# Spin-Based Logic and Memory Technologies for Low-Power Systems

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Jongyeon Kim

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Chris H. Kim

February 2016

# Acknowledgements

First, I am deeply indebted and thankful to my advisor, Professor Chris H. Kim. I feel very fortunate to have been a member of his group. He offered me great opportunities to explore the interdisciplinary topics and collaborate with brilliant people from academia and industry. It was absolutely an exciting and rewarding experience for me. He was undoubtedly a great advisor but also a thoughtful senior that inspires my life. All the lessons he has delivered through his kind favor and consideration will remain in my mind for a long time. I appreciate all that he has done for me during my Ph.D study, and this dissertation would not have been finished without his generous support and endless encouragement.

I would also like to say a very big thank you to the members of my final defense committee: Professors Sachin Sapatnekar, Steven Koester, and Paul Crowell. I appreciate you taking time out of your busy schedules for my dissertation and final presentation. I am also grateful to Professor Jian-Ping Wang and David Lilja for serving on my preliminary oral defense committee.

I also thank all my colleagues in VLSI research lab for making my school life so much fun. Especially, I thank Ki Chul Chun and Seung-Hwan Song for their technical guidance to my research. I also thank all my friends: Wei Zhang, Pulkit Jain, Xiaofei Wang, Ayan Paul, Bongjin Kim, Won Ho Choi, Hoonki Kim, Weichao Xu, Qianying Tang, Saroj Satapathy, Somnath Kundu, Chen Zhou, Paul Mazanec, Saurabh Kumar, Ibrahim Ahmed, Muqing Liu, Luke Everson, Po Wei Chiu, Gyusung Park, Ahmed

Ammar, Nakul pande, Dan Liu, Abhishek Deshpande, Deepak Kumar Tagare, Woong Choi and Gyuseong Kang for all the helps and fun.

I am grateful for the unending support and trust from my whole family. First, I would like to thank my parents for everything they have done for me. I deeply respect all the lifelong sacrifices and efforts they have devoted to support me and my sister. Thank you for being hearty friends, great mentors, and who I want to be like. All the words fall short in describing their love to me. I thank my sister, Hye-Min, for making our home full of cheerful energy and filling my missing pieces in our family. I couldn't be happier with her marriage and I truly welcome a new family member, Dong-Young. I want to thank my parents-in-law and brother-in-law for all their pray for our peace, prosperity, and happiness.

Finally, thank you to my wife Shin Ae Lee. I feel so happy and lucky since you have made me realize that my choice was perfect. Your patience and support with me is amazing, and it was the greatest source of my energy during the Ph.D study. While facing unexpected adversities and overcoming challenging hardships together, I realize that I can't imagine my life without you by my side. Wherever we are or whatever we have, as long as we are together, I'll be happy with that. I hope we enjoy a new chapter of our life with our adorable Yerim and soon-to-be-born Yedong. As we always wish, may our life be full of more laughter, more joy, and more love.

# Abstract

As the end draws near for Moore’s law, the search for low-power alternatives to CMOS technology is intensifying. Among the various post-CMOS candidates, spintronic devices have gained special attention due to its unique features such as zero static power, compact size, and instant wakeup, while enabling an entirely new class of architectures such as processor-in-memory, logic-in-memory, and neuromorphic computing. However, traditional spintronics research has been mainly limited to the materials and single device level, so the main aim of this dissertation is to clearly describe spin-based logic and memory technologies by exploring the trade-off points across different levels of design abstraction (i.e. device, circuit, and architecture).

For spin-based logic, we benchmark the system-level capability of spin-logic technology using a hypothetical spintronic-based Intel Core i7 as a test vehicle. We describe how spin-based components are integrated into a computing system and the advantages that result. Even with early promises such as zero static power, lower device count, and lower supply voltage, technical barriers associated with spin devices such as low spin injection, limited spin diffusion length, and intrinsically high activity factor result in higher active power than CMOS.

For spin-based memory, a key aspect of technology evaluation is the development of a reliable MTJ model, so we first propose a technology-agnostic MTJ model specifically designed for evaluating the scalability and variability of STT-MRAM circuits. Using the proposed model, we evaluate the circuit-level scalability of MTJ technologies providing the detailed scaling methods and projection scenarios down to 7nm. For use in high speed on-chip cache applications, we also explore the feasibility of non-traditional MRAMs such as spin-Hall effect (SHE) MRAM which provides superior switching efficiency.

In addition to the spintronics research, a logic-compatible eflash-based neuromorphic core is designed to provide a highly efficient architecture for neural computing. We use a logic-compatible embedded flash memory to store synaptic weights to provide a simple implementation of restricted Boltzmann machine (RBM) which is a well-known neural algorithm for digit recognition. With the proposed current-based architecture, a neuron operation can be accomplished by simply comparing two currents corresponding to excitatory and inhibitory weights without large digital neuron circuits used in previous works.

# Contents

# List of Tables

# List of Figures

# Chapter 1 Introduction

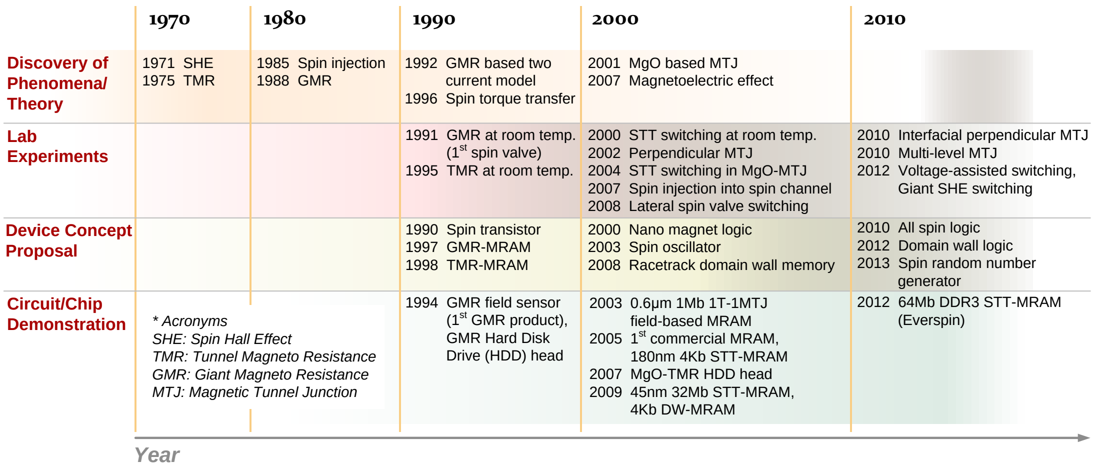
CMOS scaling, otherwise known as Moore's law, has transformed the way we create, process, communicate, and store information in the digital age [1]-[4]. As we approach the physical limits of CMOS technology however, it has become increasingly difficult to manage power dissipation issues [5]-[7]. The urgent need for a low power alternative has led to a flurry of research activity on novel post-CMOS device technologies [8], [9]. Among the various post-CMOS candidates, spintronics devices have gained momentum over the past decade as they have the potential to overcome the power and performance limitations of CMOS [10]-[12]. The Oxford dictionary defines spintronics as "a field of electronics in which electron spin is manipulated to yield a desired outcome." From a computing perspective, spintronic devices may have a profound impact on future electronic designs as they have the potential to offer unique capabilities such as zero static power, instant wake up, reduced device count, and lower switching energy, that would otherwise be difficult to simultaneously achieve using CMOS technology. Another intriguing feature of spintronic devices is that they could potentially augment existing Boolean computing methods by enabling entirely new class of architectures such as processor-in-memory, logic-in-memory, and analog/neuromorphic computing [13]-[15]. However, traditional spintronics research has been mainly limited to the materials and single device level so the actual benefits at the system level have been only superficially understood at best [16]. The main aim of this dissertation is to bring more clarity to spintronics technology by exploring the power and performance trade-offs at the device/circuit/system level focusing on logic and memory applications [17]. To provide a historical perspective, this chapter first gives an overview of the various milestones in spintronics research. We then introduce the working principles and development status of various spintronic devices targeted for logic, memory, and special functions.

## 1.1 Historical Advances in Spintronics Research

Fig. 1.1 shows the key milestones in spintronics research classified into the following four categories: new discoveries, key experiments, device concept proposals, and chip level demonstrations. Tunnel magnetoresistance (TMR) effect was first predicted in 1975 opening up the possibility that electron flow tunneling through a thin insulator can be controlled by manipulating the relative magnetization of two adjacent ferromagnet layers, which, in turn, induces two states of electrical resistance [19]. In 1988, a similar form of spin valve effect called giant magnetoresistance (GMR) was discovered in a multi-layer structure composed of ferromagnets and a metallic spacer layer [20]. The main difference between TMR and GMR is that TMR uses an insulating tunnel barrier to transmit current while GMR uses a metallic layer. In general, a larger impedance change between parallel and anti-parallel states (i.e. higher magnetoresistance ratio) can be obtained using TMR while GMR enables a lower stack resistance.

Demonstration of both GMR and TMR at room temperature led to rapid deployment of these concepts to commercial data storage products such as hard disk drive (HDD) and random access memory (RAM) [21]-[25]. In 1996, Slonczewski at IBM theoretically predicted that the magnetization of a free layer can be toggled using spin-polarized current rather than an external magnetic field. This effect commonly referred to as spin transfer torque (STT) has since been experimentally verified and proven to lower energy consumption and simplify the memory cell design compared to previous field-based switching [26]. Fig. 1.2 illustrates STT based switching in a magnetic tunnel junction (MTJ), a device composed of two ferromagnetic layers, a free layer and a fixed layer, separated by an ultra-thin tunneling barrier [27]. When electrons enter from the bottom fixed layer terminal as shown in Fig. 1.2(left), only the ones with the same magnetization manage to tunnel through, exerting spin torque on the free layer. Once the switching is complete, the magnetization directions of the two layers are in parallel to each other resulting in a low resistance state. When electrons enter from the top free layer terminal on the other hand, the ones with the opposite spin direction get reflected back to the free layer, switching the relative magnetization to an anti-parallel state. The difference in tunneling current between parallel state (low resistance) and anti-parallel state (high resistance) is utilized to encode binary data. Typically, the fixed layer is pinned by a single antiferromagnetic layer or a trilayer forming a synthetic antiferromanget (SAF) structure that does not rotate or switch during operation [28]. Experimental research on STT-based magnetization switching led to the actual demonstration of STT at room temperature validating the predictions made by theorists [29], [30]. With the advent of new tunnel barriers such as MgO, STT-MTJ devices have now become mature enough to be considered for commercial magnetoresistive RAM (MRAM) products [31], [32]. Recent trends in STT-MTJ research focus on reducing the switching energy using novel perpendicular anisotropy material, voltage-assisted switching, and spin-Hall effect [33]-[36]. Further details on each of these phenomena will be presented in following sections.

