limitation of two-terminal MTJ device is that it has the same reading and writing

current path. If we replace all the transistors in the CMOS logic tree (A, $\overline{A}$, C_i and $\overline{C_i}$) of the FA, a lot more CMOS transistors need to be added in order to separate the MTJ during the writing operation. Therefore, the PCSA-based reading circuit is no more practical. The spin-Hall-assisted STT writing approach cannot solve this problem because it still needs writing current (I_{STT}) passing through the MTJ stack during writing operation. An optimized sensing circuit should be designed. Otherwise, other writing mechanisms or spintronics devices that can completely separate the read and write current paths need to be developed.

References

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics,* vol. 3, no. 20, pp. 33-35, 1965.

[2] M. N. Baibich, J. M. Broto, A. Fert, F. N. Van Dau and F. Petroff, "Giant magnetoresistance of (001) Fe/(001) Cr magnetic superlattices," *Physical review letters,* vol. 61, no. 21, p. 2472, 1988.

[3] M. Julliere, "Tunneling between ferromagnetic films," *Physics letters A,* vol. 54, no. 3, pp. 225-226, 1975.

[4] S. Ikeda, J. Hayakawa, Y. Ashizawa et al., "Tunnel magnetoresistance of 604% at 300 K by suppression of Ta diffusion in CoFeB/MgO/CoFeB pseudo-spin-valves annealed at high temperature," *Applied Physics Letters,* vol. 93, no. 8, p. 2508, 2008.

[5] T. M. Maffitt, J. K. DeBrosse, J. A. Gabric et al., "Design considerations for MRAM," *IBM Journal of Research and Development,* vol. 1, no. 25, p. 50, 2006.

[6] W. J. Gallagher and S. S. P. Parkin, "Development of the magnetic tunnel junction MRAM at IBM: from first junctions to a 16-Mb MRAM demonstrator chip," *IBM Journal of Research and Development,* vol. 50, no. 1, pp. 5-23, 2006.

[7] I. L. Prejbeanu, M. Kerekes, R. C. Sousa et al., "Thermally assisted MRAM," *Journal of Physics: Condensed Matter,* vol. 19, no. 16, p. 165218, 2007.

[8] B. Razavi, Design of analog CMOS integrated circuits, McGraw-Hill Higher Education, 2001.

[9] N. H. E. Weste and D. M. Harris, CMOS VLSI Design: A Circuits and Systems Perspective, India: Pearson Education, 2006.

[10] P. A. M. Dirac, "The quantum theory of the electron," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 1928.

[11] P. M. Tedrow and R. Meservey, "Spin-dependent tunneling into ferromagnetic nickel," *Physical Review Letters,* vol. 26, no. 4, p. 192, 1971.

[12] G. Binasch, P. Grünberg, F. Saurenbach and W. Zinn, "Enhanced magnetoresistance in layered magnetic structures with antiferromagnetic interlayer exchange," *Physical review B,* vol. 39, no. 7, p. 4828, 1989.

[13] S. M. Thompson, "The discovery, development and future of GMR: The Nobel Prize 2007," *Journal of Physics D: Applied Physics,* vol. 41, no. 9, p. 093001, 2008.

[14] N. F. Mott, "The resistance and thermoelectric properties of the transition metals," in *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 1936.

[15] J. Daughton, J. Brown, E. Chen et al., "Magnetic field sensors using GMR multilayer," *IEEE Transactions on magnetics,* vol. 30, no. 6, pp. 4608-4610, 1994.

[16] C. Reig, M.-D. Cubells-Beltrán and D. Ramírez Muñoz, "Magnetic field sensors based on giant magnetoresistance (GMR) technology: Applications in electrical current

sensing," *Sensors,* vol. 9, no. 10, pp. 7919-7942, 2009.

[17] C. Tsang, R. E. Fontana, T. Lin and D. E. Heim, "Design, fabrication and testing of spin-valve read heads for high density recording," *IEEE Transactions on Magnetics,* vol. 30, no. 6, pp. 3801-3806, 1994.

[18] G. A. Prinz, "Magnetoelectronics," *Science,* vol. 282, no. 5394, pp. 1660-1663, 1998.

[19] K. Takanashi, "Fundamentals of Magnetoresistance Effects," in *Spintronics for Next Generation Innovative Devices,* 2015, pp. 1-20.

[20] T. Miyazaki and N. Tezuka, "Giant magnetic tunneling effect in Fe/Al2O3/Fe junction," *Journal of Magnetism and Magnetic Materials,* vol. 139, no. 3, pp. L231-L234, 1995.

[21] J. S. Moodera, L. R. Kinder, T. M. Wong and R. Meservey, "Large magnetoresistance at room temperature in ferromagnetic thin film tunnel junctions," *Physical review letters,* vol. 74, no. 16, p. 3273, 1995.

[22] C. Chappert, A. Fert and F. N. Van Dau, "The emergence of spin electronics in data storage," *Nature materials,* vol. 6, no. 11, pp. 813-823, 2007.

[23] W. Zhao, C. Chappert, V. Javerliac and J.-P. Nozière, "High Speed, High Stability and Low Power Sensing Amplifier for MTJ/CMOS Hybrid Logic Circuits," *IEEE Transactions on Magnetics,* vol. 45, no. 10, pp. 3784-3787, 2009.

[24] D. Wang, C. Nordman, J. M. Daughton, Z. Qian and J. Fink, "70% TMR at Room Temperature for SDT Sandwich Junctions With CoFeB as Free and Reference Layers," *IEEE Transactions on Magnetics,* vol. 40, no. 4, pp. 2269-2271, 2004.

[25] H. X. Wei, Q. H. Qin, M. Ma, R. Sharif and X. F. Han, "80% tunneling magnetoresistance at room temperature for thin Al–O barrier magnetic tunnel junction with CoFeB as free and reference layers," *Journal of applied physics,* vol. 101, no. 9, 2007.

[26] S. Yuasa et al., "Tunnel Magnetoresistance Effect and Its Applications," http://www.jst.go.jp/sicp/ws2009_sp1st/presentation/15.pdf.

[27] AIST, "Development of MgO-MTJ devices,"

https://unit.aist.go.jp/src/cie/en_teams/en_teams_metal.html.

[28] W. H. Butler, T. C. Schulthess and X.-G. Zhang, "Spin-dependent tunneling conductance of Fe|MgO|Fe sandwiches," *Physical Review B,* vol. 63, no. 5, p. 054416, 2001.

[29] J. Mathon and A. Umerski, "Theory of tunneling magnetoresistance of an epitaxial Fe/MgO/Fe (001) junction," *Physical Review B,* vol. 63, no. 22, p. 220403, 2001.

[30] S. S. P. Parkin, C. Kaiser, A. Panchula et al., "Giant tunnelling magnetoresistance at room temperature with MgO (100) tunnel barriers," *Nature materials,* vol. 3, no. 12, pp. 862-867, 2004.

[31] S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki and K. Ando, "Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions," *Nature Materials,* vol. 3, no. 12, pp. 868-871, 2004.

[32] D. D. Djayaprawira, K. Tsunekawa, M. Nagai, H. Maehara, S. Yamagata and N. Watanabe, "230% room-temperature magnetoresistance in CoFeB/MgO/CoFeB

magnetic tunnel junctions," *Applied Physics Letters,* vol. 86, no. 9, p. 092502, 2005.

[33] J. Hayakawa , S. Ikeda , F. Matsukura , H. Takahashi and H. Ohno, "Dependence of giant tunnel magnetoresistance of sputtered CoFeB/MgO/CoFeB magnetic tunnel junctions on MgO barrier thickness and annealing temperature," *Japanese Journal of Applied Physics,* vol. 44, no. 4L, p. L587, 2005.

[34] S. Yuasa, A. Fukushima, H. Kubota, Y. Suzuki and K. Ando, "Giant tunneling magnetoresistance up to 410% at room temperature in fully epitaxial Co/MgO/Co magnetic tunnel junctions with bcc Co (001) electrodes," *Applied Physics Letters,* vol. 89, no. 4, 2006.

[35] Y. M. Lee, J. Hayakawa, S. Ikeda, F. Matsukura and H. Ohno, "Effect of electrode composition on the tunnel magnetoresistance of pseudo-spin-valve magnetic tunnel junction with a MgO tunnel barrier," *Applied Physics Letters,* vol. 90, no. 21, p. 2507, 2007.

[36] S. A. Wolf, D. D. Awschalom, R. A. Buhrman et al., "Spintronics: a spin-based electronics vision for the future," *Science,* vol. 294, no. 5546, pp. 1488-1495, 2001.

[37] S. Tehrani, J. M. Slaughter, M. Deherrera et al., "Magnetoresistive random access memory using magnetic tunnel junctions," *Proceedings of the IEEE,* vol. 91, no. 5, pp. 703-714, 2003.

[38] B. N. Engel, J. Åkerman, B. Butcher et al., "A 4-Mb toggle MRAM based on a novel bit and switching method," *IEEE Transactions on Magnetics,* vol. 41, no. 1, pp. 132-136, 2005.

[39] J. Wang and P. P. Freitas , "Low-current blocking temperature writing of double barrier magnetic random access memory cells," *Applied physics letters,* vol. 84, no. 6, pp. 945-947, 2004.

[40] I. L. Prejbeanu, W. Kula, K. Ounadjela et al., "Thermally assisted switching in exchange-biased storage layer magnetic tunnel junctions," *IEEE Transactions on Magnetics,* vol. 40, no. 4, pp. 2625-2627, 2004.

[41] I. L. Prejbeanu, S. Bandiera, J. Alvarez-Hérault, R. C. Sousa, B. Dieny and J. -P. Nozières, "Thermally assisted MRAMs: ultimate scalability and logic functionalities," *Journal of Physics D: Applied Physics,* vol. 46, no. 7, p. 074002, 2013.

[42] W. Zhao, E. Belhaire, C. Chappert, B. Dieny and G. Prenat, "TAS-MRAM-based low-power high-speed runtime reconfiguration (RTR) FPGA," *ACM Transactions on Reconfigurable Technology and Systems (TRETS),* vol. 2, no. 2, p. 8, 2009.

[43] J. C. Slonczewski, "Current-driven excitation of magnetic multilayers," *Journal of Magnetism and Magnetic Materials,* vol. 159, no. 1, pp. L1-L7, 1996.

[44] L. Berger, "Emission of spin waves by a magnetic multilayer traversed by a current," *Physical Review B,* vol. 54, no. 13, p. 9353, 1996.

[45] J. Z. Sun, "Spin-current interaction with a monodomain magnetic body: A model study," *Physical Review B,* vol. 62, no. 1, p. 570, 2000.

[46] A. Brataas, A. D. Kent and H. Ohno, "Current-induced torques in magnetic materials," *Nature materials,* vol. 11, no. 5, pp. 372-381, 2012.

[47] J. Z. Sun, "Spin angular momentum transfer in current-perpendicular nanomagnetic

junctions," *IBM journal of research and development*, vol. 50, no. 1, pp. 81-100, 2006.

[48] D. C. Ralph and M. D. Stiles, "Spin transfer torques," *Journal of Magnetism and Magnetic Materials*, vol. 320, no. 7, pp. 1190-1216, 2008.

[49] E. Chen, D. Apalkov, Z. Diao et al., "Advances and future prospects of spin-transfer torque random access memory," *IEEE Transactions on Magnetics*, vol. 46, no. 6, pp. 1873-1878, 2010.

[50] D. Apalkov, S. Watts, A. Driskill-Smith et al., "Comparison of scaling of in-plane and perpendicular spin transfer switching technologies by micromagnetic simulation," *IEEE transactions on magnetics*, vol. 46, no. 6, pp. 2240-2243, 2010.

[51] K. C. Chun, H. Zhao, J. D. Harms et al., "A scaling roadmap and performance evaluation of in-plane and perpendicular MTJ based STT-MRAMs for high-density cache memory," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 2, pp. 598-610, 2013.

[52] T. Kishi, H. Yoda, T. Kai et al., "Lower-current and Fast switching of A Perpendicular TMR for High Speed and High density Spin-Transfer-Torque MRAM," in *IEEE International Electron Devices Meeting*, 2008.

[53] K. J. Lee, O. Redon and B. Dieny, "Analytical investigation of spin-transfer dynamics using a perpendicular-to-plane polarizer," *Applied Physics Letters*, vol. 86, no. 2, p. 022505, 2005.

[54] S. Mangin, D. Ravelosona, J. A. Katine and E. E. Fullerton, "Current-induced magnetization reversal in nanopillars with perpendicular anisotropy," *Nature Materials*, vol. 5, no. 3, pp. 210-215, 2006.

[55] H. Meng and J.-P. Wang, "Spin transfer in nanomagnetic devices with perpendicular anisotropy," *Applied physics letters*, vol. 88, no. 17, p. 2506, 2006.

[56] M. Yoshikawa, E. Kitagawa, T. Nagase et al., "Tunnel magnetoresistance over 100% in MgO-based magnetic tunnel junction films with perpendicular magnetic L1-FePt electrodes," *IEEE Transactions on Magnetics*, vol. 44, no. 11, pp. 2573-2576, 2008.

[57] Z. R. Tadisina, "Perpendicular magnetic anisotropy materials for reduced current switching devices," PhD thesis, The University of Alabama TUSCALOOSA, 2010.

[58] S. Ikeda, K. Miur, H. Yamamot, K. Mizunuma, H. D. Gan, S. Kanai, J. Hayakawa, F. Matsukura and H. Ohno, "A perpendicular-anisotropy CoFeB–MgO magnetic tunnel junction," *Nature Materials*, vol. 9, no. 9, pp. 721-724, July 2010.

[59] Z. Wang, W. Zhao, E. Deng et al., "Magnetic non-volatile flip-flop with spin-Hall assistance," *physica status solidi (RRL)-Rapid Research Letters*, vol. 9, no. 6, pp. 375-378, 2015.

[60] J. E. Hirsch, "Spin Hall Effect," *Physical Review Letters*, vol. 83, no. 9, p. 065001, 1999.

[61] V. M. Edelstein, "Spin polarization of conduction electrons induced by electric current in two-dimensional asymmetric electron systems," *Solid State Communications*, vol. 73, no. 3, pp. 233-235, 1990.

[62] J. Kim, J. Sinha, M. Hayashi et al., "Layer thickness dependence of the current-induced effective field vector in Ta|CoFeB|MgO," *Nature materials*, vol. 12, no. 3, pp. 240-245, 2013.

[63] L. Liu, C.-F. Pai, Y. Li et al., "Spin-torque switching with the giant spin Hall effect of tantalum," *Science,* vol. 336, no. 6081, pp. 555-558, 2012.

[64] C.-F. Pai, L. Liu, Y. Li et al., "Spin transfer torque devices utilizing the giant spin Hall effect of tungsten," *Applied Physics Letters,* vol. 101, no. 12, p. 122404, 2012.

[65] M. Cubukcu, O. Boulle, M. Drouard et al., "Spin-orbit torque magnetization switching of a three-terminal perpendicular magnetic tunnel junction," *Applied Physics Letters,* vol. 104, no. 4, p. 042406, 2014.

[66] A. van den Brink, S. Cosemans, S. Cornelissen, M. Manfrini, A. Vaysset, W. Van Roy, T. Min, H. J. M. Swagten and B. Koopmans, "Spin-Hall-assisted magnetic random access memory," *Applied Physics Letters,* vol. 1, no. 012403, p. 104, 2014.

[67] K. Jabeur, G. Di Pendina, F. Bernard-Granger and G. Prenat, "Spin orbit torque non-volatile flip-flop for high speed and low energy applications," *IEEE Electron Device Letters,* vol. 35, no. 3, pp. 408-410, 2014.

[68] K.-W. Kwon, S. H. Choday, Y. Kim et al., "SHE-NVFF: spin Hall effect-based nonvolatile flip-flop for power gating architecture," *IEEE Electron Device Letters,* vol. 35, no. 4, pp. 488-490, 2014.

[69] X. Wang, Y. Chen, H. Li, D. Dimitrov and H. Liu, "Spin torque random access memory down to 22 nm technology," *IEEE transactions on magnetics,* vol. 44, no. 11, pp. 2479-2482, 2008.

[70] J.-G. Zhu , "Magnetoresistive random access memory: the path to competitiveness and scalability," in *Proceedings of the IEEE,* 2008.

[71] S. A. Wolf, J. Lu, M. R. Stan, E. Chen and D. M. Treger, "The promise of nanomagnetics and spintronics for future logic and universal memory," in *Proceedings of the IEEE,* 2010.

[72] B. Jovanović, R. M. Brum and L. Torres, "Evaluation of hybrid MRAM/CMOS cells for “normally-off and instant-on” computing," *Analog Integrated Circuits and Signal Processing,* vol. 81, no. 3, pp. 607-621, 2014.

[73] S. Senni, R. M. Brum, L. Torres et al., "Potential applications based on NVM emerging technologies," in *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*, 2015.

[74] X. Dong, X. Wu, G. Sun et al., "Circuit and microarchitecture evaluation of 3D stacking magnetic RAM (MRAM) as a universal memory replacement," in *45th ACM/IEEE Design Automation Conference (DEC)*, Anaheim, 2008.

[75] International Roadmap for Semiconductor (ITRS), 2011.

http://www.itrs.net/

[76] M. Hosomi, H. Yamagishi, T. Yamamoto et al., "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM," in *IEEE InternationalElectron Devices Meeting*, Washington, 2005.