Exploiting magnetism for logic computation is a topic of growing interest. The key difference between spintronic devices for memory and logic is that the latter requires not only data storage but also data transfer over a longer distance by means of spin. In 1985, the idea that pure spin current can be generated by non-local electrical spin injection in a metallic lateral spin-valve (LSV) structure was proposed [37]. In the 2000s, LSV switching by spin accumulation and transportation was demonstrated at room temperature [38], [39]. Recently, long spin diffusion materials such as semiconductor and single layer graphene have been explored as an way to attain longer spin interconnection lengths [40]-[42].



| | 1970 | 1980 | 1990 | 2000 | 2010 |
|---|---|---|---|---|---|
| **Discovery of Phenomena/ Theory** | 1971 SHE<br>1975 TMR | 1985 Spin injection<br>1988 GMR | 1992 GMR based two current model<br>1996 Spin torque transfer | 2001 MgO based MTJ<br>2007 Magnetoelectric effect | |
| **Lab Experiments** | | | 1991 GMR at room temp. (1st spin valve)<br>1995 TMR at room temp. | 2000 STT switching at room temp.<br>2002 Perpendicular MTJ<br>2004 STT switching in MgO-MTJ<br>2007 Spin injection into spin channel<br>2008 Lateral spin valve switching | 2010 Interfacial perpendicular MTJ<br>2010 Multi-level MTJ<br>2012 Voltage-assisted switching, Giant SHE switching |
| **Device Concept Proposal** | | | 1990 Spin transistor<br>1997 GMR-MRAM<br>1998 TMR-MRAM | 2000 Nano magnet logic<br>2003 Spin oscillator<br>2008 Racetrack domain wall memory | 2010 All spin logic<br>2012 Domain wall logic<br>2013 Spin random number generator |
| **Circuit/Chip Demonstration** | ** Acronyms*<br>*SHE: Spin Hall Effect*<br>*TMR: Tunnel Magneto Resistance*<br>*GMR: Giant Magneto Resistance*<br>*MTJ: Magnetic Tunnel Junction* | | 1994 GMR field sensor (1st GMR product), GMR Hard Disk Drive (HDD) head | 2003 0.6μm 1Mb 1T-1MTJ field-based MRAM<br>2005 1st commercial MRAM, 180nm 4Kb STT-MRAM<br>2007 MgO-TMR HDD head<br>2009 45nm 32Mb STT-MRAM, 4Kb DW-MRAM | 2012 64Mb DDR3 STT-MRAM (Everspin) |

**Fig. 1.1 Historical advances in spintronics research.**

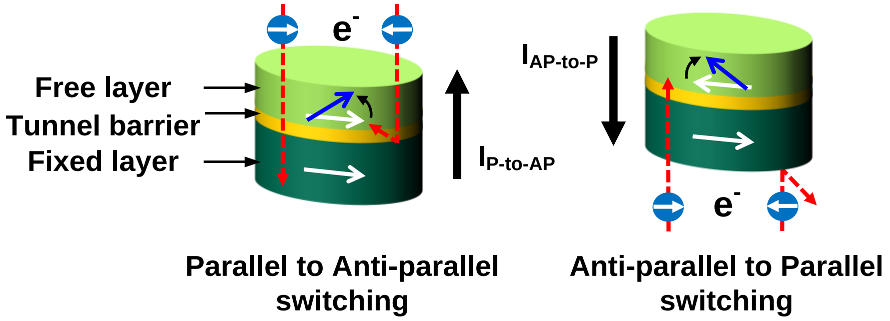

e-
Free layer
Tunnel barrier
Fixed layer
I_P-to-AP
I_AP-to-P
e-
Parallel to Anti-parallel switching
Anti-parallel to Parallel switching

**Fig. 1.2 STT switching in MTJ [27]. Electron flows for each spin polarization are shown separately for illustration purposes. We show complete transmission and reflection for both spin polarizations; however in reality, they are partially transmitted and reflected.**

## 1.2 Spintronics for Logic

The main attraction of spintronics for logic application is the nonvolatility which could enable computing systems with zero static power and instant-on-off features. The use of magnetic components to enhance the capability of conventional CMOS is also an active and fertile area of research. For instance, non-critical local circuits can be built using spin for non-volatility (=zero static power) and low device count, while speed-critical circuits can be built using CMOS for low dynamic power and good interconnect performance. In this section, we introduce key spin based logic devices that are being actively pursued by the materials and device communities.

An all-spin logic (ASL) device consists of input and output magnets connected by a channel medium (typically copper or graphene) as shown in Fig. 1.3(a). It utilizes spin injection, spin diffusion, and STT switching in a LSV structure to perform logic operation [18]. Fig. 1.3(b) shows the LSV device structure and the measured spin signal $\Delta V/I$ for a metallic channel used to demonstrate the spin current induced magnetization switching principle [39]. Here, polarized spin electrons injected and diffused through the channel give rise to a difference in the electrochemical potential between antiparallel and parallel states in the output detector. The spin torque transferred by the polarized spin electrons can then toggle the output magnetization. ASL device stores information using spin direction of the magnets and communicates using pure spin current, hence the name. Since STT switching current scales the magnet dimensions, ASL is generally thought to be a good post-CMOS candidate from a scaling perspective [43].

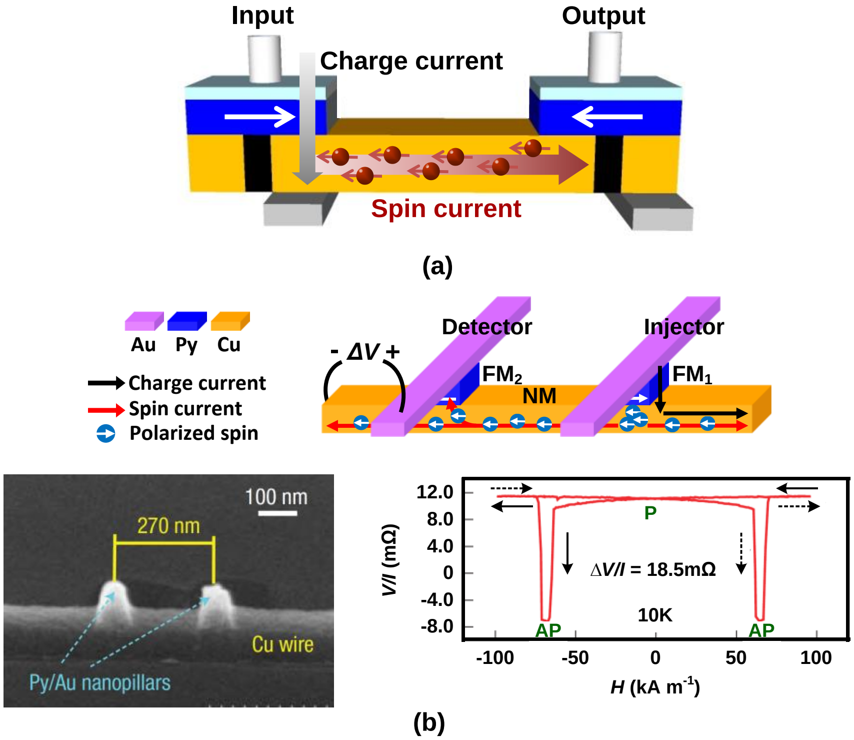


**Fig. 1.3 (a) ASL buffer circuit [18]. For illustration purposes, we only show the effective spin current (i.e. difference between majority and minority spins) assuming a positive voltage at the input terminal. (b) Lateral spin valve experiments [39]. Note that a positive voltage at the injector node results in spin current with the opposite magnetization direction as the input magnet (FM1). ΔV is defined as the difference in the spin-dependent electrochemical potential between FM1 and the channel.**

Domain wall logic (DWL) stores information in the position of a single domain wall (DW) [44]. As shown in Fig. 1.4(a), DW is the interface between different magnetic domains and can be shifted along a magnetic wire using spin-polarized current injected from either ends of the wire. This DW motion can be utilized for logic implementation as shown in Fig. 1.4(b). The magnetic wire works as a free layer forming a MTJ with a ferromagnet placed in the middle of DW wire. When a voltage is applied between the input and CLK terminals, the corresponding spin-polarized current causes DW motion along the free layer. Applying a voltage in the reverse direction results in a DW motion in the opposite direction. The position of DW represents the binary state information which can be read out by applying a voltage between the input and output terminals, or between the output and CLK terminals, depending on the specific timing sequence of the signals.

Nanomagnet logic (NML) utilizes magnetization direction as a state variable and processes information through magnetic dipole interaction between neighboring nanomagnets [45], [46]. At first, a NML-based circuit requires an initializing magnetic field to align the magnetization of nanomagnet chain along the hard axis (meta-stable state). As the magnetic field is removed, each nanomagnet is relaxed into a stable state with a preferred easy axis set by the input magnetization. Output magnetization is determined based on the majority logic performed by the superposition of incoming dipole fields. Fig. 1.5(a) and (b) shows quasi-stable state initialized with magnetic field and final stable state after removal of magnetic field, respectively. Despite benefits from nano-sized dimensions, scaling will be a challenge for NML since the initializing magnetic field will have to increase as the magnet scales as reported in [47], [48].

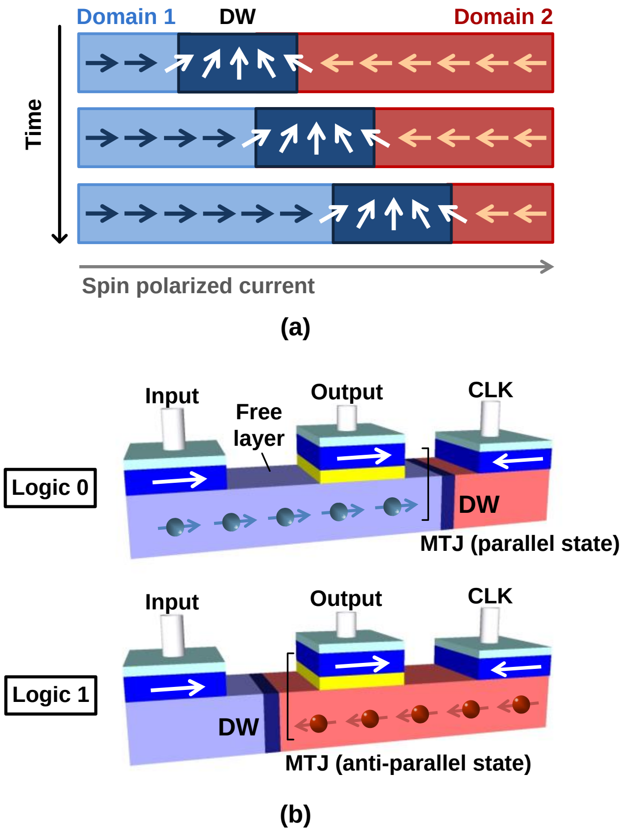


**Fig. 1.4** Domain wall logic [44]. (a) Domain wall motion induced by spin polarized current. (b) Logic gate concept and logic implementation using domain wall motion.