[77] M. Aoki, H. Iwasa and Y. Sato, "A novel voltage sensing 1T/2MTJ cell with resistance ratio for highly stable and scalable MRAM," in *Symposium on VLSI Circuits*, 2005.

[78] H. Tanizaki, T. Tsuji, J. Otani et al., "A high-density and high-speed 1T-4MTJ MRAM with Voltage Offset Self-Reference Sensing Scheme," in *IEEE Asian Solid-State*

Circuits Conference (ASSCC), Hangzhou, 2006.

[79] R. Patel, E. Ipek and E. G. Friedman, "2T–1R STT-MRAM memory cells for enhanced on/off current ratio," *Microelectronics Journal*, vol. 45, no. 2, pp. 133-143, 2014.

[80] H. Noguchi, K. Kushida, K. Ikegami et al., "A 250-MHz 256b-I/O 1-Mb STT-MRAM with advanced perpendicular MTJ based dual cell for nonvolatile magnetic caches to reduce active power of processors," in *Symposium on VLSI Technology (VLSIT)*, Kyoto, 2013.

[81] T. Ohsawa, H. Koike, S. Miura et al., "A 1 Mb nonvolatile embedded memory using 4T2MTJ cell with 32 b fine-grained power gating scheme," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 6, pp. 1511-1520, 2013.

[82] W. Zhao, S. Chaudhuri, C. Accoto et al., "Cross-point architecture for spin-transfer torque magnetic random access memory," *IEEE Transactions on Nanotechnology*, vol. 11, no. 5, pp. 907-917, 2012.

[83] A. W. Burks, H. H. Goldstine and J. Von Neumann, Preliminary discussion of the logical design of an electronic computing instrument, Springer Berlin Heidelberg, 1982, pp. 399-413.

[84] W. H. Kautz, "Cellular Logic-in-Memory Arrays," *IEEE Transactions on Computers*, vol. 100, no. 8, pp. 719-727, 1969.

[85] G. Prenat , M. El Baraji, W. Guo, R. Sousa, V. Javerliac, J.-P. Nozieres, W. Zhao and E. Belhaire, "CMOS/Magnetic Hybrid Architectures," in *IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Marrakech, 2007.

[86] W. Zhao, E. Belhaire, C. Chappert, F. Jacquet and P. Mazoyer, "New non-volatile logic based on spin-MTJ," *Physica Status Solidi a-Applications and Materials Science*, vol. 205, no. 6, pp. 1373-1377, 2008.

[87] Y. Lakys, W. Zhao, J.-O. Klein and C. Chappert, "Magnetic Look-Up Table (MLUT) featuring radiation hardness, high performance and low power," in *International Symposium on Applied Reconfigurable Computing*, 2011.

[88] D. Chabi, W. Zhao, E. Deng, Y. Zhang, N. Ben Romdhane, J.-O. Klein and C. Chappert, "Ultra Low Power Magnetic Flip-Flop Based on checkpointing/Power Gatingand Self-Enable Mechanisms," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 6, pp. 1755-1765, 2014.

[89] S. Onkaraiah, M. Reyboz, F. Clermidy, J.-M. Portal, M. Bocquet, C. Muller, Hraziia, C. Anghel and A. Amara, "Bipolar ReRAM based non-volatile flip-flops for low-power architectures," in *New Circuits and Systems Conference (NEWCAS)*, 2012.

[90] S. Matsunaga, J. Hayakawa, S. Ikeda et al., "Fabrication of a Nonvolatile Full Adder Based on Logic-in-Memory Architecture Using Magnetic Tunnel Junctions," *Applied Physics Express*, vol. 1, no. 9, p. 091301, 2008.

[91] D. Allwood, G. Xiong, C. C. Faulkner, D. Atkinson, D. Petit and R. P. Cowburn, "Magnetic domain-wall logic," *Science*, vol. 309, no. 5741, pp. 1688-1692, 2005.

[92] Y. Zhang, "Compact modeling and hybrid circuit design for spintronic devices based on current-induced switching," PhD thesis, Université Paris Sud-Paris, 2014.

[93] L. Berger , "Low-field magnetoresistance and domain drag in ferromagnets," *Journal of*

Applied Physics, vol. 49, no. 3, pp. 2156-2161, 1978.

[94] M. Hayashi, L. Thomas, R. Moriya, C. Rettner and S. S. P. Parkin, "Current-controlled magnetic domain-wall nanowire shift register," *Science,* vol. 320, no. 5873, pp. 209-211, 2008.

[95] S. S. P. Parkin, M. Hayashi and L. Thomas, "Magnetic Domain-Wall Racetrack Memory," *Science,* vol. 320, no. 5873, pp. 190-194, 2008.

[96] A. Annunziata, M. C. Gaidis, L. Thomas et al., "Racetrack memory cell array with integrated magnetic tunnel junction readout," in *International Electron Devices Meeting,* 2011.

[97] W. Zhao, D. Ravelosona, J.-O. Klein and C. Chappert, "Domain Wall Shift Register-Based Reconfigurable Logic," *IEEE Transactions on Magnetics,* vol. 47, no. 10, pp. 2966-2969, 2011.

[98] H.-P. Trinh , W. Zhao, J.-O. Klein, Y. Zhang, D. Ravelsona and C. Chappert, "Magnetic Adder Based on Racetrack Memory," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 60, no. 6, pp. 1469-1477, 2013.

[99] B. Behin-Aein, D. Datta, S. Salahuddin and S. Datta, "Proposal for an all-spin logic device with built-in memory," *Nature nanotechnology,* vol. 5, no. 4, pp. 266-270, 2010.

[100] C. Augustine, G. Panagopoulos, B. Behin-Aein, S. Srinivasan, A. Sarkar and K. Roy, "Low-power functionality enhanced computation architecture using spin-based devices," in *IEEE/ACM International Symposium on Nanoscale Architectures,* San Diego, 2011.

[101] Q. An, L. Su, J.-O. Klein, S. Le Beux, I. O'Connor and W. Zhao, "Full-adder circuit design based on all-spin logic device," in *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH),* Boston, 2015.

[102] M. Sharad, C. Augustine, G. Panagopoulos and K. Roy, "Boolean and non-Boolean computation with spin devices," in *IEEE International Electron Devices Meeting (IEDM),* San Francisco, 2012.

[103] Z. Pajouhi, S. Venkataramani, K. Yogendra, A. Raghunathan and K. Roy, "Exploring spin-transfer-torque devices for logic applications," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 34, no. 9, pp. 1441-1454, 2015.

[104] Y. Wang, Y. Zhang, E. Deng, J.-O. Klein, L. A. B. Naviner and W. Zhao, "Compact model of magnetic tunnel junction with stochastic spin transfer torque switching for reliability analyses," *Microelectronics Reliability,* vol. 54, no. 9, pp. 1774-1778, March 2014.

[105] Y. Zhang, W. Zhao, Y. Lakys et al., "Compact modeling of perpendicular-anisotropy CoFeB/MgO magnetic tunnel junctions," *IEEE Transactions on Electron Devices,* vol. 59, no. 3, pp. 819-826, 2012.

[106] W. F. Brinkman, R. C. Dynes and J. M. Rowell, "Tunneling Conductance of Asymmetrical Barriers," *Journal of Applied physics,* vol. 41, no. 5, pp. 1915-1921, April 1970.

[107] S. Zhang, P. M. Levy, A. C. Marley and S. S. P. Parkin, "Quenching of magnetoresistance by hot electrons in magnetic tunnel junctions," *Physical Review*

Letters, vol. 79, no. 19, p. 3744, 1997.

[108] G. D. Fuchs, I. N. Krivorotov, P. M. Braganca et al., "Adjustable spin torque in magnetic tunnel junctions with two fixed layers," *Applied Physics Letters,* vol. 86, no. 15, p. 152509, 2005.

[109] J. C. Slonczewski, "Currents, torques, and polarization factors in magnetic tunnel junctions," *Physical Review B,* vol. 71, no. 2, p. 024411, 2005.

[110] J. Z. Sun, R. P. Robertazzi, J. Nowak et al., "Effect of subvolume excitation and spin-torque efficiency on magnetic switching," *Physical Review B,* vol. 84, no. 6, p. 064413, 2011.

[111] Z. Diao, Z. Li, S. Wang et al., "Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory," *Journal of Physics: Condensed Matter,* vol. 19, no. 6, p. 165209, 2007.

[112] R. H. Koch, J. A. Katine and J. Z. Sun, "Time-Resolved Reversal of Spin-Transfer Switching in a Nanomagnet," *Physical review letters,* vol. 92, no. 8, p. 088302, 2004.

[113] R. Heindl, W. H. Rippard, S. E. Russek, M. R. Pufall and A. B. Kos, "Validity of the thermal activation model for spin-transfer torque switching in magnetic tunnel junctions," *Journal of Applied Physics,* vol. 109, no. 7, p. 073910, 2011.

[114] D. C. Worledge, G. Hu, D. W. Abraham, J. Z. Sun, P. L. Trouilloud, J. Nowak, S. Brown, M. C. Gaidis, E. J. O'Sullivan and R. P. Robertazzi, "Spin torque switching of perpendicular TaCoFeBMgO-based magnetic," *Applied Physics Letters,* vol. 98, no. 2, p. 2501, 2011.

[115] Cadence Verilog-A Language Reference, 2006.

[116] Virtuoso Spectre Circuit Simulator Datasheet, Cadence.

[117] Eldo User's Manual, Mentor Graphics, 2005.

[118] http://www.ief.u-psud.fr/~zhao/spinlib.html.

[119] M. Alioto and P. Gaerano, Model and design of Bipolar and MOS Current-Mode logic: CML, ECL and SCL Digital Circuits, 2006: Springer Science & Business Media.

[120] W. Zhao, T. Devolder, Y. Lakys, J.-O. Klein, C. Chappert and P. Mazoyer, "Design considerations and strategies for high-reliable STT-MRAM," *Microelectronics Reliability,* vol. 51, no. 9, pp. 1454-1458, 2011.

[121] W. Kang, Z. Wang, W. Zhao et al., "A low-cost built-in error correction circuit design for STT-MRAM reliability improvement," *Microelectronics Reliability,* vol. 53, no. 9, pp. 1224-1229, 2013.

[122] M. Hariyama, S. Ishihara, N. Idobata and M. Kameyama, "Non-Volatile Multi-Context FPGAs Using Hybrid Multiple-Valued/Binary Context Switching Signals," in *International Conference on Engineering of Reconfigurable Systems & Algorithms (ERSA)*, Las Vegas, Nevada, USA, 2008.

[123] W. Zhao, E. Belhaire, C. Chappert and P. Mazoyer, "Spin transfer torque (STT)-MRAM-based runtime reconfiguration FPGA circuit," *ACM Transactions on Embedded Computing Systems (TECS),* vol. 9, no. 2, 2009.

[124] D. Schinkel, E. Mensink, E. Klumperink and E. Tuijl, "A double-tail latch-type voltage

sense amplifier with 18ps setup+ hold time," in *IEEE International Solid-State Circuits Conference*, 2007.

[125] H. Jeon and Y.-B. Kim, "A low-offset high-speed double-tail dual-rail dynamic latched comparator," in *Great lakes symposium on VLSI*, New York, NY, USA, 2010.

[126] W. Kang, E. Deng, J.-O. Klein et al., "Separated Pre-Charge Sensing Amplifier for Deep Submicron MTJ/CMOS Hybrid Logic Circuits," *IEEE Transactions on Magnetics*, vol. 50, no. 6, pp. 1-5, 2014.

[127] "CMOS028 Design Rules Manual," STMicroelectronics, 2011.

[128] K. Shi, "Power reduction methodology in 28nm SOC production design—What have changed?" in *IEEE Faible Tension Faible Consommation*, 2013.

[129] R. Scheuerlein, W. Gallagher, S. Parkin et al., "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*, San Francisco, 2000.

[130] M. Durlam, P. Naji, A. Omair et al., "A low power 1 Mbit MRAM based on 1T1MTJ bit cell integrated with copper interconnects," in *Symposium on VLSI Circuits Digest of Technical Papers*, Honolulu, 2002.

[131] N. S. Kim, T. Austin, D. Baauw et al., "Leakage current: Moore's law meets the static power," *computer*, vol. 36, no. 12, pp. 68-75, 2003.

[132] C. J. Lin, S. H. Kang, K. Lee et al., "45 nm low power CMOS logic compatible embedded STT MRAM Utilizing a Reverse-Connection 1T/1MTJ Cell," in *IEEE International Electron Devices Meeting (IEDM)*, 2009.

[133] W. C. Black and B. Das, "Programmable logic using giant-magnetoresistance and spin-dependent tunneling devices," *Journal of Applied Physics*, vol. 87, no. 9, pp. 6674-6679, 2000.

[134] A. Mochizuki, H. Kimura, M. Ibuki and T. Hanyu, "MR-based logic-in-memory circuit for low-power VLSI," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 88, no. 6, pp. 1408-1415, 2005.

[135] J. P. Wang and X. Yao, "Programmable spintronic logic devices for reconfigurable computation and beyond—History and outlook," *Journal of Nanoelectronics and Optoelectronics*, vol. 3, no. 1, pp. 12-23, 2008.

[136] W. Zhao, M. Moreau, E. Deng et al, "Synchronous non-volatile logic gate design based on resistive switching memories," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 2, pp. 443-454, 2014.

[137] Y. Gang, W. Zhao, J.-O. Klein, C. Chappert and P. Mazoyer, "A High-Reliability, Low-Power Magnetic Full Adder," *IEEE Transactions on Magnetics*, vol. 47, no. 11, pp. 4611-4616, 2011.

[138] A. Mochizuki, H. Kimura, M. Ibuki and T. Hanyu , "TMR-based logic-in-memory circuit for low-power VLSI," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 88, no. 6, pp. 1408-141, 2005.

[139] M. W. Allam and M. I. Elmasry, "Dynamic current mode logic (DyCML): A new low-power high-performance logic style," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 3, pp. 550-558, 2001.

[140] "CMOS40 design rule manual," STMicroelectronics, 2012.

[141] K. T. S. Oldham, "The Doctrine of Description: Gustav Kirchhoff, Classical Physics, and the" purpose of All Science" in 19th-century Germany," University of California, Berkeley, 2008.

[142] A. E. Kennelly, "The equivalence of triangles and three-pointed stars in conducting networks," *Electrical world and engineer,* vol. 34, no. 12, pp. 413-414, 1899.

[143] H. Meng , J. Wang and J.-P. Wang, "A Spintronics full adder for magnetic CPU," *IEEE Electron Device Letters,* vol. 26, no. 6, pp. 360-362, 2005.

[144] H. Yoda, S. Fujita, N. Shimomura et al., "Progress of STT-MRAM technology and the effect on Normally-Off computing systems," in *International Electron Devices Meeting,* 2012.

[145] W. Xu, T. Zhang and Y. Chen, "Spin-transfer torque magnetoresistive content addressable memory (CAM) cell structure design with enhanced search noise margin," in *IEEE International Symposium on Circuits and Systems,* Seattle, 2008.

[146] S. Paul , S. Mukhopadhyay and S. Bhunia, "A Circuit and Architecture Codesign Approach for a Hybrid CMOS–STTRAM Nonvolatile FPGA," *IEEE Transactions on Nanotechnology,* vol. 10, no. 3, pp. 385-394, 2011.

[147] I. M. Miron, K. Garello, G. Gaudin, P.-J. Zermatten et al., "Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection," *Nature,* vol. 476, pp. 189-193, 2011.

[148] Z. Wang, W. Zhao, E. Deng, J.-O. Klein and C. Chappert, "Perpendicular-anisotropy magnetic tunnel junction switched by spin-Hall-assisted spin-transfer torque," *Journal of Physics D: Applied Physics,* vol. 48, no. 6, p. 065001, 2015.

[149] J. C. Slonczewski, "Conductance and exchange coupling of two ferromagnets separated by a tunneling barrier," *Physical Review B,* vol. 39, no. 10, p. 6995, 1989.

[150] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: A tutorial and survey," *IEEE Journal of Solid-State Circuits,* vol. 41, no. 3, pp. 712-727, 2006.

[151] H. Jarollahi, V. Gripon, N. Onizawa and W. J. Gross, "Algorithm and architecture for a low-power content-addressable memory based on sparse clustered networks," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 23, no. 4, pp. 642-653, 2015.

[152] Y.-J. Chang and Y.-H. Liao, "Hybrid-type CAM design for both power and performance efficiency," *IEEE transactions on very large scale integration (VLSI) systems,* vol. 16, no. 8, pp. 965-974, 2008.

[153] S. Choi, K. Sohn and H.-J. Yoo, "A 0.7-fJ/bit/search 2.2-ns search time hybrid-type TCAM architecture," *IEEE Journal of solid-state circuits,* vol. 40, no. 1, pp. 254-260, 2005.

[154] S. Matsunaga, K. Hiyama, A. Matsumoto et al., "Standby-power-free compact ternary content-addressable memory cell chip using magnetic tunnel junction devices," *Applied Physics Express,* vol. 2, no. 2, p. 023004, 2009.