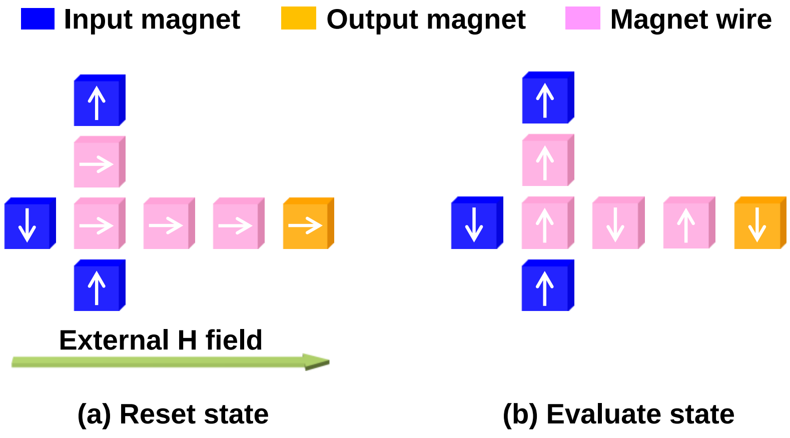


**Fig. 1.5** Nanomagnet logic [45]. (a) Reset state by applying an external magnetic field. (b) Evaluate state via dipole interaction after the external magnetic field is removed.

Spin-FET (spin field-effect transistor) is a novel device that combines an ordinary MOSFET (metal–oxide–semiconductor FET) structure with an MTJ [49], [50]. As shown in Fig. 1.6, a ferromagnet contact is placed on the source side while an MTJ is placed on the drain of the MOSFET. The MTJ on the drain side stores information via spin-polarized current. Then, the stored information is detected by output current of the transistor depending on the relative magnetization orientation between the source and the drain [51], [52]. The reconfigurable nature of spin-FET coupled with the high-integration density of CMOS makes this technology attractive for field-programmable gate array (FPGA) applications.

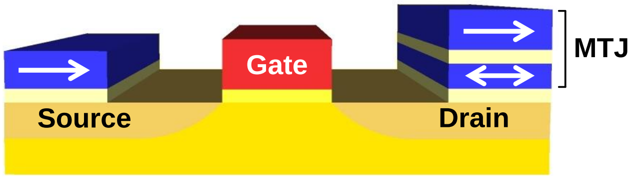
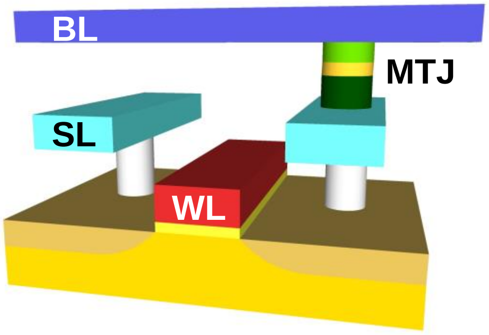


**Fig. 1.6  Basic structure of spin-FET [51].**

## 1.3  Spintronics for Memory

Spin transfer torque MRAM (STT-MRAM) has been drawing a great deal of attention as it has the potential to combine the speed of SRAM, the density of DRAM, and the nonvolatility of Flash, while providing excellent scalability, unlimited endurance, and CMOS-compatibility [53]. STT-MRAM can improve the cache access latency of last level caches (e.g. >64Mb) by reducing the global interconnect delay, a critical performance bottleneck in SRAM based L3 or L4 caches [54], [55]. STT-MTJ has been successfully integrated into advanced CMOS processes and is generally accepted as the most viable storage element for post-CMOS memories [56]-[60]. As shown in Fig. 1.7, an STT-MRAM bit-cell consists of an MTJ and an access transistor. MTJ stores information with relative magnetization and the magnetization reversal happens based on STT switching. Write operation is accomplished by alternating the voltage polarities of BL and SL while read operation is accomplished by sensing the resistance difference between the reference and the accessed cells using a small read current bias.



| | WL | BL | SL |
|---|---|---|---|
| Write$_{AP\text{-}P}$ | VDD | VDD | GND |
| Write$_{P\text{-}AP}$ | VDD | GND | VDD |
| Read | VDD | $V_{READ}$ | GND |

**Fig. 1.7 Cell structure and bias condition of STT-MRAM [53].**

One of the key directions of STT-MRAM research has been the reduction of the switching current for a given nonvolatility. To address this challenge, perpendicular anisotropy MTJs based on high crystal anisotropy material have been experimentally demonstrated [61]. Another approach is to take advantage of new switching mechanisms such as voltage-controlled magnetic anisotropy (VCMA) and spin-Hall effect (SHE). VCMA-based switching is being considered as a successor to conventional STT as the interfacial anisotropy in a CoFeB/MgO junction can be lowered when a voltage is applied to the MTJ [62], [63]. The switching sequence for VCMA is depicted in Fig. 1.8. A free layer with uniaxial anisotropy has two energetically equivalent states (i.e. parallel and anti-parallel states), separated by an energy barrier of $E_b$. In traditional STT switching, the barrier height between the two states remain unchanged so a large spin-polarized current must be injected for electrons to jump over the $E_b$ barrier and land on the other side. VCMA-based switching, on the other hand, can raise or lower the barrier height depending on the mode of operation. For example, in retention mode, no voltage is applied to the MTJ ensuring a high $E_b$ and hence good nonvolatility. During the switching however, the voltage applied to the MTJ lowers the $E_b$ and thus reducing the switching energy. When the voltage is off after the switching, $E_b$ is restored back to its former height. This novel switching method can be adopted for energy efficient MRAM without compromising nonvolatility. Note that applied voltage alone cannot switch the magnetization so an additional bias in the form of an external magnetic field or spin-polarized current is needed to complete the switching.

Giant spin-Hall effect (GSHE), the generation of large spin currents transverse to the charge current direction in specific spin-Hall metals (i.e. Pt, β-Ta, β-W, etc), is another option to realize a low-energy STT-MTJ switching [64]. Fig. 1.9 illustrates the generation of pure spin current by GSHE, along with the cell structure of a spin-Hall torque (SHT) MRAM cell. SHT-MRAM requires three terminals for separate read and bidirectional write operation. Note that read current can flow through the MTJ stack from /BL to GND when write transistor is turned off showing a large off-state resistance. Although this three-terminal device potentially results in an area penalty, it offers several advantages over the traditional 1T-MTJ STT-MRAM including (i) a spin injection efficiency ($I_{spin}/I_{charge}$) higher than 100% using optimal metal dimension, which enables a significantly low switching current without impacting nonvolatility; (ii) Separate read and write paths allowing longer device lifetime and disturb-free read operation. This is because only the small read current flows through the tunnel oxide as the write current flows through the spin hall metal itself [65], [66].

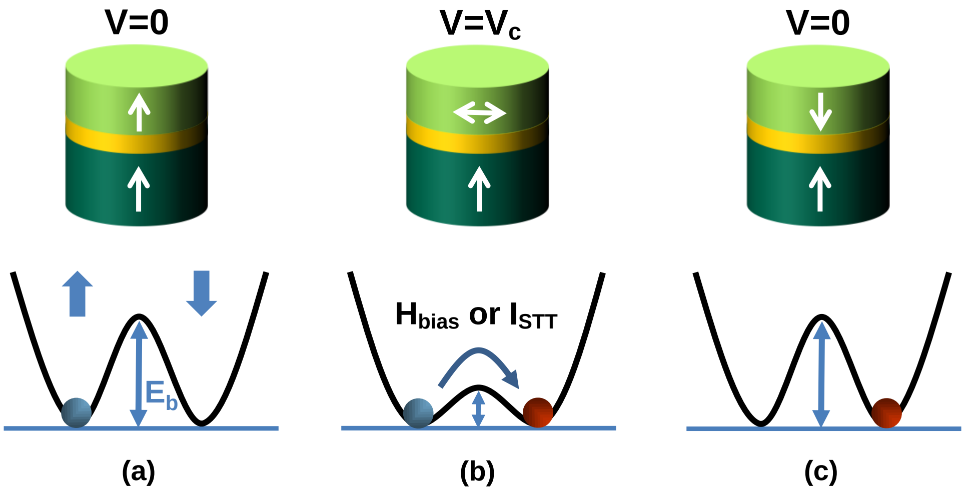
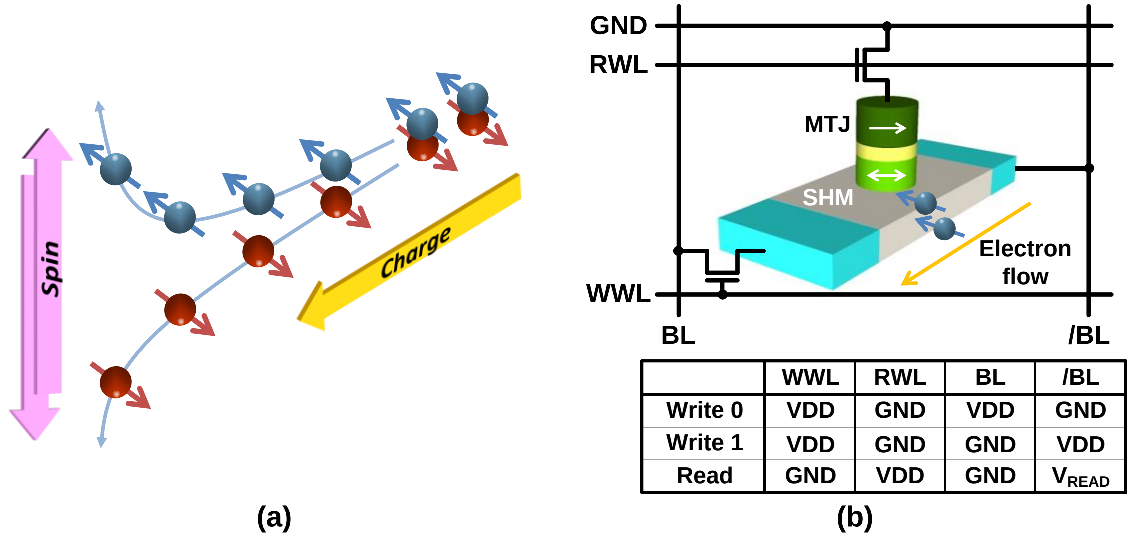


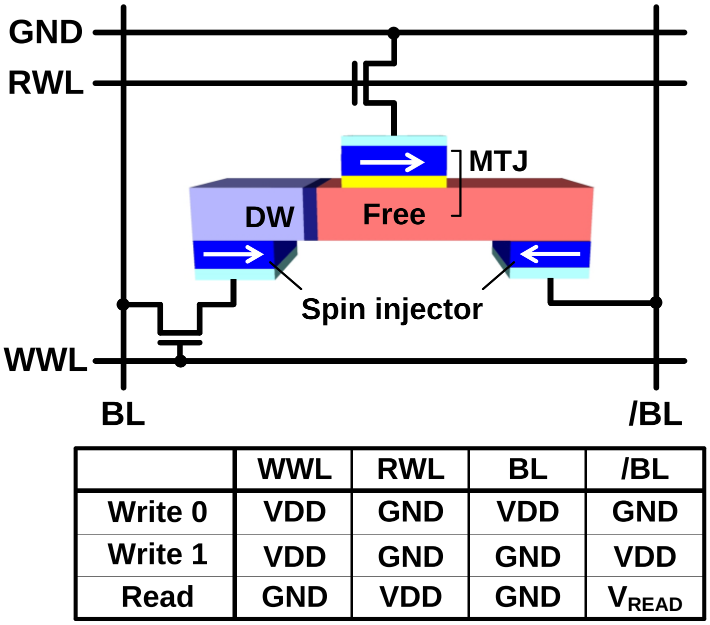
**Fig. 1.8 VCMA-based switching [63]. (a) High energy barrier before switching (retention mode). (b) Voltage-induced energy barrier lowering during the switching which requires additional stimuli to determine the switching direction. (c) Restored energy barrier after switching.**



|  | WWL | RWL | BL | /BL |
|---|---|---|---|---|
| Write 0 | VDD | GND | VDD | GND |
| Write 1 | VDD | GND | GND | VDD |
| Read | GND | VDD | GND | $V_{READ}$ |

(a) (b)

**Fig. 1.9 Spin-Hall Effect based STT-MRAM. (a) Transverse spin current generation by giant spin-Hall effect [64]. (b) Memory cell configuration and bias conditions for write and read [66].**



|  | WWL | RWL | BL | /BL |
|---|---|---|---|---|
| Write 0 | VDD | GND | VDD | GND |
| Write 1 | VDD | GND | GND | VDD |
| Read | GND | VDD | GND | $V_{READ}$ |

**Fig. 1.10 Domain Wall MRAM: cell structure and bias conditions [69].**

Similar to the domain wall logic, there has also been a proposal for utilizing the position of the DW for memory applications [53], [67]. A typical three-terminal DW memory employs two fixed layers in antiparallel configuration for spin injection, which enables a bidirectional DW motion along the free layer to encode binary information [68]. Depending on the position of the DW, two possible relative magnetization orientations of the MTJ are translated to either low or high current during the read operation. Since the current paths for read and write are separated, high-speed operation with improved reliability is possible [69]. Bit-cell configuration and basic operations are shown in Fig. 1.10.

## 1.4 Spintronics for Special Functions

Precessional motion and physical randomness in spintronic devices may offer new ways to design special functional blocks. For example, the steady-state magnetization precession induced by spin torque effect can be used as a spin oscillator to generate a microwave signal [70]. The main advantages of spin oscillator over CMOS-based voltage-controlled oscillator (VCO) are compact size, large frequency tuning range, and good scalability. Fig. 1.11(a) shows the working principle based on both STT and TMR effects. When a charge current is applied to the MTJ, the spin torque excites the free layer magnetization into steady-state oscillation cancelling out the damping torque. Note that the frequency of the oscillation can be tuned by the amount of charge current applied to the MTJ. The oscillating magnetization of the free layer relative to that of the fixed layer induces a change in resistance generating a time-varying output voltage as shown in Fig. 1.11(b) [71]. Spin oscillators are being explored as an alternative to conventional ring oscillator based VCOs or LC-VCOs [72], and may enable new capabilities such as high-density parallel signal demodulators and inter/intra chip wireless communication.

The random thermal fluctuation present in a nanomagnet can be amplified for generating random bits [73]. Fig. 1.12(a) shows the operation sequence to collect physical random bits from a single MTJ. First, a negative reset ($I_{reset}$) current initialize a MTJ to an anti-parallel state assuming a bottom-pinned MTJ structure. Then, by applying a perturbation current ($I_{perturb}$, an intermediate write current) that will force the magnetization direction to a neutral state and turning off the bias, a random output can be generated according to the thermal noise in the device. Finally, the MTJ state can be read out using a read bias current ($I_{read}$) and a sensing circuit. Energy diagram for each sequence is presented in Fig. 1.12(b).

Summary of the post-CMOS spintronic devices reviewed in this section is presented in Table 1.1.

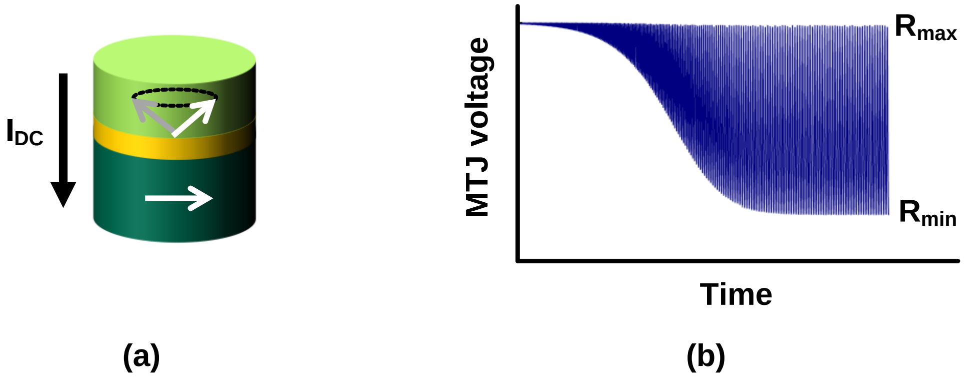


**Fig. 1.11 Spin-based oscillator [70]. (a) Spin precession in MTJ with an in-plane magnetized fixed layer and an out-of-plane magnetized free layer. (b) Time-varying MTJ voltage generated by the free layer oscillation.**

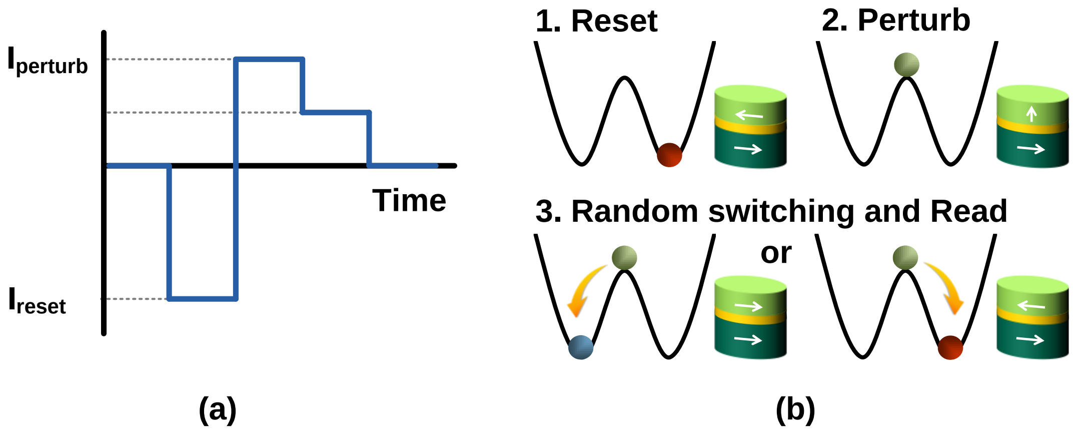


**Fig. 1.12 Spin-based random number generator [73]. (a) Operation sequence for collecting physical random bits from a single MTJ. (b) Working principle with energy diagram and corresponding magnetization orientation.**

| | Device name | State variable | Operating principle | Key Features | Status/ Maturity |
|---|---|---|---|---|---|
| Logic | All spin logic | Absolute magnetization | Non-local spin transport, STT switching | High scalability, Low switching energy | Non local switching verified |
| | Self-contained logic | Relative magnetization | STT by GSHE and TMR | Charge-based data communication | Concept only |
| | Domain wall logic | Domain wall position | Current-driven DW motion | Output evaluation by voltage clock | Concept only |
| | Nanomagnet logic | Absolute magnetization | Magnetic dipole interaction | Energy overhead for initializing field | Single logic gate demo-ed |
| | Spin-FET | Relative magnetization | Magnetization-dependent output current | High integration FPGA with spin-MOSFET | Junction structure demo-ed |
| Memory | STT-MRAM | Relative magnetization | Write: STT, Read: TMR | Highly scalable universal memory | Product prototyping stage |
| | VCMA-MRAM | Relative magnetization | Write: Energy barrier control by voltage, Read: TMR | Additional stimuli for switching direction | MTJ switching demonstrated |
| | SHE-MRAM | Relative magnetization | Write: STT by GSHE, Read: TMR | SHE efficiency depends on SHM dim. | MTJ switching demonstrated |
| | DW-MRAM | Domain wall position | Write: DW motion, Read: TMR | Separate paths for read and write | Low density array demo-ed |
| Others | Spin oscillator | Relative magnetization | Resistance change by spin precession | Compact size, Wide frequency range | Single device tested |
| | Random Num. Generator | Relative magnetization | Randomness of MTJ switching probability | Compact size, tunability | Concept only |

**Table 1.1 Summary of key spintronics devices.**

# Chapter 2 Spin-Based Logic: a Case Study on a High Performance Microprocessor

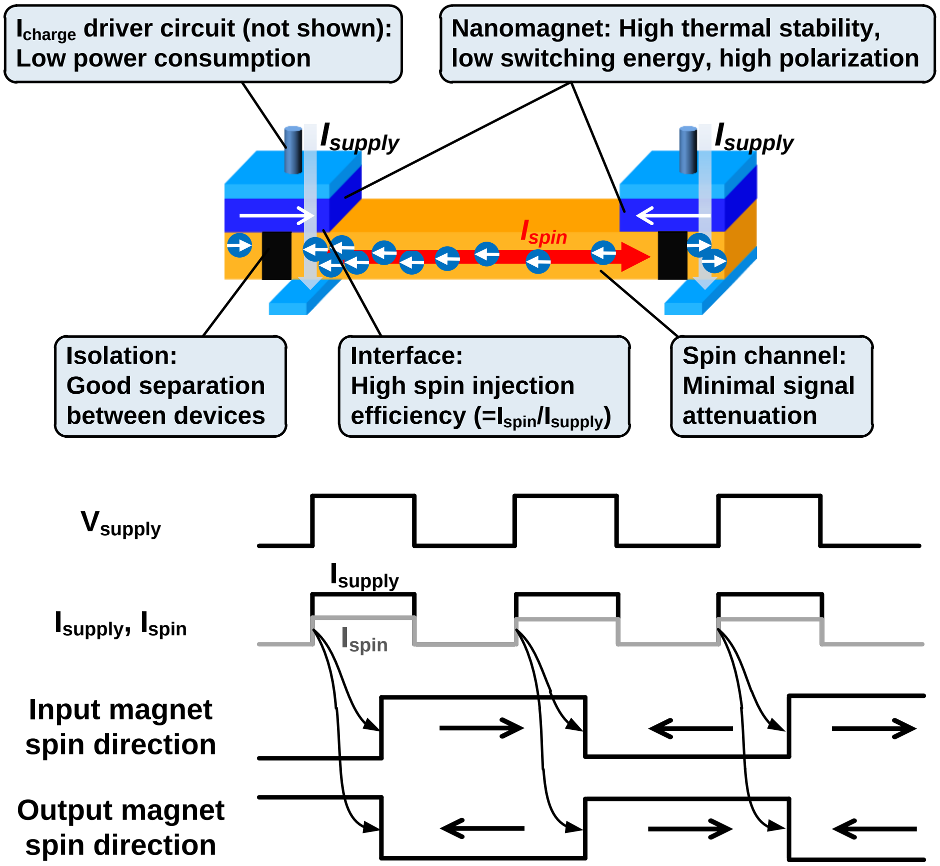
This chapter provides more clarity to spin-based logic technology by exploring the power and performance trade-offs at the system level using a spintronics based Intel's Core i7 processor as the test vehicle [17]. We choose All Spin Logic (ASL) device as the technology platform for this case study although a similar methodology could be applied to other spintronic devices [18]. We describe our benchmarking methodology whereby a simple method for estimating the device count and switching energy is proposed. We also address the signal attenuation issue in spin based interconnects and present guidelines for assessing and optimizing the total interconnect power. Finally, the power consumption of an ASL based processor is compared with its CMOS counterpart for various device parameters and operating scenarios (e.g. all cores active, one core active). We believe that the fundamental principles and perspectives gained from this study will help guide future spintronics device research and pave the way for a more rapid deployment of spintronics technology.