[155] Y. Zhang, W. Zhao, J.-O. Klein, D. Ravelsona and C. Chappert, "Ultra-high density

content addressable memory based on current induced domain wall motion in magnetic track," *IEEE Transactions on Magnetics,* vol. 48, no. 11, pp. 3219-3222, 2012.

[156] R. Nebashi, N. Sakimura, Y. Tsuji et al., "A content addressable memory using magnetic do- main wall motion cells," in *Symposium on Circuits (VLSIC),* 2001.

[157] W. Xu, T. Zhang and Y. Chen, "Design of spin-torque transfer magnetoresistive RAM and CAM/TCAM with high sensing and search speed," *IEEE transactions on very large scale integration (VLSI) systems,* vol. 18, no. 1, pp. 66-74, 2010.

[158] M. Zackriya V and H. M. Kittur, "Precharge-Free, Low-Power Content-Addressable Memory," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 24, no. 8, pp. 2614 - 2621, 2016.

[159] Z. Wang, "Compact modeling and circuit design based on ferroelectric tunnel junction and spin-Hall-assisted spin-transfer torque," PhD thesis, Université Paris-Saclay, 2015.

List of publications

Journals:

[1] **E. Deng**, G. Prenat, L. Anghel, W. Zhao, “Non-volatile magnetic decoder based on MTJs”, Electronics Letters, vol. 52, no. 21, pp. 1774-1776, Oct. 2016.

[2] **E. Deng**, Z. Wang, J. O. Klein, G. Prenat, B. Dieny and W. Zhao, “High-Frequency Low-Power Magnetic Full-Adder Based on Magnetic Tunnel Junction With Spin-Hall Assistance,” IEEE Transactions on Magnetics, vol. 51, no. 11, pp. 1-4, Nov. 2015.

[3] **E. Deng** et al., "Synchronous 8-bit Non-Volatile Full-Adder based on Spin Transfer Torque Magnetic Tunnel Junction," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 62, no. 7, pp. 1757-1765, July 2015.

[4] **E. Deng**, W. Kang, Y. Zhang, J. O. Klein, C. Chappert and W. Zhao, "Design Optimization and Analysis of Multicontext STT-MTJ/CMOS Logic Circuits," IEEE Transactions on Nanotechnology, vol. 14, no. 1, pp. 169-177, Jan. 2015.

[5] **E. Deng**, Y. Zhang, J. O. Klein, D. Ravelsona, C. Chappert and W. Zhao, "Low Power Magnetic Full-Adder Based on Spin Transfer Torque MRAM," IEEE Transactions on Magnetics, vol. 49, no. 9, pp. 4982-4987, Sept. 2013.

[6] Y. Wang, H. Cai, L. A. B. Naviner, Y. Zhang, X. Zhao, **E. Deng** et al., "Compact Model of Dielectric Breakdown in Spin-Transfer Torque Magnetic Tunnel Junction," IEEE Transactions on Electron Devices, vol. 63, no. 4, pp. 1762-1767, April 2016.

[7] Z. Wang, W. Zhao, **E. Deng**, Y. Zhang and J. O. Klein, "Magnetic non-volatile flip-flop with spin-Hall assistance," physica status solidi (RRL)-Rapid Research Letters, vol. 9, no. 6, pp. 375-378, June 2015.

[8] Z. Wang, W. Zhao, **E. Deng**, J. O. Klein and C. Chappert, "Perpendicular-anisotropy magnetic tunnel junction switched by spin-Hall-assisted spin-transfer torque," Journal of Physics D: Applied Physics, vol. 48, no. 6, p. 065001, Jan. 2015.

[9] W. Kang, Z. Li, Z. Wang, **E. Deng** et al., "Variation-Tolerant High-Reliability Sensing Scheme for Deep Submicrometer STT-MRAM," IEEE Transactions on Magnetics, vol. 50, no. 11, pp. 1-4, Nov. 2014.

[10] W. Kang, W. Zhao, **E. Deng** et al., "A radiation hardened hybrid spintronic/CMOS

nonvolatile unit using magnetic tunnel junctions," Journal of Physics D: Applied Physics, vol. 47, no. 40, p. 405003, Sept. 2014.

[11] W. Kang, **E. Deng** et al., "Separated Precharge Sensing Amplifier for Deep Submicrometer MTJ/CMOS Hybrid Logic Circuits," IEEE Transactions on Magnetics, vol. 50, no. 6, pp. 1-5, June 2014.

[12] D. Chabi, W. Zhao, **E. Deng** et al., "Ultra Low Power Magnetic Flip-Flop Based on Checkpointing/Power Gating and Self-Enable Mechanisms," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 61, no. 6, pp. 1755-1765, June 2014.

[13] Y. Wang, Y. Zhang, **E. Deng**, J. O. Klein, L. A. B. Naviner et W. Zhao, "Compact model of magnetic tunnel junction with stochastic spin transfer torque switching for reliability analyses," Microelectronics Reliability, vol. 54, no. 9, pp. 1774-1778, 2014. (**Best paper** in 25th European Symposium on Reliability of Electron Devices, Failure Physics and Analysis)

[14] W. Zhao, M. Moreau, **E. Deng** et al., "Synchronous Non-Volatile Logic Gate Design Based on Resistive Switching Memories," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 61, no. 2, pp. 443-454, Feb. 2014.

Conferences with publications:

[1] **E. Deng**, G. Prenat, L. Anghel and W. Zhao, "Multi-context Non-volatile Content Addressable Memory Using Magnetic Tunnel Junctions," IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH'16), Beijing, July 2016, pp. 103-108. (**Best paper**)

[2] **E. Deng** et al., "Robust magnetic full-adder with voltage sensing 2T/2MTJ cell," IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH'15), Boston, MA, July 2015, pp. 27-32.

[3] J. M. Portal, M. Moreau, M. Bocquet, H. Aziza, D. Deleruyelle, C. Muller, Y. Zhang, **E. Deng** et al., "Analytical study of complementary memristive synchronous logic gates," IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH'13), New York, July 2013, pp. 70-75.

[4] Y. Zhang, **E. Deng** et al., "Synchronous full-adder based on complementary resistive switching memory cells," IEEE 11th International New Circuits and Systems Conference (NEWCAS), Paris, June 2013, pp. 1-4. (**Best paper**)

Other conferences:

[1] Non-Volatile Memory Technology Symposium (NVMTS), Beijing, China, 2015.

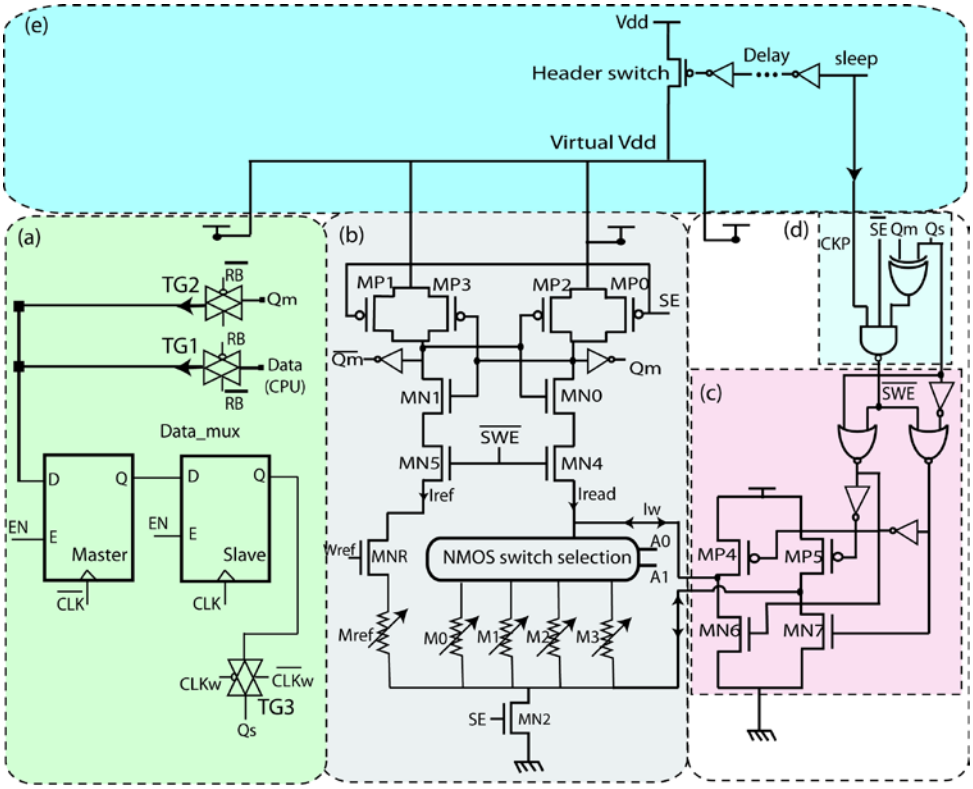
[2] IEEE International Magnetics Conference (INTERMAG), Dresden, Germany, 2014.

[3] Journées Nationales du Réseau Doctoral en Micro-nanoélectronique (JNRDM), Lille, France, 2014.

Book chapter:

Y. Zhang, W. Zhao, W. Kang, **E. Deng**, J. O. Klein and D. Revelosona, "Current-Induced Magnetic Switching for High-Performance Computing," Spintronics-based Computing, by Springer International Publishing, pp. 1-51, 2015. DOI 10.1007/978-3-319-15180-9.

Appendix A Schematic of the multi-context magnetic flip-flop (MFF)



(a) CMOS Volatile D-latch (b) 4-bits STT-MRAM including sense amplifier circuit (c) STT writing circuit (d) Self-enable control circuit (e) Power gating (PG) cell using sleep p-MOS transistor (header switch) [88]

(a) Three transmitter gates (TG_{1-3}) dedicate to transfer data between the CMOS flip-flop (CFF) and the STT-MRAM. CFF receive data from the computing unit if $RB = '0'$, and load data stored in one of the MTJs (M_{0-3}) if $RB = '1'$.

(b) Sensing operation is performed when $SE = W_{ref} = '1'$ and $\overline{SWE} = '1'$. One of the NMOS switches is turned ON by A_0 and A_1.

(c) 4T writing circuit is applied to the MFF. During the writing operation, $\overline{SWE} = '0'$, transistors MN_4 and MN_5 are turned OFF.

(d) The fixed writing pulse is replaced by a sequence of self-enabled read/write operations. The writing operation is activated only when $CKP = '1'$, $\overline{SE} = '1'$ and the input data Qs is equal to the current stored data Qm.

(e) During the standby mode, signal *sleep* is activated, disabling both writing and reading operations. It should be noted that the current CFF state should be stored in the local storage memory cell (MTJ) before completely powering off.

Appendix B Resistance comparison of the logic network

Table B.1.a Truth table of the optimized NV-AND/NV-NAND structure-2

A	B	R_L	R_R	Equivalent resistance comparison	Qm
0	0	$R_{OFF}+R_{AP}$	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_P$	$R_L>R_R$	0
0	1	$R_{OFF}+R_P$	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{AP}$	uncertain	uncertain
1	0	$R_{ON}+R_{AP}$	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_P$	$R_L>R_R$	0
1	1	$R_{ON}+R_P$	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{AP}$	uncertain	uncertain

Table B.1.b Resistance condition of the optimized NV-AND/NV-NAND structure-2

A	B	Condition	Qm
0	1	$R_{AP}-R_P=TMR\times R_P<\frac{R_{OFF}^2}{R_{OFF}+R_{ON}}$	0 (correct)
		$R_{AP}-R_P=TMR\times R_P>\frac{R_{OFF}^2}{R_{OFF}+R_{ON}}$	1 (error)
1	1	$R_{AP}-R_P=TMR\times R_P>\frac{R_{ON}^2}{R_{OFF}+R_{ON}}$	1 (correct)
		$R_{AP}-R_P=TMR\times R_P<\frac{R_{ON}^2}{R_{OFF}+R_{ON}}$	0 (error)

Table B.1.c Truth table of the optimized NV-AND/NV-NAND structure-3

A	B	R_L	R_R	Equivalent resistance comparison	Qm
0	0	$R_{OFF}+R_{AP}$	R_P	$R_L>R_R$	0
0	1	$R_{OFF}+R_P$	R_{AP}	uncertain	uncertain
1	0	$R_{ON}+R_{AP}$	R_P	$R_L>R_R$	0

1	1	$R_{ON}+R_P$	R_{AP}	uncertain	uncertain

Table B.1.d Resistance condition of the optimized NV-AND/NV-NAND structure-3

A	B	Condition	Qm
0	1	$R_{AP}-R_P=TMR\times R_P<R_{OFF}$	0 (correct)
		$R_{AP}-R_P=TMR\times R_P>R_{OFF}$	1 (error)
1	1	$R_{AP}-R_P=TMR\times R_P>R_{ON}$	1 (correct)
		$R_{AP}-R_P=TMR\times R_P<R_{ON}$	0 (error)

Table B.2.a Truth table of the optimized NV-OR/NV-NOR structure-2

A	B	R_L	R_R	Equivalent resistance comparison	Qm
0	0	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{AP}$	$R_{ON}+R_P$	uncertain	uncertain
0	1	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_P$	$R_{ON}+R_{AP}$	$R_L<R_R$	1
1	0	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{AP}$	$R_{OFF}+R_P$	uncertain	uncertain
1	1	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_P$	$R_{OFF}+R_{AP}$	$R_L<R_R$	1

Table B.2.b Resistance condition of the optimized NV-OR/NV-NOR structure-2

A	B	Condition	Qm
0	0	$R_{AP}-R_P=TMR\times R_P>\frac{R_{ON}^2}{R_{OFF}+R_{ON}}$	0 (correct)
		$R_{AP}-R_P=TMR\times R_P<\frac{R_{ON}^2}{R_{OFF}+R_{ON}}$	1 (error)
1	0	$R_{AP}-R_P=TMR\times R_P<\frac{R_{OFF}^2}{R_{OFF}+R_{ON}}$	1 (correct)
		$R_{AP}-R_P=TMR\times R_P>\frac{R_{OFF}^2}{R_{OFF}+R_{ON}}$	0 (error)

Table B.2.c Truth table of the optimized NV-OR/NV-NOR structure-3

A	B	R_L	R_R	Equivalent resistance comparison	Qm
0	0	R_{AP}	$R_{ON}+R_P$	uncertain	uncertain
0	1	R_P	$R_{ON}+R_{AP}$	$R_L<R_R$	1
1	0	R_{AP}	$R_{OFF}+R_P$	uncertain	uncertain
1	1	R_P	$R_{OFF}+R_{AP}$	$R_L<R_R$	1

Table B.2.d Resistance condition of the optimized NV-OR/NV-NOR structure-3

A	B	Condition	Qm
0	0	$R_{AP}-R_P=TMR\times R_P>R_{ON}$	0 (correct)
		$R_{AP}-R_P=TMR\times R_P<R_{ON}$	1 (error)
1	0	$R_{AP}-R_P=TMR\times R_P<R_{OFF}$	1 (correct)
		$R_{AP}-R_P=TMR\times R_P>R_{OFF}$	0 (error)

Table B.3.a Truth table and resistance configuration of CARRY/C_o logic

A	B	C_i	Resistance Comparison	C_o	Left sub-branch AC_i	Right sub-branch $\overline{AC_i}$
0	0	0	$R_L>R_R$	0	$2R_{OFF}$	$2R_{ON}$
0	0	1	$R_L>R_R$	0	$R_{OFF}+R_{ON}$	$R_{ON}+R_{OFF}$
0	1	0	$R_L>R_R$	0	$2R_{OFF}$	$2R_{ON}$
0	1	1	$R_L<R_R$	1	$R_{OFF}+R_{ON}$	$R_{ON}+R_{OFF}$
1	0	0	$R_L>R_R$	0	$R_{ON}+R_{OFF}$	$R_{OFF}+R_{ON}$
1	0	1	$R_L<R_R$	1	$2R_{ON}$	$2R_{OFF}$
1	1	0	$R_L<R_R$	1	$R_{ON}+R_{OFF}$	$R_{OFF}+R_{ON}$
1	1	1	$R_L<R_R$	1	$2R_{ON}$	$2R_{OFF}$

Table B.3.b Truth table of the optimized CARRY/C_o structure-1

A	B	C_i	R_L	R_R	Equivalent resistance comparison	Qm

0	0	0	$\frac{R_{OFF}}{2}+R_{AP}$	$\frac{R_{ON}}{2}+R_{P}$	$R_L > R_R$	0
0	0	1	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{AP}$	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{P}$	$R_L > R_R$	0
0	1	0	$\frac{R_{OFF}}{2}+R_{P}$	$\frac{R_{ON}}{2}+R_{AP}$	uncertain	uncertain
0	1	1	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{P}$	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{AP}$	$R_L < R_R$	1
1	0	0	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{AP}$	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{P}$	$R_L > R_R$	0
1	0	1	$\frac{R_{ON}}{2}+R_{AP}$	$\frac{R_{OFF}}{2}+R_{P}$	uncertain	uncertain
1	1	0	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{P}$	$\frac{R_{OFF}R_{ON}}{R_{OFF}+R_{ON}}+R_{AP}$	$R_L < R_R$	1
1	1	1	$\frac{R_{ON}}{2}+R_{P}$	$\frac{R_{OFF}}{2}+R_{AP}$	$R_L < R_R$	1

Table B.3.c Resistance condition of the optimized CARRY/C_o structure-1

A	B	C_i	Condition	Qm
0	1	0	$R_{AP}-R_P = TMR \times R_P < \frac{R_{OFF}-R_{ON}}{2}$	0 (correct)
			$R_{AP}-R_P = TMR \times R_P > \frac{R_{OFF}-R_{ON}}{2}$	1 (error)
1	0	1	$R_{AP}-R_P = TMR \times R_P < \frac{R_{OFF}-R_{ON}}{2}$	1 (correct)
			$R_{AP}-R_P = TMR \times R_P > \frac{R_{OFF}-R_{ON}}{2}$	0 (error)

Table B.3.d Truth table of the optimized CARRY/C_o structure-2

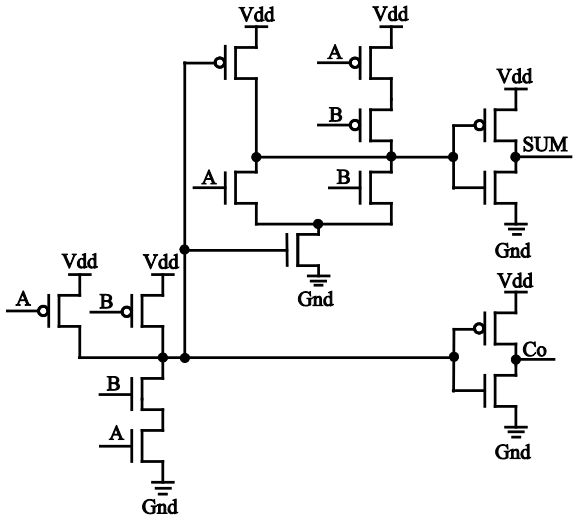
A	B	C_i	R_L	R_R	Equivalent resistance comparison	Qm
0	0	0	$2R_{OFF}+R_{AP}$	$2R_{ON}+R_P$	$R_L > R_R$	0
0	0	1	$R_{ON}+R_{OFF}+R_{AP}$	$R_{ON}+R_{OFF}+R_P$	$R_L > R_R$	0
0	1	0	$2R_{OFF}+R_P$	$2R_{ON}+R_{AP}$	uncertain	uncertain
0	1	1	$R_{ON}+R_{OFF}+R_P$	$R_{ON}+R_{OFF}+R_{AP}$	$R_L < R_R$	1

1	0	0	$R_{ON}+R_{OFF}+R_{AP}$	$R_{ON}+R_{OFF}+R_{P}$	$R_L > R_R$	0
1	0	1	$2R_{ON}+R_{AP}$	$2R_{OFF}+R_{P}$	uncertain	uncertain
1	1	0	$R_{ON}+R_{OFF}+R_{P}$	$R_{ON}+R_{OFF}+R_{AP}$	$R_L < R_R$	1
1	1	1	$2R_{ON}+R_{P}$	$2R_{OFF}+R_{AP}$	$R_L < R_R$	1

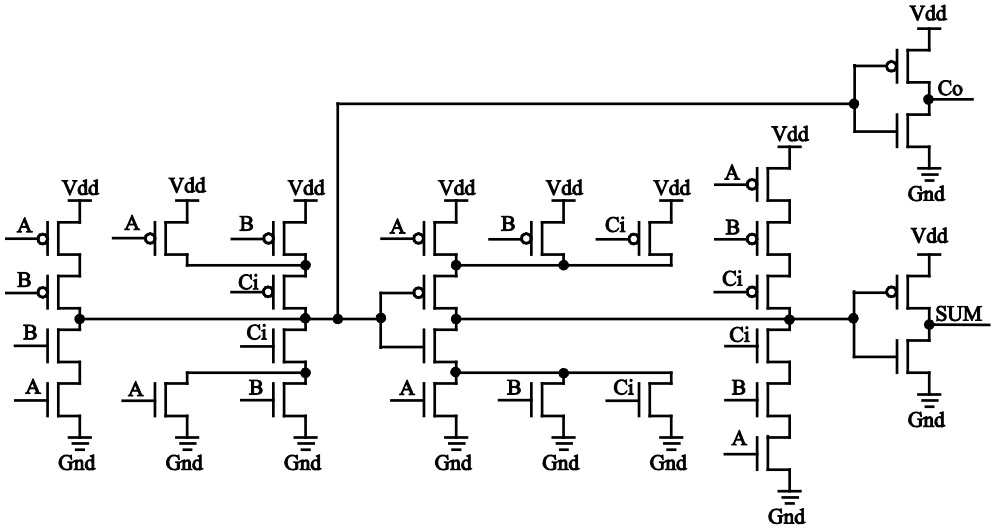
Table B.3.e Resistance condition of the optimized CARRY/C_o structure-2

A	B	C_i	Condition	Qm
0	1	0	$R_{AP}-R_P = TMR \times R_P < 2(R_{OFF}-R_{ON})$	0 (correct)
			$R_{AP}-R_P = TMR \times R_P > 2(R_{OFF}-R_{ON})$	1 (error)
1	0	1	$R_{AP}-R_P = TMR \times R_P < 2(R_{OFF}-R_{ON})$	1 (correct)
			$R_{AP}-R_P = TMR \times R_P > 2(R_{OFF}-R_{ON})$	0 (error)

Appendix C Basic addition cells used in the 8-bit NVFA (Structure-1)



(a)



(b)

(a) CMOS-only half-adder (HA) (b) CMOS-only full-adder (FA)

Appendix D Source code of the spin-Hall-assisted STT MTJ model

---

```
/*-----------------------------------------
Property: IEF, UMR8622, Univ.Paris Sud-CNRS
Authors: Zhaohao WANG, Weisheng ZHAO, Jacques-Olivier Klein and Claude
Chappert [159]
-----------------------------------------*/

`resetall
`include "constants.vams"
`include "disciplines.vams"

//MTJ Shape definition
`define rec 1
`define ellip 2
`define circle 3

/*-------------------------------------Electrical Constants-------------------------------------*/

`define e 1.6e-19    // Elementary Charge
`define m 9.11e-31    // Electron Mass
`define uB 9.274e-24    // Bohr Magnetron Constant
`define u0 1.256637e-6    // Vacuum permeability
`define hbas 1.0545e-34    // Planck's Constant
`define kB 1.38e-23    // Boltzmann Constant

module model(T1,T2,T3,Tmz);

inout T1,T2,T3;
electrical T1,T2,T3;    // Actual terminals corresponding to pinned layer and W strip

output Tmz;
electrical Tmz;    // Virtual terminal for monitoring the Magnetization

electrical Tx;    // Internal node amongst T1~T3

/*-----------------------------MTJ and W Technology Parameters-----------------------------*/

parameter real    alpha=0.03;    // Gilbert Damping Coefficient
parameter real    P=0.62;    // Electron Polarization Percentage
parameter real    eta = 0.3;    // Spin Hall angle
parameter real    Hk=8e4;    // Out of plane Magnetic Anisotropy in A/m
parameter real    Ms=9e5;    // Saturation Field in the Free Layer in A/m
parameter real    PhiBas=0.4;    // The Energy Barrier Height for MgO in electron-volt
parameter real    Vh=0.5;    // Voltage bias when the TMR(real) is 1/2TMR(0) in Volt

/*-----------------------------------MTJ Device Parameters-----------------------------------*/
```

```
parameter integer    SHAPE=2 from[1:3];    // Shape of MTJ
parameter real    tsl=0.7e-9 from[0.5e-9:3.0e-9];    // Height of the Free Layer in meter
parameter real    a=100.0e-9;    // Length in meter
parameter real    b=100.0e-9;    // Width in meter
parameter real    r=50e-9;    // Radius in meter
parameter real    tox=8.5e-10 from[8e-10:15e-10];    // Height of the Oxide Barrier in meter
parameter real    TMR=1.2;    // TMR(0) with Zero Volt Bias

/*-------------------------------------W strip Parameters-----------------------------------------*/

parameter real    d=3e-9;    // Thickness in meter
parameter real    l=120e-9;    // Length in meter
parameter real    w=100e-9;    // Width in meter

/*-------------------------------State Parameters of MTJ and W-------------------------------*/

parameter integer    PAP=1 from[0:1];    // Initial state of the MTJ, 0 = parallel, 1 =
                                            anti-parallel
parameter real      T= 300;    // Room temperature in Kelvin
parameter real      RA=10e-12 from[5e-12:15e-12];    // Resistance area product of MTJ in
                                                        ohm-m2
parameter real    rho = 2e-6;    // Resistivity of W in ohm-m
parameter real    sim_step = 1e-12;    // Simulation time step in second

/*-----------------------------------------Variables-----------------------------------------*/

real    FA;      //Coefficient used in Brinkman model
real    gamma;      //GyroMagnetic Ratio
real    surface;        //MTJ surface area
real    V12,V13,V23;    //Voltages across two terminals
real    Rp;    //MTJ Resistance when the relative magnetization is parallel
real    R_MTJ;      //Real resistance of MTJ
real    R_W;    //Resistance of W
real    theta,phi;    //Angle of magnetization
real    delta_phi,delta_theta;        //Change of angle
real    delta_aver;    //Root square average value of theta deviation
real    V_MTJ;    //Voltage across the MTJ from top layer to bottom layer
real    ksi;    //Coefficient used in LLG equation
real    J_STT;        //Current density for STT
real    J_SHE;        //Current density for SHE
real    mz;    //Magnetization in z direction
real    E_thermal;    //Thermal stability energy
real    t_previous;    //Recording the simulation time

analog begin

/*-----------------------------------------initial conditions-----------------------------------------*/

@(initial_step)begin
     if (SHAPE==1)
          surface=a*b;    //SQUARE
```

```
    else if (SHAPE==2)
        surface=`M_PI*a*b/4.0;    //ELLIPSE
    else begin
        surface=`M_PI*r*r;    //ROUND
    end

    gamma = 2*`uB/`hbas;    //GyroMagnetic Ratio
    ksi = gamma*`hbas/(2*`e*tsl*Ms);    //Coefficient used in LLG equation
    E_thermal = 0.5*`u0*Ms*Hk*tsl*surface;    //Thermal stability energy
    delta_aver = sqrt(2.0*`kB*T/E_thermal);      //Root square average value of theta
                                                   deviation

    //MTJ resistance under zero bias
    FA=3.3141e-7/RA;
    Rp=(tox/(FA*sqrt(PhiBas)*surface))*exp(2*sqrt(2*`m*`e*PhiBas)*tox/`hbas);

    R_W = rho*l/(d*w); //W strip resistance

    //Initial angle and mz
    phi = `M_PI;
    if (PAP==0) begin
        theta = delta_aver;
        end
    else begin
        theta = `M_PI-delta_aver;
    end
    mz = cos(theta);
    $display("PAP=%d,mz=%g,theta=%g,Rp=%g,R_W=%g",PAP,mz,theta,Rp,R_W);
    t_previous = $realtime;

end

/*----------------------------------------------Simulation-----------------------------------------*/

//Calculation of MTJ resistance
V(Tx) <+ (0.5*V(T1)*R_W+(Rp*(1+(V(T1)-V(Tx))*(V(T1)-V(Tx))/(Vh*Vh)+TMR)/
        (1+(V(T1)-V(Tx))*(V(T1)-V(Tx))/(Vh*Vh)+0.5*(1+mz)*TMR))*(V(T2)+V(T3)))/
        (2*(Rp*(1+(V(T1)-V(Tx))*(V(T1)-V(Tx))/(Vh*Vh)+TMR)/(1+(V(T1)-V(Tx))*
        (V(T1)-V(Tx))/(Vh*Vh)+0.5*(1+mz)*TMR))+0.5*R_W);
R_MTJ = Rp*(1+(V(T1)-V(Tx))*(V(T1)-V(Tx))/(Vh*Vh)+TMR)/(1+(V(T1)-V(Tx))*
        (V(T1)-V(Tx))/(Vh*Vh)+0.5*(1+mz)*TMR);

//Calculation of current
I(T1,Tx) <+ ((V(T1)-V(Tx))/R_MTJ);
I(T2,Tx) <+ (2*(V(T2)-V(Tx))/R_W);
I(T3,Tx) <+ (2*(V(T3)-V(Tx))/R_W);

//Calculation of STT and SHE current density
J_STT = -I(T1,Tx)/surface;
if (V(T2)>V(T3))
    J_SHE = min(abs(I(T2,Tx)),abs(I(Tx,T3)))/(w*d);
else if (V(T2)<V(T3))
```

```
        J_SHE = -min(abs(I(T3,Tx)),abs(I(Tx,T2)))/(w*d);
    else begin
        J_SHE = 0;
    end

    //LandauLifshitzGilbert equation including STT torque and SHE torque
    if (analysis("tran")) begin
        delta_phi = ($realtime-t_previous)*(1.0/(1+alpha*alpha))*(gamma*`u0*Hk*
                    cos(theta)-alpha*ksi*P*J_STT-ksi*eta*J_SHE*(alpha*cos(theta)*
                    cos(phi)-sin(phi))/sin(theta));
        delta_theta = ($realtime-t_previous)*(1.0/(1+alpha*alpha))*(-alpha*gamma*`u0*
                     Hk*cos(theta)*sin(theta) - ksi*P*J_STT*sin(theta)-ksi*eta*J_SHE*
                     (alpha*sin(phi)+cos(theta)*cos(phi)));
        phi = phi + delta_phi;
        theta = theta + delta_theta;
        t_previous = $realtime;
        $bound_step(sim_step);
    end

    //Limit the theta
    if(theta > `M_PI-delta_aver)
        theta = `M_PI-delta_aver;
    else if (theta < delta_aver)
        theta = delta_aver;

    //Output mz
    mz = cos(theta);
    V(Tmz)<+mz;

    end
    endmodule
```

Appendix E Résumé en français

Chapitre 1 Etat de l'art

La spintronique vise à exploiter la propriété du spin des électrons et à créer de nouveaux dispositifs. L'origine de la spintronique remonte aux années 1970. Mais il n'a pas été utilisé dans des applications pratiques en raison des limites de la technologie et de l'équipement jusqu'à la découverte de l'effet de magnétorésistance géante (MRG) par Albert Fert en 1988 et Peter Grünberg en 1989. Le premier capteur commercial basé sur MRG a été annoncé en 1994. Aujourd'hui, le capteur basé sur MRG est utilisé dans le stockage de données, les applications biologiques, les applications spatiales, etc.

L'effet de magnétorésistance tunnel (MRT) a été observé par Jullière dans la jonction Fe/Ge/Co en 1975. Une jonction tunnel magnétique (JTM) est principalement constituée d'une fine couche isolante et deux couches ferromagnétiques (FM) (voir Figure R.1(a)). Une couche ferromagnétique (couche de référence) est à aimantation fixée, tandis que l'autre couche ferromagnétique (couche de stockage) est à aimantation libre. En raison de l'effet MRT, JTM peut présenter deux états, parallèle (P) et antiparallèle (AP), correspondant à une résistance faible et élevée en modifiant l'orientation relative de magnétisation de deux couches ferromagnétiques (voir Figure R.1 (c)). JTM peut être commuté entre deux états par des champs magnétiques externes ou un courant d'injection.

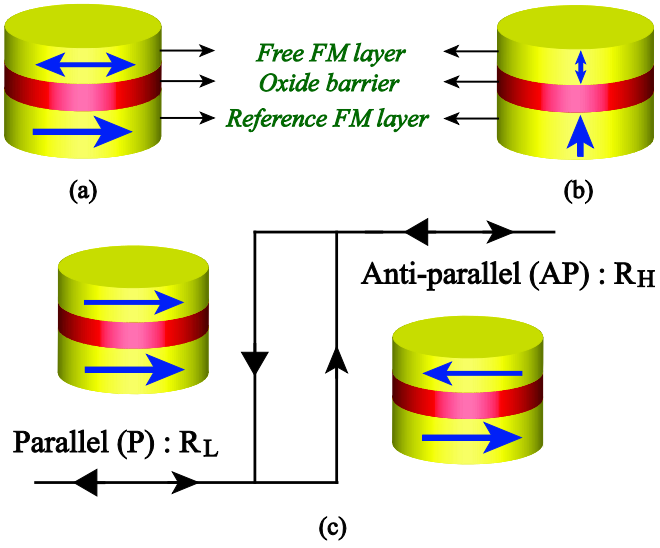


Figure R.1 Structure de la jonction tunnel magnétique (JTM) et l'effet de magnétorésistance tunnel (MRT)

Le ratio de MRT, qui caractérise l'amplitude du changement de résistance, est défini par Eq. R.1. C'est montré que 604% du ratio de MRT a été observé avec la jonction CoFe/MgO/CoFe.

$$TMR = \frac{\Delta R}{R_P} = \frac{R_{AP} - R_P}{R_P} = \frac{\Delta G}{G_{AP}} = \frac{G_P - G_{AP}}{G_{AP}} \qquad \text{Eq. R.1}$$

Le mécanisme de commutation induit par le champ magnétique est une approche d'écriture pour la première génération de MRAM. Cette approche souffre de deux problèmes principaux. Tout d'abord, le courant nécessaire à la commutation est trop élevé (~ mA), ce qui entraîne une consommation élevée, une faible densité et une faible extensibilité. En outre, la perturbation à demi-sélectivité est un autre inconvénient. La commutation assistée par thermique a été proposée pour surmonter les problèmes. Néanmoins, JTM doit être refroidi après une opération de commutation avec une durée de refroidissement relativement longue (~ ns). Pour cette raison, Cette approche ne peut pas satisfaire aux exigences de haute vitesse pour les applications de logique ou de registre.