## 2.1 All Spin Logic (ASL) Components

Of particular interest in this work is the power and performance evaluation of spin-based computing system based on ASL. Recently, ASL is being considered as a promising post-CMOS device candidate due to its nonvolatility, higher density, lower device count, and good scalability. In this section, a bottom up overview of all spin based components is provided starting from individual devices and logic gates to functional blocks and processor system.

### *2.1.1 ASL Device Basics*

A conceptual diagram of an ASL-based inverter utilizing the LSV structure is shown in Fig. 2.1. Although ASL devices come in several different flavors (for example, the injector current can be a clock pulse or a constant DC supply, the interface between the nanomagnet and the channel can be either a direct contact or a magnetic tunneling junction depending on the material type), they all share the same basic components: input and output nanomagnets to store the digital information, a channel to transfer spin information to the next stage, an isolation layer to provide separation between devices, and an interface between the nanomagnet and channel for injecting spin polarized electrons. Input and output nanomagnets have two possible magnetization states represented by left and right pointing arrows, and are connected through a channel. The input current ($I_{supply}$) provided by a supply voltage pulse ($V_{supply}$) passes through the input magnet generating spin-polarized electrons in the channel entrance. These accumulated spins induce non-equilibrium magnetization enforcing spin diffusion along the channel in the form of spin current ($I_{spin}$), which transfers only spin angular momentum without charge flow.  Note that a positive $V_{supply}$ results in $I_{spin}$ with the opposite magnetization direction as the input magnet. This can be explained as follows. Electrons injected by $I_{supply}$ flow from GND to $V_{supply}$ when a positive $V_{supply}$ is applied. Those with the same spin direction as the input magnet can easily move towards V$_{supply}$, while electrons with opposite spin polarity get reflected back into the channel contributing to an increase in I$_{spin}$. Conversely, a negative V$_{supply}$ results in I$_{spin}$ with the same magnetization direction as the input magnet. Subsequently, I$_{spin}$ propagates through the channel exerting spin-torque on the output magnet. Once I$_{spin}$ exceeds a certain switching threshold, the magnetization direction of the output magnet toggles. Thus, depending on the polarity of the V$_{supply}$, we can obtain either an INVERT function (positive V$_{supply}$), or a COPY function (negative V$_{supply}$) using the simple ASL device shown in Fig. 2.1.



I$_{charge}$ driver circuit (not shown): Low power consumption
Nanomagnet: High thermal stability, low switching energy, high polarization
I$_{supply}$
I$_{supply}$
I$_{spin}$
Isolation: Good separation between devices
Interface: High spin injection efficiency (=I$_{spin}$/I$_{supply}$)
Spin channel: Minimal signal attenuation
V$_{supply}$
I$_{supply}$
I$_{supply}$, I$_{spin}$
I$_{spin}$
Input magnet spin direction
Output magnet spin direction

**Fig. 2.1 (Above) Conceptual diagram of an ASL based inverter. The desired properties for each sub-components are listed. (Below) Waveforms illustrate the operating principle.**

One key requirement for proper operation is to ensure spin information flows from the input towards the output while information flowing in the other direction should be blocked. This directionality can be achieved by placing the GND node closer to the input terminal than the output terminal as shown in Fig. 2.1 [74]. It has been shown that a large $I_{spin}$ generated at the input can diffuse towards the output while spin injection in the opposite direction is greatly reduced. Another important point to note here is that spin can only propagate over a certain distance, known as spin diffusion length, beyond which spin transfer becomes negligible. It is, therefore, critical to use a channel material that can support longer diffusion length in order to ensure low-power and high-speed spin transport.

### *2.1.2 ASL Gate Implementation*

Fig. 2.2 shows the implementation of various Boolean operations using an ASL with positive $V_{supply}$. Note that the same configuration results in different Boolean logic functions for negative $V_{supply}$. In this work, we choose to construct gates using positive $V_{supply}$ since an INVERT operation cannot be realized efficiently using a negative $V_{supply}$. However, the same principle can be applied to the positive $V_{supply}$ case. We now describe each type of Boolean logic gate in more detail. As shown in Fig. 2.2(a), an inverter can be implemented using a single spin device comprising two magnets and a channel. A buffer (or COPY operation) can be implemented by adding another magnet at the output of the inverter, in which the second and the third magnets constitute another inverter, as shown in Fig. 2.2(b). When it comes to implementing multiple input gates, spin devices have to rely on the majority function (or inverse of majority function for a positive

$V_{supply}$), where the output value is determined based on whether the majority of the inputs are in a ‘0’ or in a ‘1’ state. For example, NAND gate based on majority logic is depicted in Fig. 2.2(c). Magnets with a fixed spin polarity, known as fixed magnets (denoted as “F”), may be used in order to achieve desired Boolean function at the output. Magnetization of the output magnet is determined by the superposition of spin polarized signals from all input magnets and fixed magnets. Note that an AND gate can be simply implemented by adding one magnet at the output node of a NAND gate. Another interesting feature of all spin gates is that they can be easily reconfigured (e.g. NAND to NOR, NOR to NAND) by switching the magnetization direction of the fixed magnets as shown in Fig. 2.2(d). Generally speaking, an N-input gate can be constructed using N free magnets and N-1 fixed magnets. These basic ASL gates are summarized in Fig. 2.3(a) and truth tables for multi-input gates are shown in Fig. 2.3(b).

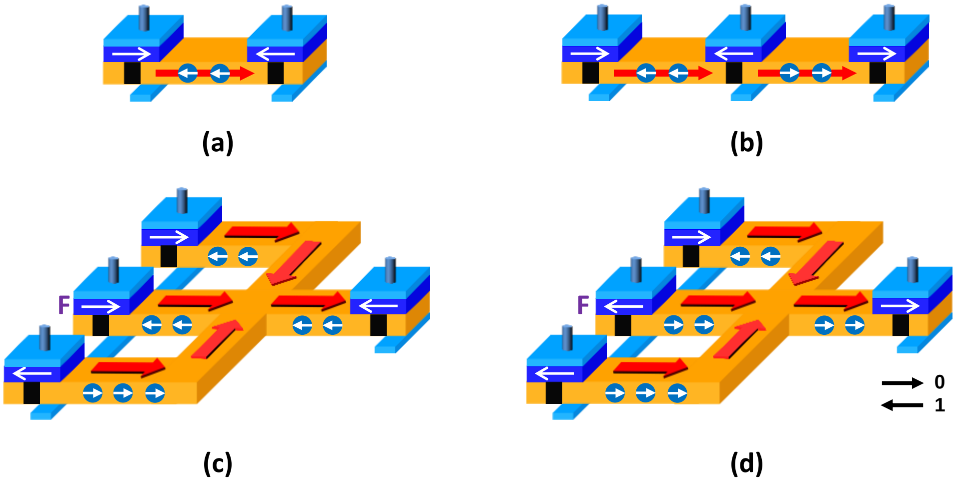
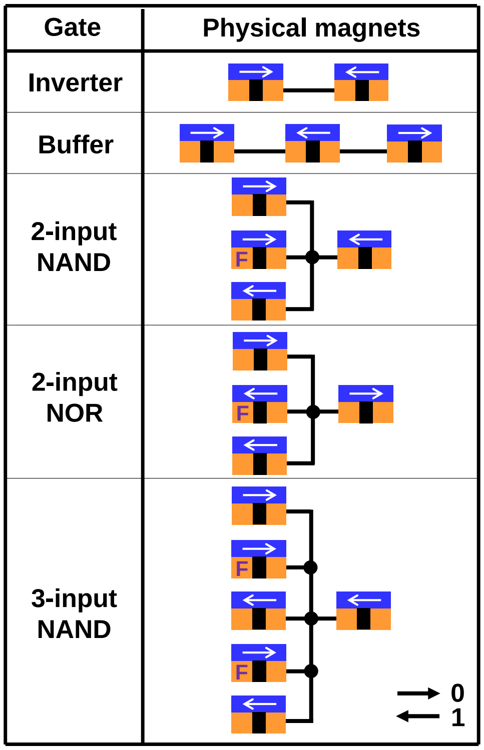


**Fig. 2.2 Implementation of ASL Boolean gates (positive $V_{supply}$). Only the net spin polarization shown for spin current. (a) Inverter (b) Buffer (c) NAND and (d) NOR. “F” denotes a magnet with fixed magnetization direction.**



(a)

| A | B | F | O |
|---|---|---|---|
| 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 |
| 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 |

2-input NAND

| A | B | C | F | F | O |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 0 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 | 1 |
| 1 | 1 | 1 | 0 | 0 | 0 |

3-input NAND

| A | B | C | D | F | F | F | O |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
|  |  |  | ⋮ |  |  |  |  |
|  |  |  | ⋮ |  |  |  |  |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |

4-input NAND

(b)

**Fig. 2.3 Construction of ASL gates. (a) Basic ASL gates. (b) Truth tables of multi-input NAND gates based on majority rule (A/B/C/D: input magnet, F: fixed magnet, O: output magnet)**

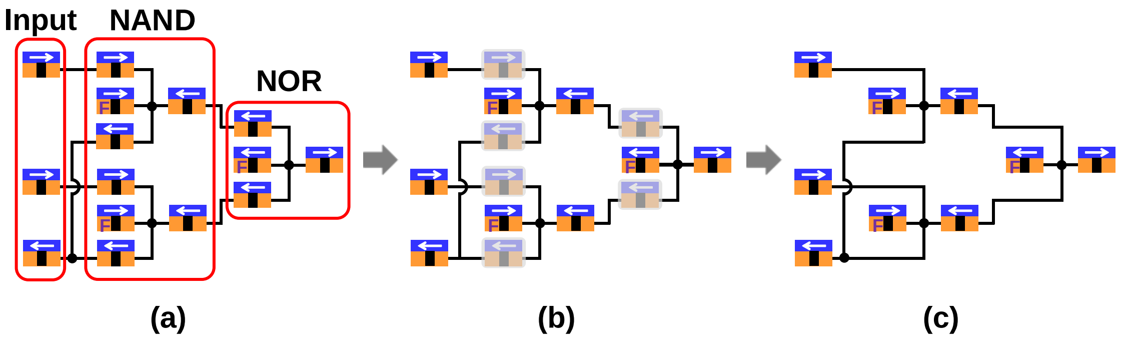


**Fig. 2.4 Cascading ASL gates. (a) Direct implementation. (b) Removal of redundant input and output magnets. (c) Final implementation using half the number of ASL devices.**