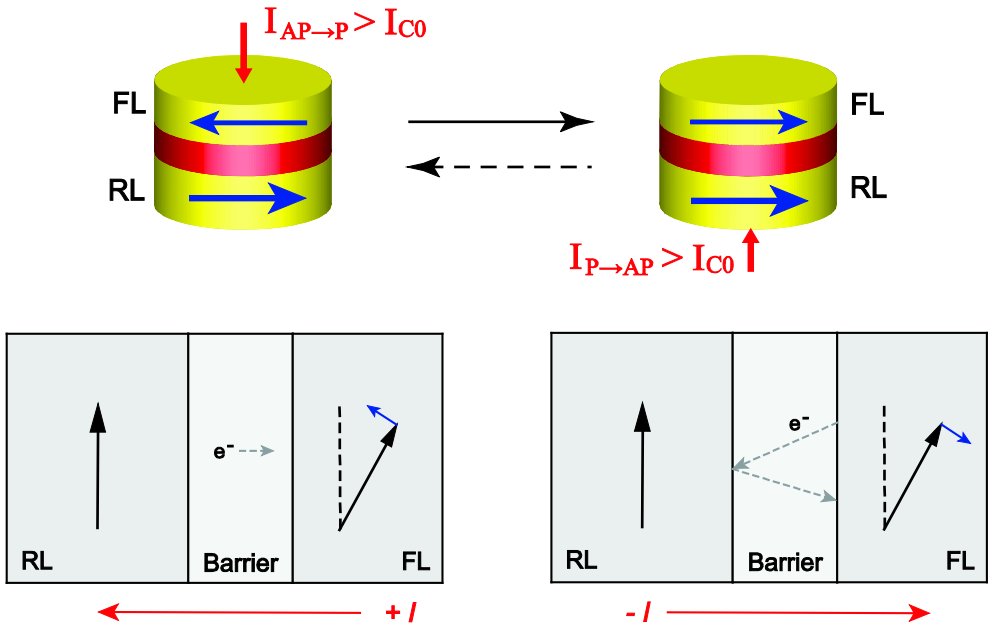


Figure R.2 Commutation de couple de transfert de spin (CTS)

Une autre approche est basée sur le couple de transfert de spin (CTS). La commutation de la magnétisation de la couche de stockage dans deux cas est illustrée à Figure R.2, en supposant que le courant injecté de couche de stockage est positif. Par rapport aux autres approches de commutation, CTS n'a besoin que d'un courant bidirectionnel et la densité de courant soit faible (10^6~10^7 A/cm^2). Ce mécanisme d'écriture simplifie le circuit d'écriture, tout en conservant une puissance plus faible et une densité plus élevée. JTM avec anisotropie magnétique perpendiculaire (p-JTM) (voir Figure R.1(b)) a un courant critique inférieur et une

vitesse de commutation supérieure à celle d'i-MTJ lors de la même stabilité thermique.

L'effet de Spin Hall (ESH) est une autre façon pour commuter la magnétisation de la couche de stockage par un courant d'injection. Un dispositif magnétique à trois électrodes a été proposé, où une bande de métal lourd (par exemple, Ta, Pt) est placé sous la couche de stockage. Lorsqu'un courant traverse le métal, les électrons avec différentes directions de spin sont dispersés dans des directions opposées. Le couplage spin-orbital convertit le courant de charge en courant de spin perpendiculaire, générant un couple appelé couple de spin-orbite (CSO) pour faciliter l'inversion de la magnétisation.

MRAM est une des applications de spintronique très importantes. Par rapport aux autres technologies de mémoire, MRAM combine les caractéristiques de non-volatilité, d'endurance illimitée pour lecture/écriture (> 10^{15} cycles), de temps rapide (<10 ns) etc.

Les systèmes d'aujourd'hui sont principalement construits sur l'architecture de John von Neumann. La partie logique et la mémoire sont des fonctions distinctes, et elles sont reliées par des interconnexions complexes avec une distance de transfert relativement longue (voir in Figure R.3(a)). Cela entraîne généralement un délai de transfert prolongé (ou une faible vitesse d'opération) et une dissipation de puissance de transfert élevée (~ 1pJ/bit/mm). En outre, comme les mémoires (par exemple, SRAM) sont volatiles, elles ont toujours besoin d'énergie pour conserver les données. En effet, les courants de fuite de sous-seuil et de grille augmentent, et les problèmes de puissance élevée sont devenus les principaux inconvénients des circuits logiques CMOS. Pour cette raison, la réduction de la puissance statique et dynamique ainsi que le délai d'interconnexion deviennent deux objectifs majeurs pour le système de prochaine génération.

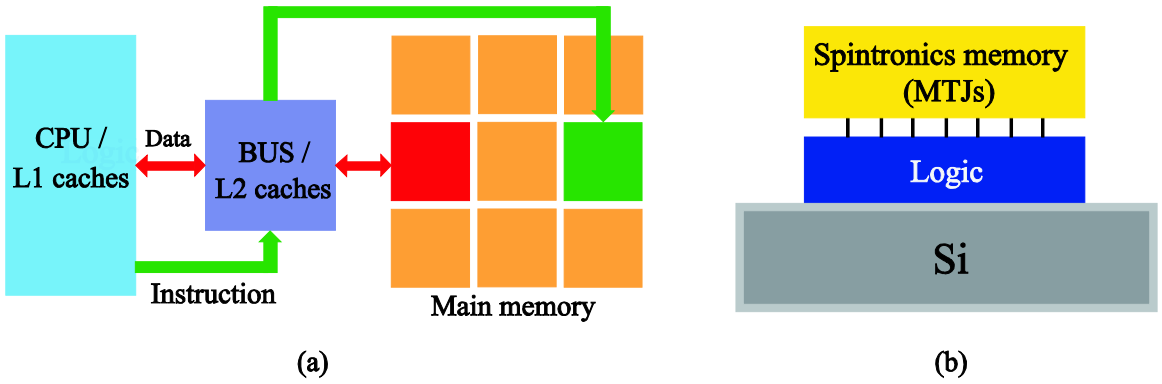


Figure R.3 (a) Diagramme de l'architecture classique de Von Neumann. Les mémoires et les unités logiques sont séparées et connectées par bus et mémoires cache (b) structure 3-D

Figure R.3(b) présente le circuit logique basé sur l'architecture logique-en-mémoire (LEM).

Afin de profiter pleinement des circuits logiques non-volatils, les dispositifs spintroniques devraient combiner les caractéristiques de la vitesse élevée, de l'endurance illimitée, de la petite taille et de la compatibilité avec la technologie CMOS. JTM est adapté pour être intégré à des circuits logiques et de mémoire conventionnels. Récemment, des circuits innovants basés sur des circuits hybrides JTM/CMOS ont été présentés. Par exemple, la table de recherche magnétique et le flip-flop magnétique ont été introduites pour le circuit logique programmable.

Chapitre 2 Conception de circuit hybrid JTM/CMOS

Afin de concevoir de nouveaux circuits mémoires et logiques basés sur la technologie hybride JTM/CMOS, nous utilisons un modèle compact développé par le groupe NANOARCHI à l'IEF (Institut d'Electronique Fondamentale). Figure R.4 montre le symbole JTM. L'électrode virtuelle *State* est utilisée pour identifier la configuration de magnétisation de JTM. La sortie de cette électrode (V_{state}) sera de 0 V (ou 1 V) si le JTM est en état parallèle (ou état antiparallèle).

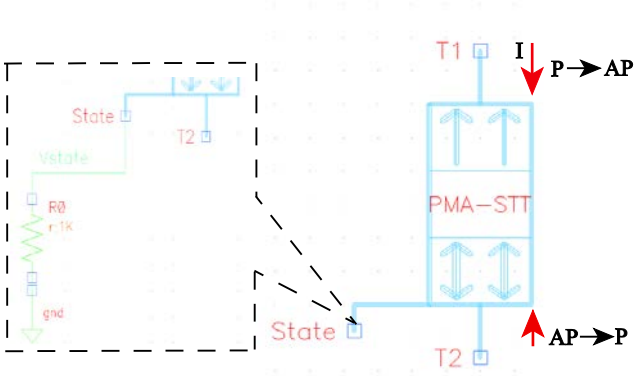


Figure R.4 Symbol du modèle JTM

Les JTMs peuvent être intégrés dans un amplificateur en mode de courant qui détecte ses configurations magnétiques. Il a été confirmé que l'amplificateur de détection de pré-charge (ADPC) effectue la meilleure vitesse de détection, consommation d'énergie, surface et fiabilité, en comparaison avec d'autres amplificateurs de détection. Figure R.5 montre le schématique du circuit de lecture basé sur l'ADPC. Deux JTMs à états complémentaires sont placés dans deux branches et stockent de donnée binaire. Les données stockées dans les JTMs sont détectées et amplifiées aux sorties Qm et $\overline{Qm}$.

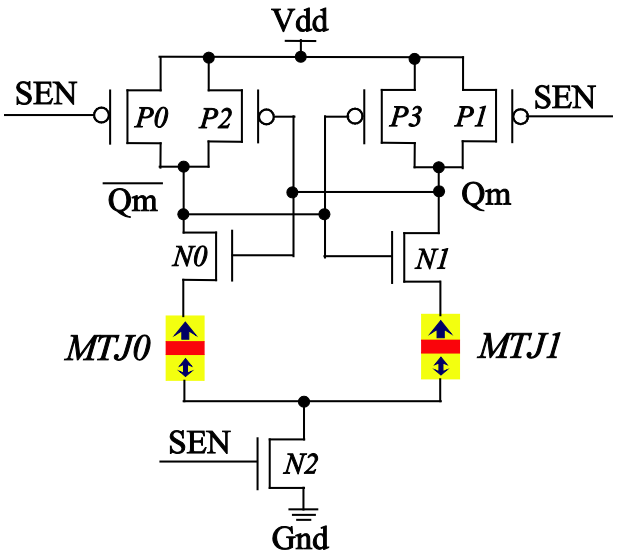
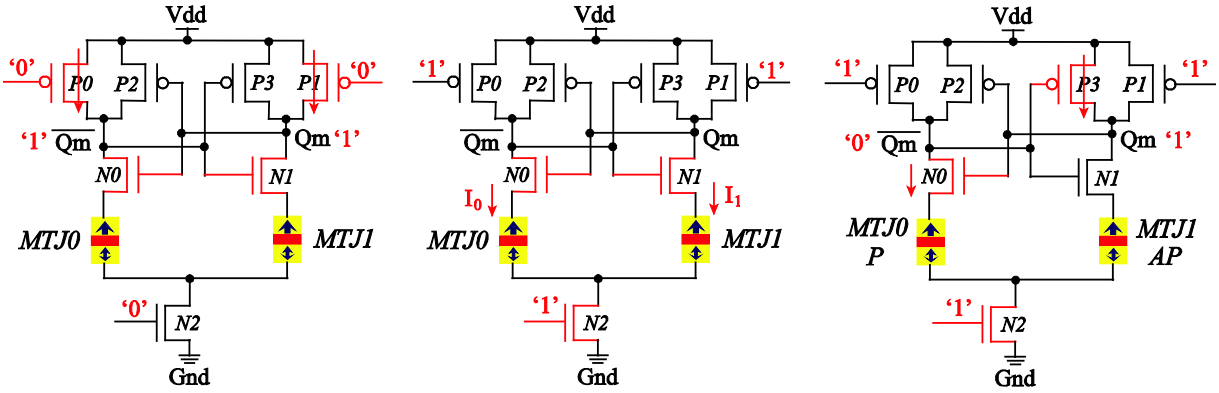


Figure R.5 Schéma de l'amplificateur de détection de pré-charge (ADPC)

Selon la valeur du signal de commande SEN, le circuit ADPC fonctionne en deux phases: phase de pré-charge et phase d'évaluation. Pendant la phase de pré-charge ($SEN = '0'$), les deux nœuds Qm et $\overline{Qm}$ sont tirés jusqu'à Vdd (voir Figure R.6 (a)). Pendant la phase d'évaluation ($SEN = '1'$), N_2 est activé, ce qui permet au courant de lecture de passer les deux JTMs (voir Figure R.6(b)). Nous supposons que, MTJ_0 et MTJ_1 sont respectivement initialisés à l'état parallèle et antiparallèle, Qm seront tirés jusqu'à V_{dd} ou logique '1', alors que $\overline{Qm}$ continuera à se décharger (voir Figure R.6(c)).



(a) (b) (c)

Figure R.6 Trois états pour l'opération de détection d'ADPC

Les résultats de simulation montrent que le circuit de détection basé sur ADPC a un délai de lecture inférieur à 200 ps et la dissipation d'énergie aussi faible (~2 fJ) avec une fréquence de 500 MHz. Le temps de lecture peut être encore réduit en augmentant la taille des transistors

ou le ratio MRT de JTM. En outre, grâce à la détection dynamique et aux petits courants traversant les JTMs, qui sont beaucoup plus bas que le courant critique de commutation (~50 μA), une écriture erronée pendant l'opération de détection peut être évitée. La fiabilité du circuit de lecture peut être améliorée en augmentant la valeur du ratio MRT ou la largeur des transistors CMOS.

Afin de réaliser l'opération d'écriture, un courant d'écriture bidirectionnel doit être généré par un circuit CMOS. Le circuit d'écriture à quatre transistors (4T) est illustré à Figure R.7(a). Pendant l'opération d'écriture, un seul transistor PMOS et un transistor NMOS seront ouverts, générant un courant d'écriture. Le circuit d'écriture à six transistors (6T) est illustré à la Figure R.7(b). Lors de l'exécution de l'opération d'écriture, P_{0-1} et N_2 sont activés alors que N_{0-1} et P_2 sont désactivés, ou vice versa. Les tensions V_{0-3} sont générées par deux signaux et les portes logiques (voir Figure R.7(c)). *WE* est le signal d'activation et *Din* détermine la direction du courant d'écriture. Le temps d'écriture peut être réduit en augmentant la tension d'alimentation ou la taille des transistors.

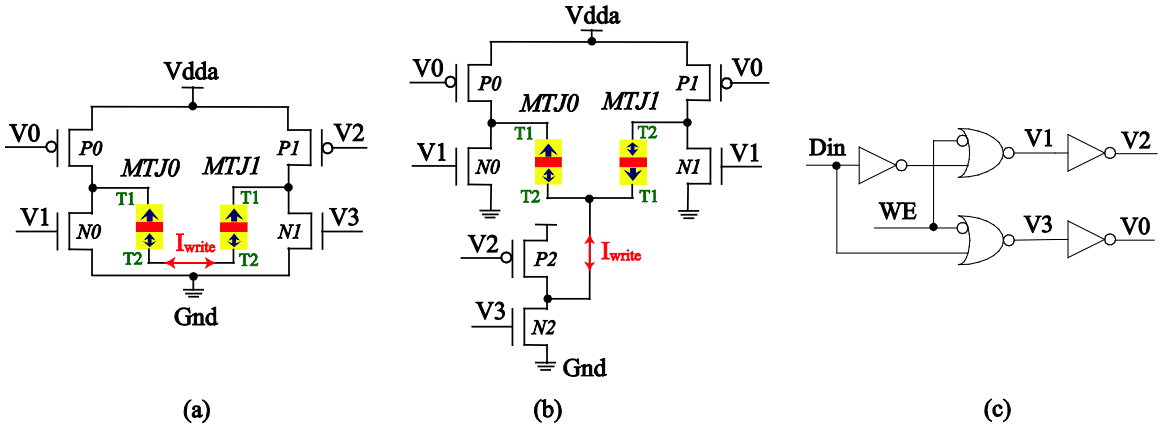


Figure R.7 (a) 4T circuit d'écriture (b) 6T circuit d'écriture (c) Portes logiques pour contrôler l'activation et la direction du courant d'écriture

En combinant le circuit de lecture et le circuit d'écriture, le circuit hybride JTM/CMOS peuvent être conçus (voir Figure R.8). Afin de réaliser l'opération d'écriture sans perturber les sorties, deux transistors de séparation contribuent à isoler les JTMs de la partie de détection et empêchent ainsi le courant d'écriture de traverser la partie de détection.

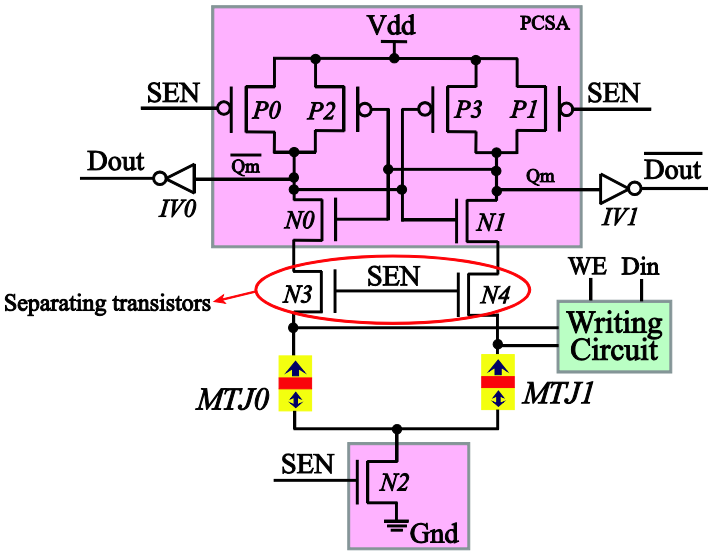
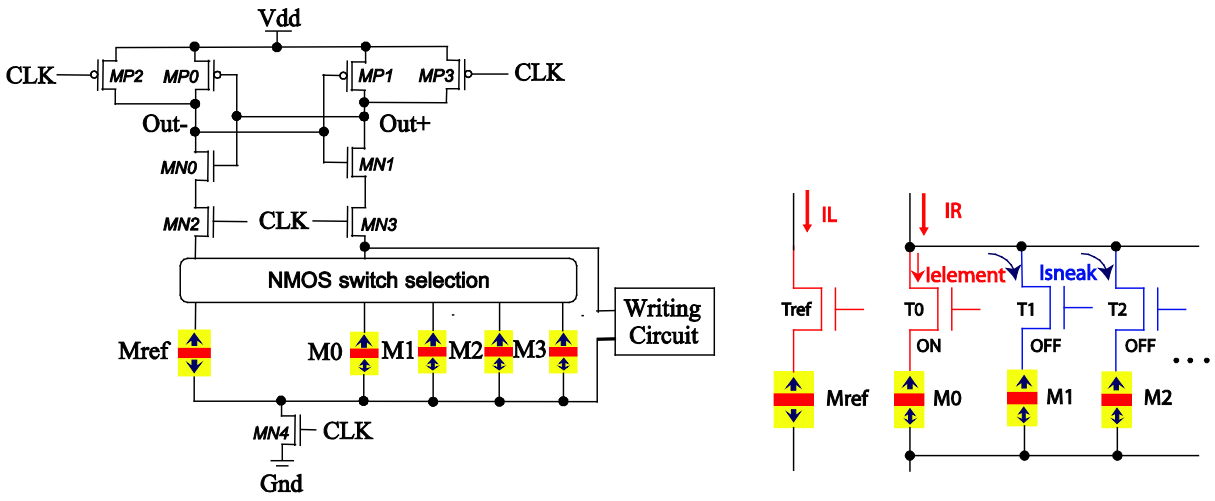


Figure R.8 Schéma complet du circuit de lecture/écriture

L'architecture logique hybride à plusieurs contextes (ou multi-bit), qui comporte de multiples bits non-volatiles pour la commutation rapide entre les contextes, fournit une propriété supplémentaire en fonction de la surface en raison de l'intégration des JTMs au-dessus des circuits logiques CMOS. En outre, la sécurité des données peut être améliorée.

Figure R.9(a) montre la structure multi-bit intégrant quatre contextes. Cette structure utilise une JTM de référence (M_{ref}) pour détecter les données non-volatiles stockées dans les JTMs de stockage (M_{0-3}). La résistance de la JTMs de référence (R_{ref}) devrait être entre R_P et R_{AP} . La commutation entre quatre contextes peut être réalisée en configurant un décodeur 2-4, ce qui permet d'ouvrir un seul transistor de sélection tandis que les trois autres transistors sont fermés.

La structure asym-ADPC fait face à plusieurs problèmes critiques de fiabilité: 1) Pendant l'opération de lecture, à l'exception du courant ($I_{element}$) traversant la sous-branche adressée (par exemple, M_0), des courants parasites (I_{sneak}) traversant les sous-branches fermées (M_{1-3}) ne sont pas négligeables en raison de capacités parasites (voir Figure R.9(b)). 2) Les variations de processus croissantes entraînent une déviation significative des paramètres du transistor et de JTM, menant à un grand décalage du circuit de détection.



(a) (b)

Figure R.9 (a) Schéma de structure asymétrique multi-bit (asym-ADPC) (b) problème des courants parasites

Pour surmonter le problème de la détection asymétrique et atténuer l'influence des courants parasites, nous proposons une structure de détection symétrique. Il existe M JTMs de stockage (par exemple, M=2 dans Figure R.10) et une JTM de référence de chaque côté. Cette conception permet d'atténuer radicalement la perturbation des courants parasites.

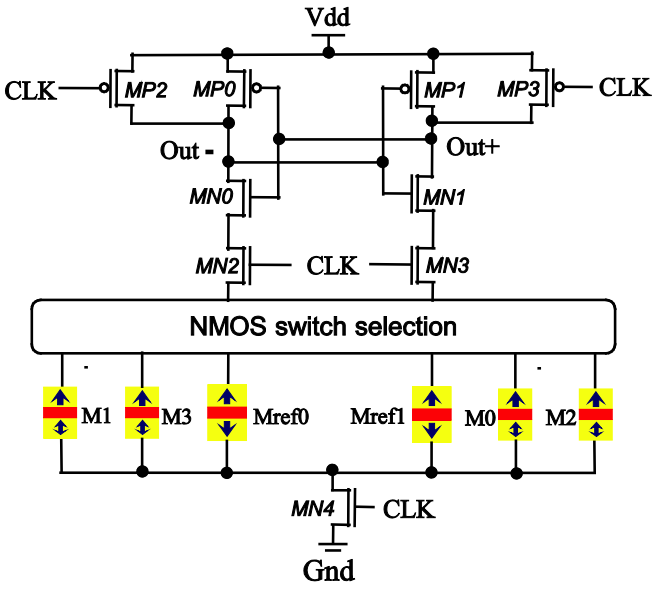


Figure R.10 Schéma de la structure symétrique multi-bit basée sur ADPC (sym-ADPC)

Pour surmonter les problèmes liés au processus technologique, nous proposons une nouvelle structure (voir Figure R.11). L'amplificateur de détection de pré-charge séparé (ADPCS) a deux queues de décharge pour séparer la phase de décharge de la phase d'évaluation. Pendant la phase de décharge ($CLK = CLKP = '1'$), les deux nœuds ($A+$ et $A-$) commencent à se

décharger mais avec un taux de temps différent. En conséquence, une tension différentielle (ΔA) entre $A+$ et $A-$ est créée, ce qui génère une tension différentielle aux nœuds B. Pendant la phase d'évaluation, MN_2 et MN_3 maintient ON, ce qui permet aux nœuds de sortie de se décharger. Les ondes simulées, où les différences de tension ΔA et ΔB sont clairement illustrées, sont représentées sur Figure R.12.

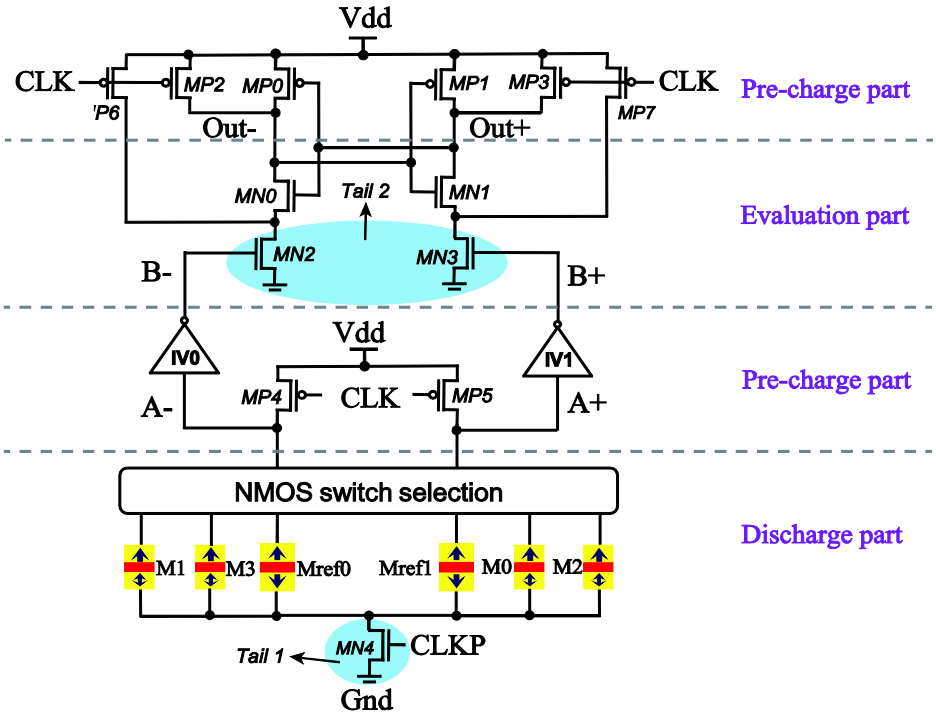


Figure R.11 Schéma de la structure symétrique multi-bit basée sur un amplificateur de détection de pré-charge séparé (sym-ADPCS)

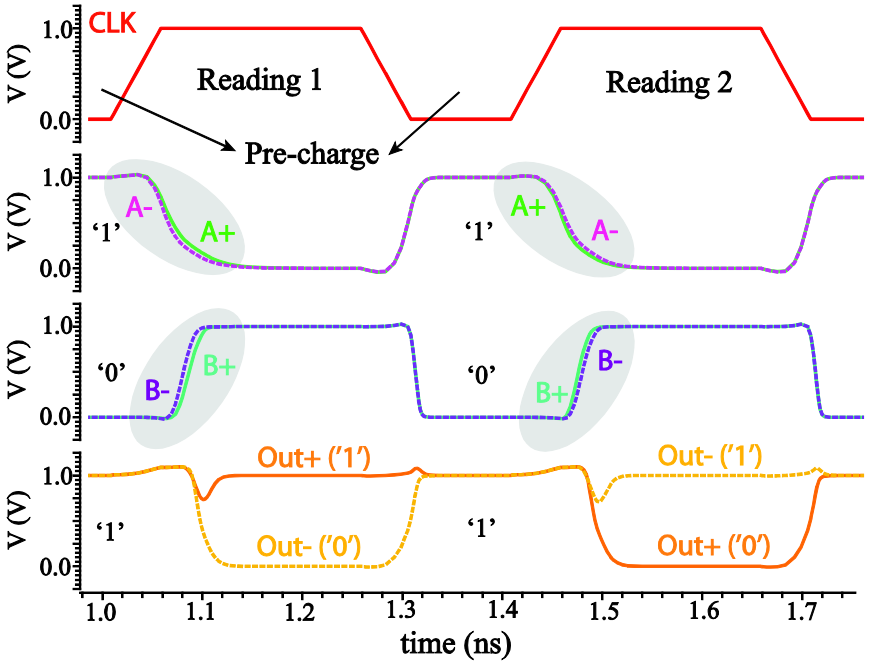


Figure R.12 Simulation du circuit sym-ADPCS

Les trois structures peuvent fonctionner à haute fréquence car elles maintiennent un délai de propagation inférieur à 200 ps, grâce à l'approche de détection dynamique rapide. En outre, ils produisent une faible puissance de détection, qui atteint un niveau presque négligeable (~ fJ). La structure asymétrique (asym-ADPC) présente une faible évolutivité et au maximum cinq JTMs peuvent être intégrés, tandis que les structures symétriques (sym-ADPC et sym-ADPCS) présentent de bonnes perspectives pour intégrer un grand nombre de JTM, par exemple 32 JTMs. La structure sym-ADPCS a presque la moitié moins d'erreurs et 14.2% plus petit temps de détection par rapport à la structure asym-ADPC, avec tous les transistors restant dans la taille minimale (voir Table R.1). Il présente la meilleure fiabilité et la meilleure vitesse de détection. Cependant, son énergie de lecture est presque quatre fois plus grande que les structures basées sur ADPC (asym-ADPC et sym-ADPC) en raison de ses deux chemins de courant.

Table R.1 Comparaison de trois structures multi-bit

Performances		**asym-ADPC**	**sym-ADPC**	**sym-ADPCS**
Temps (ps/bit)		160	162.7	139.6
Énergie (fJ/bit)		1.21	1.24	5.32
Taille		14T	15T	23T
Limitation du nombre de JTM		< 6 JTMs	> 30 JTMs	> 30 JTMs
Taux d'erreur (TD)	JTM_AP	30.4%	29.8%	15%
	JTM_P	32.2%	34.6%	19.5%
	TD moyen	31.3%	32.2%	17.25%

Chapitre 3 Conception de circuits logiques non-volatiles

L'architecture générale de logique-en-mémoire (LEM) se compose principalement de trois parties: 1) un amplificateur de détection (AD) pour détecter les courants de deux branches et pour évaluer le résultat logique sur les sorties, 2) un bloc d'écriture pour programmer les données stockées dans des cellules de mémoire non-volatiles, 3) un réseau logique (RL) qui effectue le calcul (voir Figure R.13). RL contient des JTMs qui conservent les entrées non-volatiles et un arbre logique CMOS pour des entrées volatiles. JTM est utilisée non seulement comme élément de stockage mais aussi comme opérande.

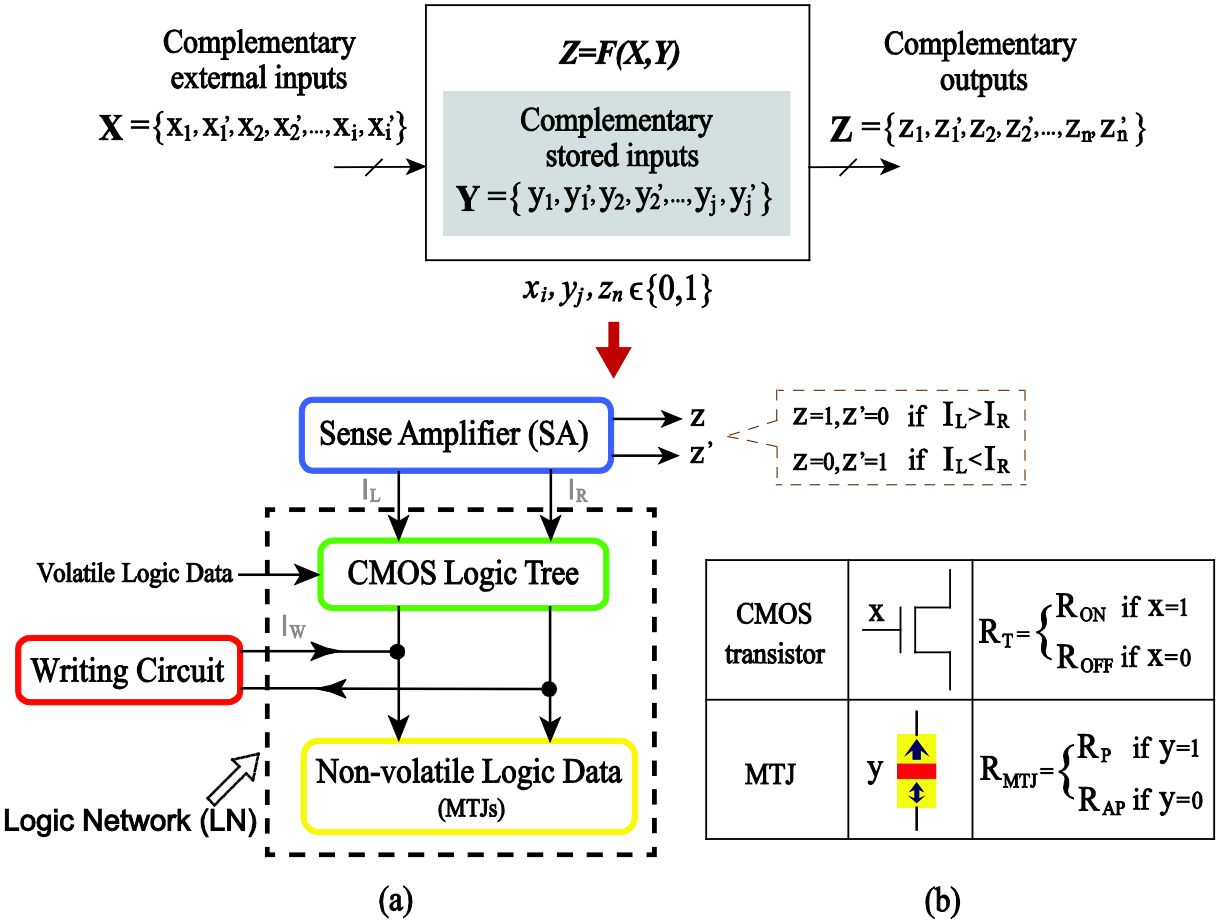


Figure R.13 (a) Schéma de l'architecture logique-en-mémoire (LEM) (b) Composants dans le réseau logique (RL)

En configurant le RL, différentes fonctions logiques peuvent être réalisées telles que la porte ET, OU et XOR, comme montré dans Figure R.14. Les JTMs sont toujours dans des états opposés pour assurer la vitesse de détection élevée et ils sont connectés en série avec un point central commun.