In cascaded spin logic implementation, each output magnet of a gate becomes the input magnet of the next gate, so one of the redundant magnets can be removed without affecting the logic function as can be seen in Fig. 2.4. It is obvious that the gates connected to the primary inputs will require one input magnet for each input signals. However, all subsequent gates in the cascaded structure can simply be implemented with fixed magnets and an output magnet only. Therefore, the total number of ASL devices required for the entire logic block implementation can be calculated as follows:

$$\begin{aligned} Total\, device\, count &= (\#\, of\;\; primary\, input\, magnets) \\ &+ \sum_{All\, gates} (\#\, of\;\; fixed + output\, magnets)_{gate} \end{aligned} \qquad (1)$$

Table 2.1 shows the device count comparison of a logic block using CMOS gates, individual spin gates, and cascaded spin gates. The number of devices for the cascaded ASL configuration can be calculated by subtracting the number of primary input magnets from the individual ASL gate’s total device count. Interestingly, the number of devices for a cascaded ASL configuration is half the number of devices required for CMOS implementation. This is indeed valid for typical logic blocks where the number of magnets connected to the primary input is small enough compared to the total device count including the input, output, and fixed magnets. Consequently, large combinational logic block can be implemented by using primarily the fixed and the output magnets only. This device count estimation method is based on a drop-in replacement scenario in which each CMOS gate is replaced by an equivalent ASL gate. However, the ASL implementation could be made more efficient if the circuit block can be re-synthesized to take advantage of the inherent majority function of ASL [43], [75].

| Function | Device count | | |
|---|---|---|---|
| | CMOS | Individual ASL gate | Cascaded ASL gate |
| Inverter | 2 | 2 | 1 |
| Buffer | 4 | 3 | 2 |
| 2-input NAND | 4 | 4 | 2 |
| 2-input NOR | 4 | 4 | 2 |
| 2-input AND | 6 | 5 | 3 |
| 2-input OR | 6 | 5 | 3 |
| 3-input NAND | 6 | 6 | 3 |
| 3-input NOR | 6 | 6 | 3 |
| 3-input AND | 8 | 7 | 4 |
| 3-input OR | 8 | 7 | 4 |

**Table 2.1 Device count comparison between CMOS, individual ASL, and cascaded ASL gates. Cascaded ASL gates can be implemented with half the number of devices compared to that of CMOS.**

### *2.1.3 ASL Pipeline Implementation*

We can leverage the inherent nonvolatility of spin technology to efficiently implement sequential logic elements such as latches and flip-flops as shown in Fig. 2.5 [76]. This is achieved by serially connecting ASL devices while carefully manipulating the CLK and $V_{const}$ signals. In Fig. 2.5(a), a level-sensitive positive latch is demonstrated using a pair of magnets. The first magnet controlled by CLK behaves like a switch, while the second magnet with a constant bias $V_{const}$ acts as a storage device. When the CLK goes high, the latch becomes transparent, and the pair of magnets transfer spin signal from input to output. On the other hand, a low CLK signal disables the spin signal propagation through the first magnet, and hence, the output retains its original state. This construction of the ASL latch closely resembles that of a conventional CMOS latch. Also cascading two latches and making them work in a master and slave fashion leads to an edge-triggered ASL flip-flop as illustrated in Fig. 2.5(b). The device count for the ASL flip-flop is 4 while a CMOS flip-flop would typically require 20 or more transistors. As such, the design of sequential elements can be drastically simplified in spin resulting in considerable savings in area and power.

A more significant advantage of the inherent nonvolatility is that it provides the possibility of removing sequential elements all together from the circuit. In a conventional CMOS pipeline, sequential elements are inserted between pipeline stages that are clocked in a synchronized manner. For realization of pipeline, CMOS requires separate supply voltage and clock as shown in Fig. 2.6(a). In contrast, ASL utilizes single input terminal as supply voltage and CLK at the same time. By proper manipulation of

CLK applied at the input node, data propagation can be controlled without explicit sequential elements. As illustrated in Fig. 2.6(b), non-overlapping dual phase clock applied to alternate stages of ASL pipeline enables sequential operation, since data propagation only happens when the CLK is enabled. For instance, when CLK2 is low, the first magnet of each ASL pipeline stage denoted as ‘B’ in Fig. 2.6(b), stores the final outcome from the previous pipeline stage. When CLK2 goes high, magnet ‘B’ launches the data to the following stage. Applying the dual-phase clocking mentioned above to every other logic gate enables ultra-deep pipeline increasing the throughput of system as shown in Fig. 2.6(c). Deeper pipelining in CMOS usually suffers from large power consumption in the sequential elements since the number of sequential elements has an exponential dependency on pipeline depth [77]. However, in the case of an ASL based pipeline, no sequential elements are present in the system so the power overhead for realizing an ultra-deep pipeline becomes negligible.

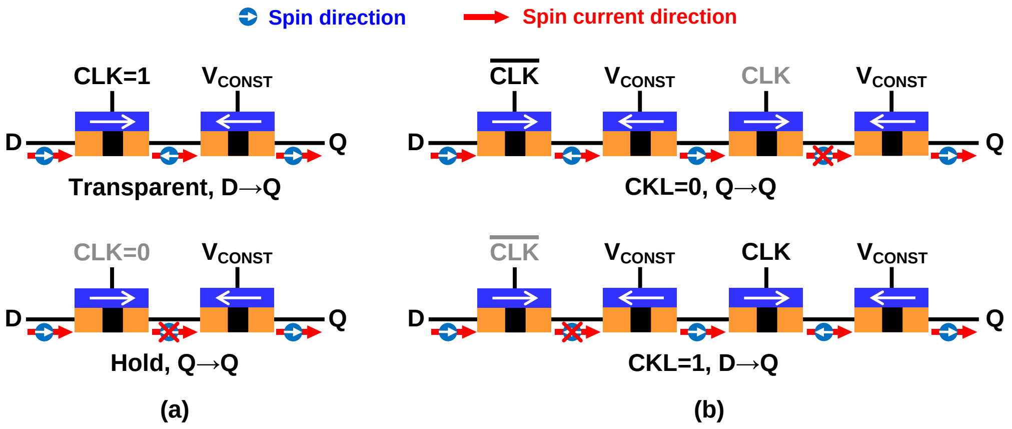


**Fig. 2.5 Implementation of ASL-based sequencing elements. (a) Level-sensitive positive latch. (b) Edge-triggered flip-flop. Clocked magnets control the spin signal propagation.**

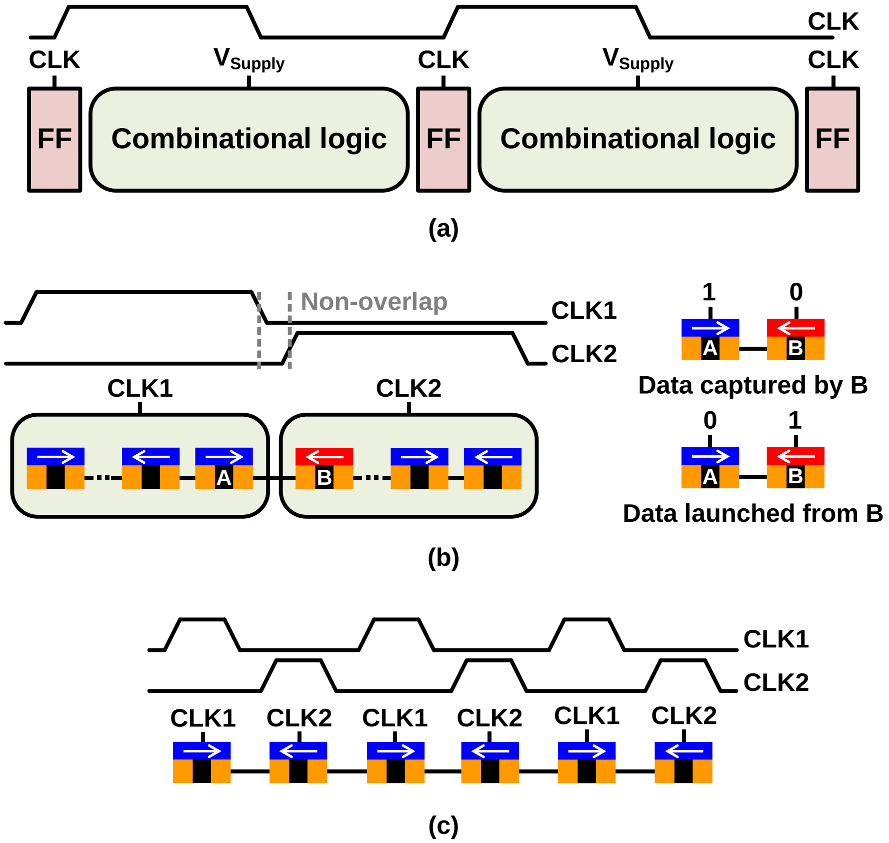


**Fig. 2.6 Construction of ASL based pipeline. (a) Conventional CMOS pipeline. (b) Pipeline architecture can be implemented in ASL without any sequencing elements by simply employing non-overlapping dual phase clocks. (c) Example of an ultra-deep pipeline with one logic gate per pipeline stage.**

### *2.1.4 Device Count Comparison*

In this section, we compare the device count between ASL and CMOS using Intel's Core i7 processor as the test vehicle. The system specifications are listed in Table 2.2 [78]. We consider the processor built in a 32nm high-k metal-gate CMOS technology. Our initial focus is on gate level power and performance, so for the time being, we will assume that the global interconnects between sub-blocks for spin are charge-based (not spin-based). Furthermore, we will assume no spin attenuation in the local interconnects, which removes the need for local ASL buffers. In reality, spin current cannot travel over a long distance (e.g. several micrometers), and as a result, numerous ASL buffers are needed to amplify the attenuated spin signal. As described in the previous sub-section, the total device count for a given ASL block is the sum of the number of fixed and output magnets for the ASL gates and the number of primary inputs for that block. The device count for ASL gates was shown to be roughly half that of CMOS. Intel's Core i7 processor consists of roughly 1 billion CMOS transistors out of which approximately 0.46 billion are used for SRAM caches while the remaining 0.54 billion are used in random logic. An ASL implementation of the logic part can be simply estimated as 0.54/2=0.27 billion based on (1). For a more accurate estimate, we need to check if indeed the number of input magnets is negligible compared to the total device count. To estimate the number of input magnets, a well-established empirical relationship known as the Rent's rule is utilized. According to Rent's rule, the relationship between the number of I/O terminals of a logic block (T) and the number of gates in the logic block (N) is given as [79]

$$T = kN^{p} \tag{2}$$

Where k is the average number of terminals per gate and p is the connectivity of the gates (0<p<1). It has been reported that Intel's microprocessor family follows this Rent's relationship closely using parameters of k=2.09, p=0.36 [80]. N, which is the total number of gates in a logic block, can be roughly estimated using a known k value. Since k, which is the average number of terminals per gate, is approximately equal to 2, equivalent logic gate for this particular k value can be assumed to be an inverter. Now since an inverter has 2 transistors and total number of transistors present in the Core i7 processor is approximately 1 billion, total number of equivalent logic gates present in the processor can be given as N=1billion/2=0.5 billion. With known k and p values and previously estimated N value, as per Rent's rule, total number of pins for the ASL based Core i7 processor is found to be 2830, which is negligible compared to the number of devices used for the ASL gates. This, therefore, confirms that the random logic portion of the Core i7 chip can be implemented with only 0.27 billion devices.

| Parameter | Characteristic |
|---|---|
| Architecture | Sandy Bridge |
| Channel length | 32nm |
| Power supply | 0.7 ~ 1.15V |
| Die size | 216mm$^2$ |
| Transistor count | 1 billion |
| Power | 95W @ 3.4GHz |
| # of cores | 4 cores |
| Cache | 64KB L1 cache per core<br>256KB L2 cache per core<br>8MB L3 shared cache |

**Table 2.2 Summary of an Intel Core i7 processor chosen as the test vehicle for our system level study [78].**

## 2.2 Methodology for Estimating ASL Power Dissipation

In this section, we present a methodology to estimate the switching energy of ASL gates considering design space options under process constrains and specific system level requirements.