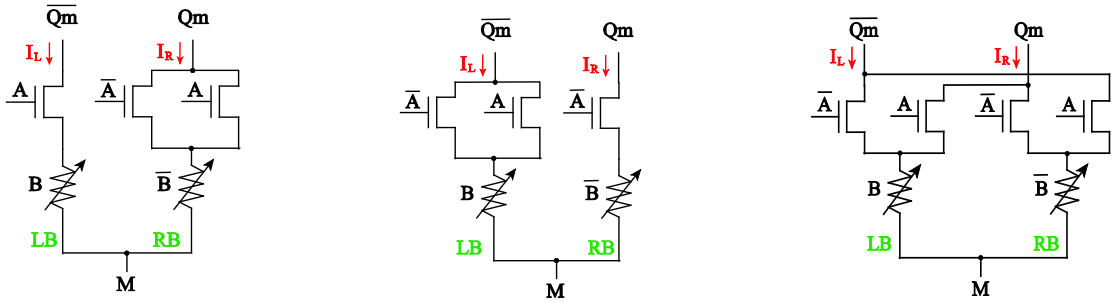
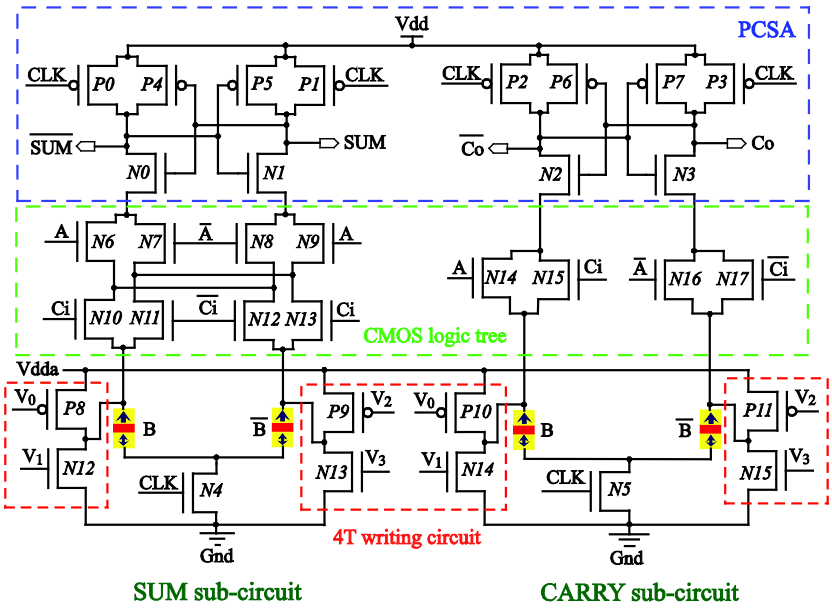


Figure R.14 Structure générale du réseau logique (RL) pour (a) la porte ET non-volatile (b) la porte OU non-volatile (c) la porte XOR non-volatile

Figure R.15 montre le circuit complet d'additionneur complet non-volatile (ACNV) 1-bit en combinant le sous-circuit SUM et le sous-circuit CARRY. Nous comparons l'ACNV avec

l’AC basé sur la technologie CMOS (voir Table R.2). Grâce à l'intégration 3-D de JTM, le surface de cette conception est avantageuse par rapport à celle de l’AC basé sur CMOS. L'énergie de transfert de données devient beaucoup plus faible grâce à la distance plus courte entre la mémoire et l’unité logique.



Truth table of SUM and CARRY logic

A	B	Ci	SUM	Co
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

Figure R.15 Schéma complet de l'additionneur complet non-volatile (ACNV) 1-bit

Table R.2 Comparaison d’ACNV 1-bit avec AC basé sur CMOS

Performances	**CMOS AC**	**ACNV**
Temps de détection	75 ps	87.4 ps
Puissance dynamique (@500MHz)	2.17 µW	1.98 µW
Puissance standby	~ nW	~ 0 [144]
Energie de transfert de données	> pJ/bit	< fJ/bit
Surface	46T	38T + 4 JTM

Pour étendre l’ACNV 1-bit à la structure multi-bit et à réaliser une non-volatilité complète, trois ACNVs 8-bit sont proposés. Les schémas structurels complets ainsi que les distributions de données non-volatiles sont illustrés à Figure R.16. L'architecture d’ACNV 8-bit est composée d'un demi-additionneur (DA) et de sept ACs en connexion en série, exécutant une opération d'addition de deux mots de 8 bits. Il convient de noter que la première structure (Structure-1) est conçue en fonction de DA et AC basé sur CMOS, tandis que les autres structures utilisent les DAs et les ACs non-volatiles pour effectuer une opération d'addition.

Par rapport à Structure-1 et Structure-2, Structure-3 montre des avantages en termes de surface car elle présente moins de flip-flops que d'autres structures. Cet avantage devient plus

significatif avec l'augmentation de bit puisque plus d'additionneurs non-volatiles peuvent partager le même décodeur. Structure-2 et Structure-3 consomment respectivement 16.1% et 34.1% moins d'énergie dynamique que Structure-1 (voir Table R.3).

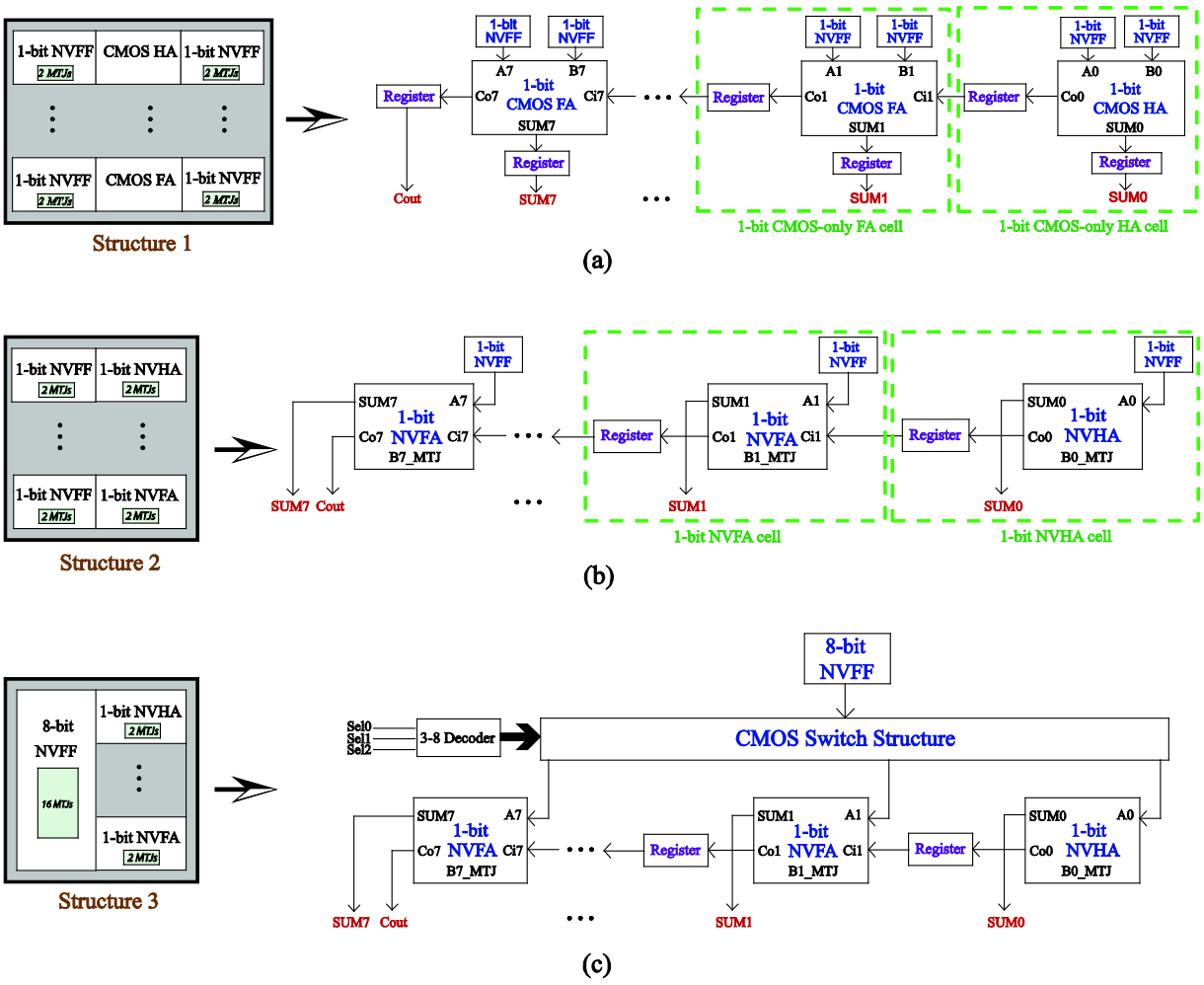


Figure R.16 Les distributions locales des données non-volatiles et des schémas complets des structures ACNV 8-bit (a) Structure-1: A et B sont stockées dans des flip-flops non-volatiles (b) Structure-2: B sont stockées dans des JTMs intégrés alors que A sont stockées dans 8 flip-flops non-volatiles (c) Structure 3: A sont stockées dans un flip-flop non-volatile de 8 bits

Table R.3 Comparaison de différents l'additionneur complet de 8 bits

Paramètre	**Surface (μm^2)**	**Temps (ns)**	**Energie dynamique (pJ/8 bits)**
Structure-1	218.74	0.14	1.039
Structure-2	219.46	0.15	0.8718
Structure-3	194.96	0.18	0.6845

La cellule de mémoire proposée (2T/2JTM) est composée de deux JTMs, un transistor NMOS

et un transistor PMOS connecté en série (voir Figure R.17). M_0 et M_1 ont la même configuration, sauf qu'ils sont dans les états complémentaires, c'est-à-dire qu'une JTM a une résistance élevée tandis qu'une autre a une faible résistance. V_M dépend des caractéristiques de deux JTMs. Pour lire la donnée de stockage d'un bit, une tension d'alimentation est appliquée, générant un courant de lecture statique I_S. V_M est soit élevé lorsque la résistance de M_0 (R_0) est inférieure à celle de M_1 (R_1), ou faible quand R_0 est supérieure à R_1.

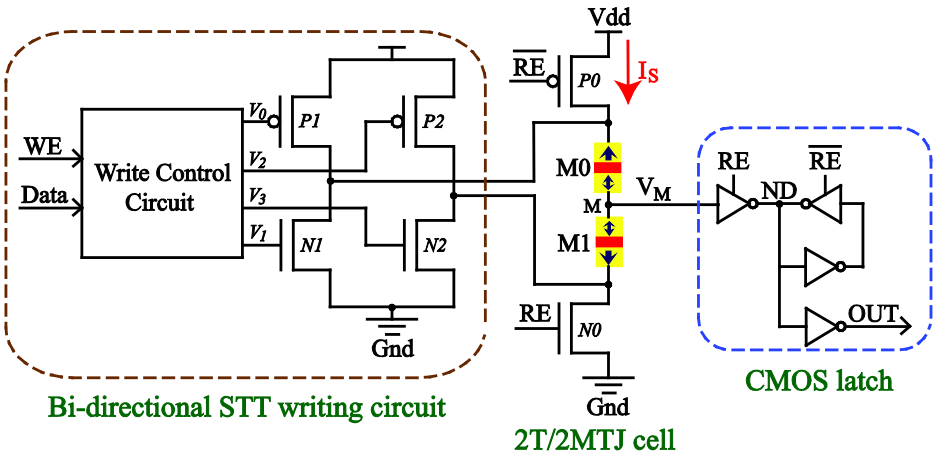


Figure R.17 Circuit de détection de mode de tension (CDMT)

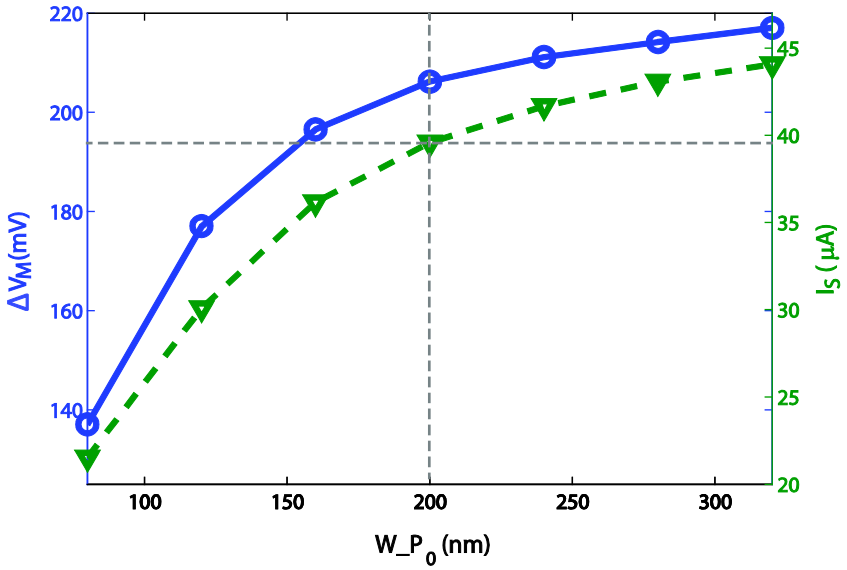


Figure R.18 Marge de détection et courant de détection de la cellule 2T/2JTM par rapport à la largeur de P_0

Figure R.18 démontre l'influence de la largeur du transistor PMOS sur le courant de détection statique et la marge de détection. Il montre que le plus grand W_P_0 conduit à un plus grand ΔV_M, ce qui est avantageux pour une détection fiable. Cependant, la résistance du transistor

devient plus petite et I_S se rapproche du courant d'écriture critique de JTM (~50 μA). Une écriture involontaire peut se produire pendant l'opération de lecture en raison des variations de processus. Cela entraîne également une forte énergie. Afin de résoudre ces problèmes, nous proposons le circuit optimisé avec un circuit de contrôle (voir Figure R.19). Pendant l'opération de lecture, une fois que les sorties SUM (ou C_o) et $\overline{SUM}$ (ou $\overline{C_o}$) sont différentes, les transistors seront fermés, puis l'opération de détection est désactivée.

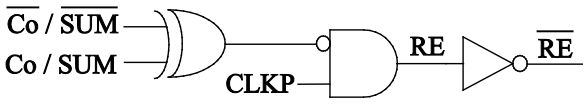


Figure R.19 Circuit de contrôle

Les ACNVs précédemment proposés sont principalement basés sur les JTMs commutés par CTS. Bien qu'ils présentent des avantages dans la vitesse de lecture et l'énergie de lecture, ils souffrent d'une faible vitesse d'écriture et d'une forte dissipation de la puissance d'écriture car la commutation CTS nécessite un grand temps d'incubation lors du processus initial. La commutation CTS assistée par Spin-Hall a été proposée pour obtenir une opération d'écriture à grande vitesse. L'ACNV basé sur JTM commuté par CTS assistée par spin-Hall (CTS+ESH ACNV) est illustré dans Figure R.20. Le circuit de lecture (Partie 1) du CTS+ESH ACNV est identique à ACNV basé sur CTS. Mais il a un circuit d'écriture plus complexe. V_{STT} et V_{SHE} contrôler la direction des courants d'écriture.

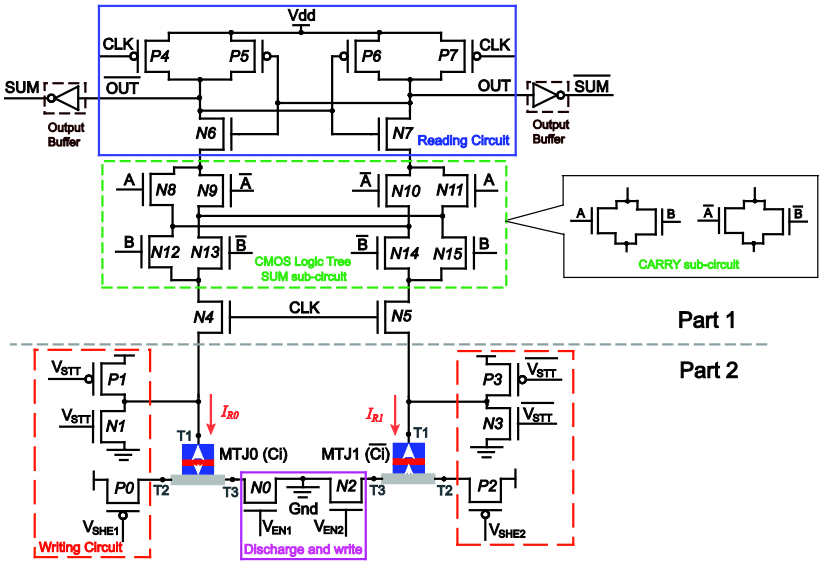


Figure R.20 Schéma de CTS + ESH ACNV

Les résultats de la simulation montrent que CTS + ESH ACNV présente des avantages en temps et en énergie lors du maintien de la même taille de circuit. Pour effectuer une opération

comprenant l'écriture et la lecture, l'ACNV proposée nécessite 38% moins de temps de fonctionnement (temps de lecture et temps d'écriture) et 30,8% moins d'énergie.

Chapitre 4 Mémoire adressable par contenu non-volatile (MACNV)

La mémoire adressable par contenu (MAC) est largement utilisée dans de nombreuses applications telles que les routeurs réseau, les processeurs, etc. Elle compare le mot de recherche avec sa mémoire puis renvoie l'adresse ou le mot a été trouvé. Le mot de recherche est d'abord chargé sur les lignes de recherche. La tension sur la ligne de sortie sera déchargée s'il existe un ou plusieurs bits qui sont différent du mot de recherche. La tension reste à un niveau élevé si tous les bits correspondent au mot de recherche. La MAC basée sur CMOS souffre d'un problème de puissance élevée et d'une faible densité en raison du courant de fuite. MAC non-volatile (MACNV) basé sur des dispositifs spintroniques tels que JTM est une solution efficace pour ces problèmes.

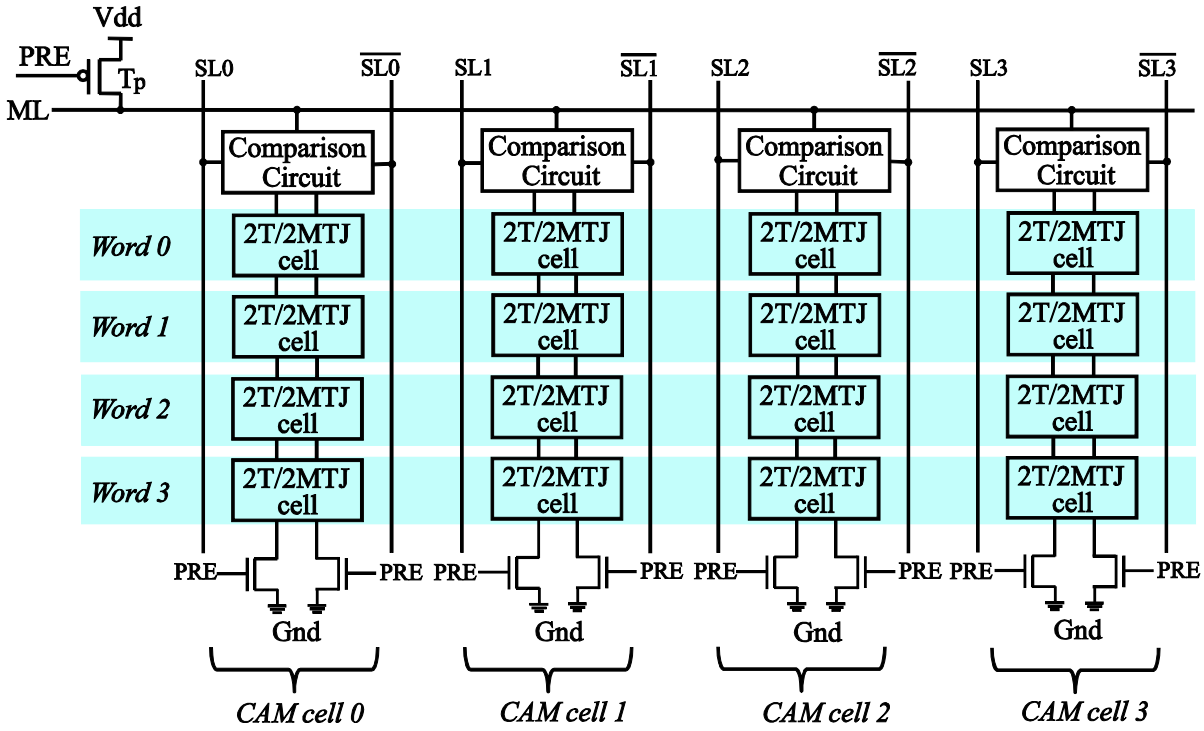


Figure R.21 Structure de la mémoire adressable par contenu non-volatile (MACNV)

Nous proposons un MACNV comme une des applications de l'architecture LEM. Plusieurs JTMs utilisées pour le stockage et la fonction logique partagent le même circuit de comparaison pour assurer l'efficacité de surface. Figure R.21 illustre la structure de MACNV. ML est pré-chargée par le transistor PMOS (T_p) lorsque le signal est activé. Quand un mot

(par exemple, "0100") est recherché, le premier mot ($Word_0$) sera chargé. S'il est différent du mot de recherche, les mots suivants seront adressés jusqu'à ce qu'un correspond mot soit trouvé.

Le circuit de comparaison et le circuit d'écriture sont partagés par des cellules de stockage dans la même colonne (voir Figure R.22). L'opération de recherche (ou l'opération de comparaison) est effectuée en comparant le mot de recherche avec les données stockées dans les JTMs. ML tiendra la charge lorsque tous les bits correspondent aux lignes de recherche $SL_3 - SL_0$. Sinon, ML sera déchargé.

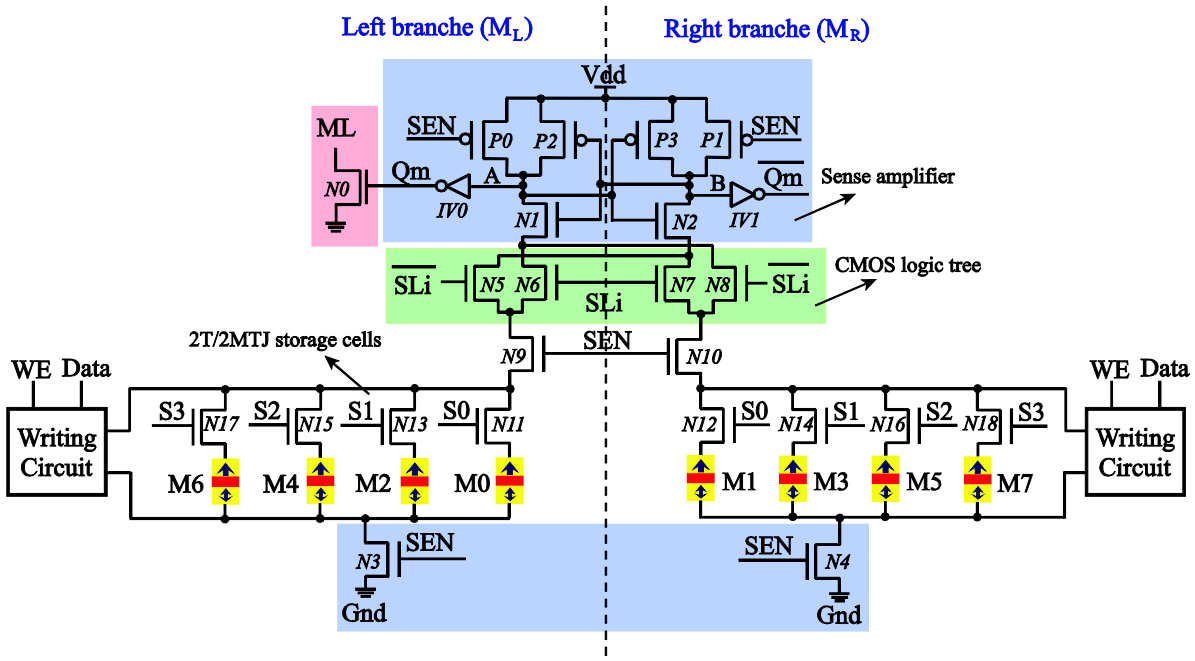


Figure R.22 Schéma de la cellule MAC basique. SL_i représente la ligne de recherche, où i est le nombre de lignes de mots.

Table R.4 Mécanisme d'opération de la cellule MAC

Donnée stockée		**Donnée de recherche**	**Qm**	**N_0**	**Résultat**
(M_L, M_R)	Donnée non-volatile				
(P, AP)	0	0	Gnd	Fermé	*Match*
		1	V_{dd}	Ouvert	*Mismatch*
(AP, P)	1	0	V_{dd}	Ouvert	*Mismatch*
		1	Gnd	Fermé	*Match*

Table R.4 résume la relation entre les données stockées, le mot de recherche et le résultat de comparaison. L'opération de recherche de la MACNV ne nécessite que 110 ps. La

consommation d'énergie est aussi faible que 3.2 fJ/bit/search. La MACNV multi-bit promet une commutation rapide du contexte car tous les éléments de stockage sont directement connectés au circuit de comparaison. L'efficacité de la surface devient plus significative lors de l'augmentation du nombre de mots.

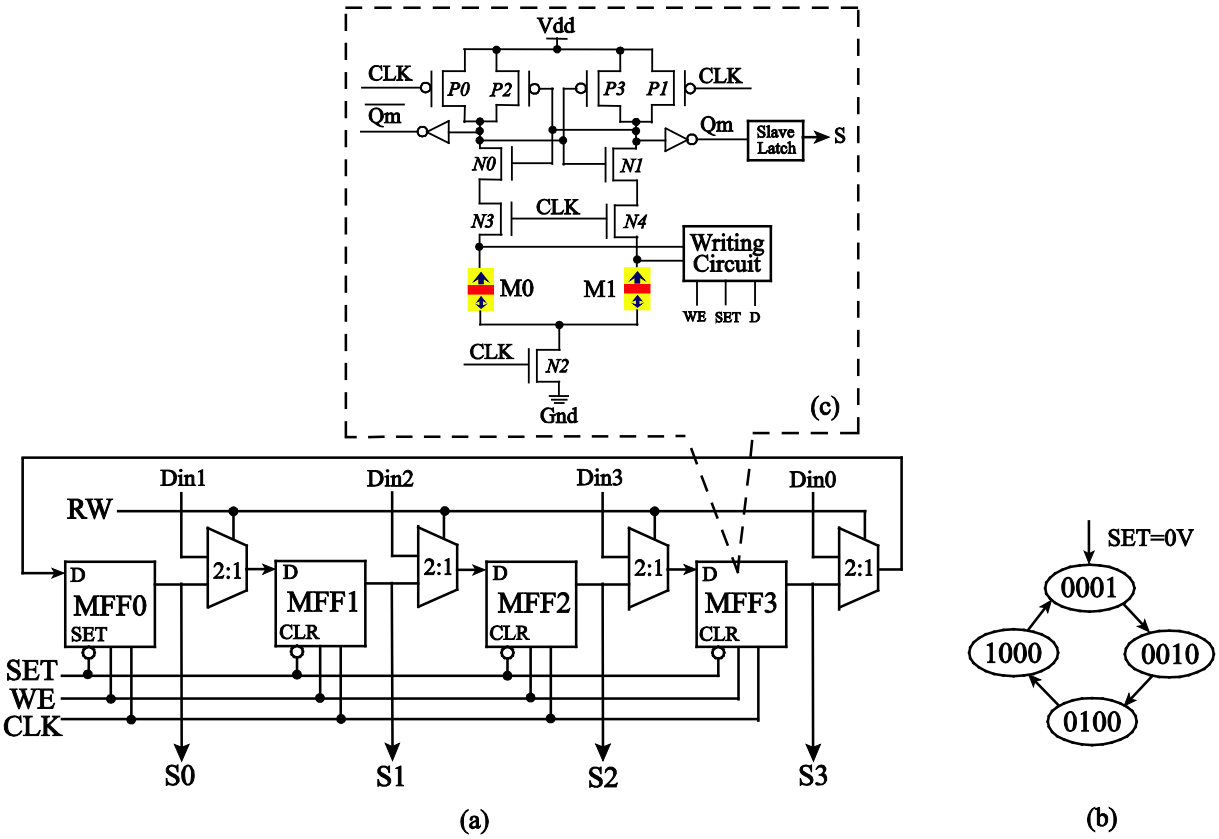


Figure R.23 (a) Schéma du décodeur magnétique basé sur le registre à décalage (RDDM) (b) Diagramme d'état de RDDM ($S_3S_2S_1S_0$) (c) Flip-flop magnétique utilisant deux JTMs qui sont toujours dans les états complémentaires

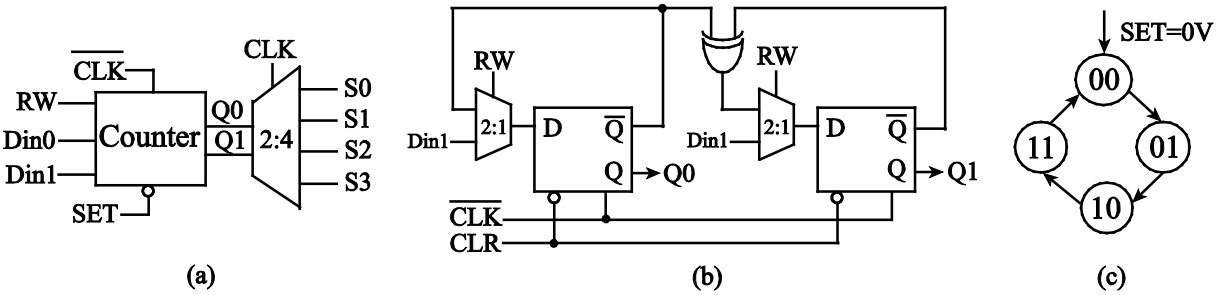


Figure R.24 (a) Schéma du décodeur magnétique basé sur le compteur (CDM) (b) Structure du compteur basé sur CMOS (c) Schéma d'état du compteur basé sur CMOS (Q_1Q_0)

Deux décodeurs magnétiques (DMs), c'est-à-dire un décodeur basé sur le registre à décalage (RDDM) et le décodeur basé sur le compteur (CDM), servent de circuit de commutation pour la MACNV (voir Figure R.23 et Figure R.24). Les deux DMs permettent de conserver l'adresse du mot sélectionné, même en état de coupure. En outre, les concepteurs peuvent

choisir une ligne particulière pour comparer avec le mot de recherche.

Conclusion

Cette thèse vise à concevoir et à simuler des circuits logiques non-volatiles intégrant les JTMs non seulement comme cellules de stockage mais aussi comme opérandes. Pour l'application logique, l'architecture LEM ouvre la voie d'intégrer les mémoires non-volatiles directement dans le circuit logique. Cette architecture réduit considérablement la distance de communication, réduisant ainsi le délai de transfert et l'énergie.

La structure hybride JTM/CMOS a été analysée, y compris les circuits de la lecture et d'écriture. Le circuit multi-bit a été conçu pour une plus grande efficacité de la surface, où plusieurs cellules de stockage partagent le même circuit de lecture et d'écriture. Afin de résoudre les problèmes causés par la structure asymétrique, la structure symétrique et les circuits de lecture ADPCS ont été proposés.

Les circuits logiques et arithmétiques basés sur l'architecture LEM ont été conçus en utilisant le modèle de JTM. La structure de l’ACNV 1-bit a été détaillée, suivie de l'étude de l'effet de différents facteurs sur le temps de fonctionnement et l'énergie. Par rapport à l’AC basée sur CMOS, l’ACNV a montré des avantages sur la consommation d'énergie statique et la surface en raison de la technologie d'intégration 3-D. Ensuite, nous avons proposé et comparé trois structures d’ACNV 8-bit qui réalisent une non-volatilité complète. Enfin, nous avons optimisé l’ACNV en termes de fiabilité et de performance d'écriture.

L'architecture LEM a également été appliquée à la conception de MACNV. Il a eu des avantages dans la vitesse de recherche et la consommation d'énergie par rapport à d'autres MACs. Deux décodeurs magnétiques ont été conçus pour la sélection de ligne de mots.

Conception et développement de circuits logiques de faible consommation et fiables basés sur des jonctions tunnel magnétiques à écriture par transfert de spin

Résumé - Les dispositifs de spintronique, tels que la jonction tunnel magnétique (JTM) écrite par transfert de spin, sont largement étudiés comme une solution pour aider à repousser les limites à venir dans la miniaturisation des circuits électroniques, en particulier la consommation statique causée par la diminution de la taille des dispositifs CMOS. L'architecture logique-en-mémoire (LEM) hybride permet de réduire le temps et la consommation dynamique de transfert entre la mémoire et la logique. Cette thèse consiste à concevoir des circuits logiques et mémoires, en combinant les technologies JTM et CMOS.

En utilisant un modèle compact JTM et le design-kit CMOS de STMicroelectronics, nous étudions des circuits hybrides JTM/CMOS 1-bit et multi-bit. Une mémoire MRAM basée sur la structure JTM/CMOS hybride est proposée. Puis, basés sur le concept de LEM, des circuits logiques/arithmétiques non-volatiles (NOT, AND, OR, XOR, ainsi qu'un additionneur complet) sont conçus, analysés et optimisés. Enfin, une mémoire adressable par contenu non-volatile (MACNV) et deux architectures de décodeurs magnétiques pour la sélection de ligne sont proposées.

Mots Clés: Spintronique, transfert de spin, jonction tunnel magnétique, circuits hybrides JTM/CMOS, circuits logiques/arithmétiques non-volatiles.

Design and development of low-power and reliable logic circuits based on spin-transfer torque magnetic tunnel junctions

Abstract - Spintronics devices, such as spin transfer torque based magnetic tunnel junction (STT-MTJ), are under intensive investigation to overcome the static power issue caused by the shrinking of CMOS technology. Hybrid logic-in-memory (LIM) architecture allows reducing latency and dynamic power due to long data traffic. This thesis focuses on the design of hybrid MTJ/CMOS logic circuits and memories.

By using a compact STT-MTJ model and the STMicroelectronics CMOS design kit, we design and optimize the single-bit and multi-bit hybrid MTJ/CMOS circuits. Magnetic random access memory (MRAM) based on the multi-context hybrid MTJ/CMOS structure is proposed. Then, based on the LIM concept, non-volatile logic/arithmetic circuits are designed and analyzed including NOT, AND, OR, XOR and full-adder (FA). Furthermore, we optimize the FA from the circuit-, structure- and device-level. Finally, LIM-based non-volatile content addressable memory (CAM) and magnetic decoders are designed.

Keywords: Spintronics, spin transfer torque, magnetic tunnel junction, hybrid MTJ/CMOS circuits, non-volatile logic/arithmetic circuits.

Thèse préparée au laboratoire TIMA (Techniques de l'Informatique et de la Microélectronique pour l'Architecture des ordinateurs), 46 Avenue Félix Viallet, 38031, Grenoble Cedex, France

ISBN: 978-2-11-129224-6