### *2.2.1 Strategy for Switching Energy Calculation*

Fig. 2.7 shows the overview of our switching energy estimation strategy of a single ASL gate. Switching energy can be expressed as $E=V_{supply}\cdot Q_{total}$, where $Q_{total}$ is the total amount of charge applied at the input magnet of the ASL gate for switching the state of the output magnet, and it can be expressed as $Q_{total}=I_{c,critical}\cdot t_{sw}$. Here, $I_{c,critical}$ is defined as the critical charge current for the given switching time $t_{sw}$. Only a fraction of $I_{c,critical}$ known as critical spin current, $I_{s,critical}$, is responsible for switching, and the corresponding fraction is known as spin injection ratio and denoted by $I_s/I_c$. Therefore, switching energy can be expressed as $E=V_{supply}\cdot I_{s,critical}\cdot(I_c/I_s)\cdot t_{sw}$. This final equation suggests that switching energy of an ASL gate can be reduced either by increasing $I_s/I_c$, or by lowering $I_{s,critical}$. $I_{s,critical}$ required for a successful switching of output magnet is estimated by a physical simulation framework based on a Landau-Lifshitz-Gilbert (LLG) solver. The inputs to the LLG solver are functions of the material and the dimension of the magnets. These dimensions and material parameters are, in turn, determined by the thermal stability factor ($\Delta$), which is set by the degree of nonvolatility of the system. Spin injection ratio ($I_s/I_c$) is a device parameter that represents a spin transport capability of the LSV structure, which is governed by materials and dimensions of magnet and channel. More details on how each of these parameters can be optimized for minimum chip power will be discussed in the following sections.

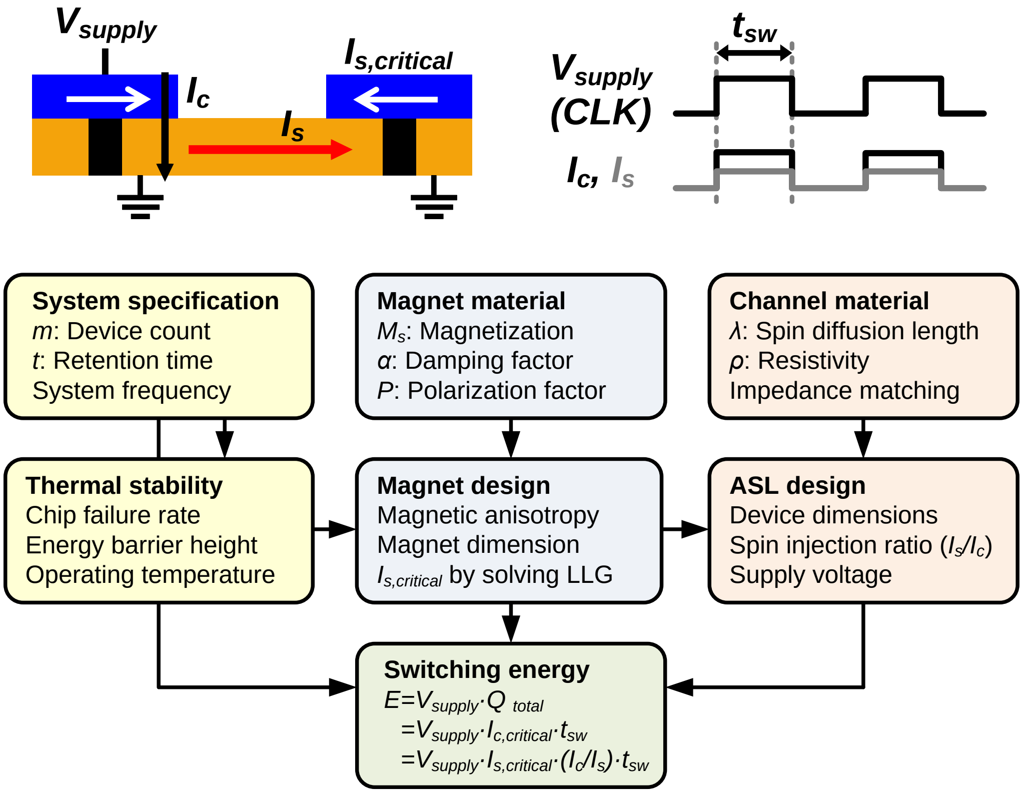


**Fig. 2.7 Work flow for calculating ASL switching energy.**

### *2.2.2 Thermal Stability Requirements*

In this section, we discuss how to determine the thermal stability factor in the context of a realistic microprocessor system. Thermal stability ($\Delta=E_b/k_BT$) is a measure of how much energy is required to flip the magnetization direction under thermal fluctuation, where $E_b$ is the energy barrier between two states, $k_B$ is the Boltzmann constant, $T$ is the temperature in absolute scale. To realize a practical nonvolatile system, thermal stability of each magnet must be high enough so that thermally assisted magnetization reversal can be prevented during the lifetime of the data (e.g. 10 years for storage data or 1 clock cycle for computation data). On the other hand, thermal stability of a magnet should be minimized for low switching energy. To satisfy these two conflicting requirements, the thermal stability must be determined based on the nonvolatility and switching energy requirements at the system level. To this end, we present a systematic methodology for calculating the optimal thermal stability value in this section. The derivation starts from the equation describing the thermal switching probability of a magnet [81]

$$P(t) = 1 - \exp(-t/\tau). \tag{3}$$

Here, $\tau$ is the relaxation time defined by Néel-Arrhenius equation

$$\tau = \tau_0 \exp(\frac{E_b}{k_B T}) , \tag{4}$$

where $\tau_0$ is the attempt cycle time (typically of the order of 1ns). Equation (4) can be further extended to the probability of an entire chip fail as [82]

$$F_{chip} = 1 - \exp\{-m \frac{t}{\tau_0} \exp(-\frac{E_b}{k_B T})\}. \tag{5}$$

where $m$ is the total number of devices in the system and $t$ is the retention time period. Fig. 2.8 plots the required thermal stability for an ASL Core i7 with a 10 year data retention time as a function of chip failure rate at room temperature (300K). Note that 0.27 billion of device count is used as estimated in the previous section. We see that a thermal stability greater than $69k_BT$ is needed to guarantee a chip failure rate lower than 0.01% (or 1 FIT). Here, FIT stands for failure in time and is equivalent to one failure in $10^9$ device-hours of operation.

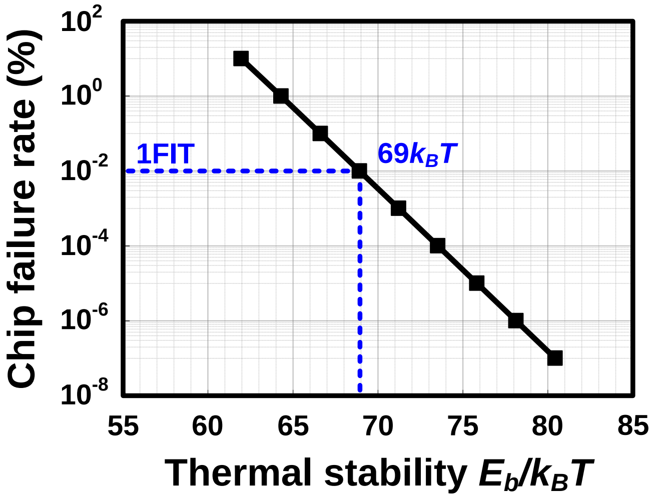


**Fig. 2.8 Thermal stability required for an ASL Core i7 with 0.27 billion devices to meet a 10 year retention time. Here, we assume a retention error of 1 FIT (=1 failure in $10^9$ device-hours of operation).**

### *2.2.3 Magnet Dimensions for Ensuring Nonvolatility*

We have already seen that degree of nonvolatility is determined by the system-level thermal stability criterion, which, in turn, sets the value of $E_b$ required of the magnetic material. $E_b$ can be expressed as

$$E_b = K_u V = H_k M_s V / 2 \tag{6}$$

where $K_u$ is the uniaxial magnetic anisotropy energy density, and $V$ is the volume of magnet. $H_k$ is the magnetic anisotropy field, which decides the energetic preference of the magnetization direction often referred to as the easy axis. $M_s$ is the saturation magnetization, which occurs when all domains are aligned. Depending on the orientation of easy axis, magnetic anisotropy can be classified into following two categories: in-plane magnetic anisotropy (IMA) and perpendicular magnetic anisotropy (PMA). The easy-axis of IMA lies in the x-y plane of the magnet while that of PMA is perpendicular to the x-y plane of the magnet. Fig. 2.9(a) and (b) show the dynamic spin motion during switching for IMA and PMA, respectively.

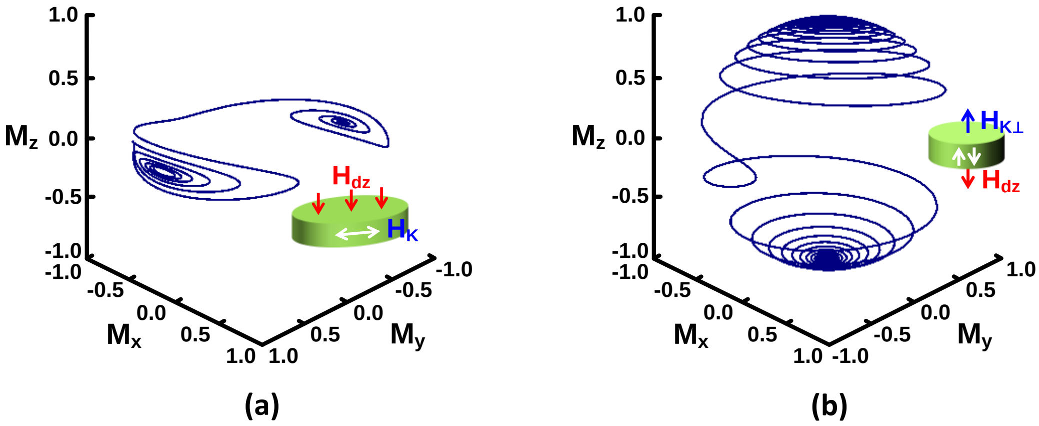


**Fig. 2.9 Dynamic spin motion simulated using macro-spin model. (a) In-plane magnetic anisotropy (easy axis = y-direction). (b) Perpendicular magnetic anisotropy (easy axis = z-direction).**

For IMA, thermal stability is primarily determined by shape anisotropy. The surface poles of magnet produce not only outward field but also counter field inside the magnet which acts against the magnetization thereby demagnetizing the magnet. This internal field is known as the demagnetizing field ($H_d$). Basically, $H_d$ along a short axis is stronger than along a long axis generating the magnetization along the longest axis. Therefore, shape alone can be a source of magnetic anisotropy. $H_d$ can be given by

$$H_d = -4\pi N_d M_s \tag{7}$$

where $N_d$ is the demagnetizing factor. Assuming a hexahedron-shaped magnet, $N_d$ values in x, y, and z direction can be calculated based on the dimension of a magnet in each direction (Typically, $N_{dx}+N_{dy}+N_{dz}=1$) [83]. In case of thin film elongated along x and y dimensions as shown in Fig. 2.9(a), $N_{dx}$ and $N_{dy}$ are very small compared to $N_{dz}$. Thus, $H_{dz}$ is the strongest and the magnetization tries to stay in the x-y plane resulting in in-plane magnetization. The shape anisotropy field, $H_{k,shape}$, is proportional to the difference between $N_{dx}$ and $N_{dy}$, and is governed by aspect ratio of magnet as follows:

$$H_{k,shape} = 4\pi(N_{dx} - N_{dy})M_s \tag{8}$$

Finally, the Δ of the IMA can be expressed as

$$\Delta_{IMA} = \frac{K_u V}{k_B T} = \frac{H_{k,shape} M_s V}{2k_B T} = \frac{2\pi(N_{dx} - N_{dy})M_s^2 V}{k_B T} \tag{9}$$

In terms of spin motion, as shown in Fig. 2.9(a), IMA shows limited trajectory in the z-direction. This indicates that IMA has to overcome a large $H_{dz}$ field, which attempts to keep the magnetization within the x-y plane, resulting in a large switching current.

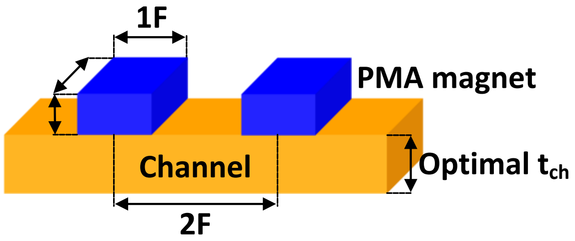
As an alternative to IMA, PMA has been extensively investigated recently to realize low current switching while maintaining sufficient thermal stability. As shown in Fig. 2.9(b), $H_{dz}$ assists the magnetization switching by partially canceling out the perpendicular anisotropy field ($H_{k\perp}$) resulting in a lower switching current. However, $H_{k\perp}$ must be larger than the $H_{dz}$ in order to maintain the orientation of the magnetization [59]. This can be achieved by using either high crystal anisotropy from $L1_0$-phase alloys (e.g. FePt, CoPt, FePd, etc) or interface anisotropy from thin CoFeB layer [84]-[86]. The effective perpendicular anisotropy field ($H_{k\perp eff}$) is determined by a difference between $H_{k\perp}$ and $H_{dz}$ as follows:

$$H_{k\perp eff} = H_{k\perp} - H_{dz} = 2K_{\perp} / M_s - 4\pi N_{dz} M_s \tag{10}$$

The resultant Δ of the PMA can be expressed as

$$\Delta_{PMA} = \frac{K_{\perp eff}V}{k_B T} = \frac{H_{k\perp eff}M_s V}{2k_B T} = \frac{(K_{\perp} - 2\pi N_{dz}M_s^2)V}{k_B T} \tag{11}$$

Note that the Δ of PMA is also affected by magnet dimensions due to $N_{dz}$. Therefore, thermal stability requirement for both IMA and PMA can be met by adjusting magnet dimensions according to equations (9) and (11).



| Parameters | Default value |
|---|---|
| Magnet dimensions, $W_m$×$L_m$×$t_m$ | 5nm×5nm×4nm |
| Demagnetizing factors ($N_{dx}$, $N_{dy}$, $N_{dz}$) | 0.31, 0.31, 0.38 |
| Magnet spin diffusion length, $\lambda_m$ | 4nm |
| Saturation magnetization, $M_s$ | 1.1×10$^3$A/m |
| Crystal anisotropy constant, $K_u$ | 3.15×10$^6$J/m$^3$ |
| Damping factor, $\alpha$ | 0.0055 |
| Polarization factor, $P$ | 0.5 |
| Magnet resistivity, $\rho_m$ | 17μΩ·cm |
| Channel dimensions, $W_{ch}$×$L_{ch}$×$t_{ch}$ | 5nm×10nm×6nm |
| Channel spin diffusion length, $\lambda_{ch}$ | 400nm |
| Channel resistivity, $\rho_{ch}$ | 2.35μΩ·cm |

**Table 2.3 Device parameters of PMA-based ASL for a 10 year retention time at 5nm technology node. F is the technology node (i.e. 5nm).**

In this work, we consider crystal anisotropy based PMA magnet, which utilizes a high $K_u$ (previously noted as $K_{\perp}$ for PMA) of specific materials for enhancing the thermal stability. Note that interface anisotropy based PMA requires further reduction in damping and a stronger interface anisotropy in order to be a viable contender in 5nm. Target parameters for the PMA magnet are shown in Table 2.3. The width and length of the magnet have been fixed as per the technology node (i.e. 5nm by 5nm). The thickness of the magnet is set as one spin diffusion length of the magnet material since a magnet thinner than its spin diffusion length can behave like a leaky polarizer, and cause incomplete spin polarization and relaxation in the input and output magnets, respectively [87]. Based on these magnet dimensions and the given $M_s$ value, the required $K_u$ of magnet was calculated to be $3.15\times10^6$ J/m$^3$ for a thermal stability of $69k_BT$ using equation (11).

### *2.2.4 Critical Spin Current for Magnet Switching*

The critical spin current ($I_{s,critical}$) to be provided for output magnet switching can be measured by macro-spin simulation based on LLG equation. Material parameters, magnet dimensions, temperature, and physical constants are first given as input parameters. The material parameters include $M_s$, $\alpha$, and $P$. $\alpha$ is the damping factor, which determines how fast the magnetization returns to the easy axis. $P$ is the polarization factor, which is estimated using the difference in the spin-dependent density of states (DOS). These material parameter values used in this work are listed in Table 2.3. Magnet dimensions are estimated in the previous section. Anisotropy field, which determines dynamic spin behavior, is a strong function of the magnet dimensions. Dynamic spin motion of the output magnet can be modeled with a unit vector of magnetization over time by assuming a nano-sized magnet as a macro-spin. At equilibrium temperature, thermal fluctuation induces randomly distributed initial angle between the magnetization vector ($\bar{M}$) and the easy axis. Note that, at a short switching pulse, spin precession dominates magnetization switching (i.e. precessional switching region: pulse width<3ns) and, thus, the initial position of magnetization vector has a significant influence on switching probability [88]. In this work, assuming 100% switching probability, very small initial angle (≈ 1.5°) estimated by agreement with measured data is considered to decide the initial position of the magnetization vector. When $V_{supply}$ is turned on, spin current density ($J_s$) generated from input magnet travels through channel and exerts spin torque to output magnet. Here, polarized spin direction depends on the magnetization of the input magnet ($\overline{M}_i$), which is represented as [0, 0, 1] assuming that the easy-axis is in the z-direction. This spin torque tries to flip the $\overline{M}$ in the output magnet against the $H_{k\perp eff}$. The $H_{k\perp eff}$ is mainly governed by a difference between $H_{k\perp}$ and $H_{dz}$, which can be denoted in vector notation over time as follows:

$$\overline{H}_{k\perp eff}(t) = [0,0,(2K_\perp / M_s)M_z(t)] - 4\pi M_s \cdot [N_{dx}M_x(t), N_{dy}M_y(t), N_{dz}M_z(t)] \tag{12}$$

The dynamics of $\overline{M}$ (t) is described by the LLG equation as follows:

$$\frac{1+\alpha^2}{\gamma} \cdot \frac{d\overline{M}}{dt} = -\overline{M} \times \overline{H}_{k\perp eff} - \alpha \cdot \overline{M} \times (\overline{M} \times \overline{H}_{k\perp eff}) + \frac{\hbar J_s}{2et_m M_s} \cdot \overline{M} \times (\overline{M} \times \overline{M}_i) \tag{13}$$

where $\gamma$ is the gyromagnetic ratio, $\hbar$ is the reduced Planck's constant, $e$ is the electron charge, and $t_m$ is the thickness of the magnet. For a $J_s$ exceeding the critical value, a dynamic precession is reinforced, which finally end up switching the magnetization vector to another energetically stable state. Based on the FO of 4 and switching time of 2ns, $I_{s,critical}$ for output magnet switching is measured as 51μA, which will be used to estimate $I_{c,critical}$ in the following section.

### *2.2.5 Spin Injection Ratio of ASL Gate*

The switching energy of ASL device is primarily a function of spin injection ratio ($I_s/I_c$). The spin signal ($\Delta V/I$) is proportional to the spin accumulation in the channel and can be analytically derived using the spin-diffusion equation [89]:

$$\frac{\Delta V}{I} = \frac{P_m^2 R_{s,m}^2}{2R_{s,m}\exp(L_{ch}/\lambda_{ch}) + R_{s,ch}\sinh(L_{ch}/\lambda_{ch})} \tag{14}$$

Here, $P$ is the spin polarization factor, $\lambda$ is the spin diffusion length, and $L$ is the channel length. $R_s$ is the spin resistance and can be expressed as

$$R_s = 2\rho\lambda/[(1-P^2)S] \tag{15}$$

where $\rho$ is the resistivity, and $S$ is the effective cross-sectional area. If the spin current $I_s$ generated by the charge current $I_c$ is sufficiently large, the transfer of spin angular momentum causes the magnetization of the detector magnet to reverse. When the $I_s$ is completely relaxed in the injector magnet, the $I_s$ flowing into the detector can be expressed as [90]

$$I_s = \frac{(\Delta V/I)I_c}{P_m R_{s,m}} \tag{16}$$

Eventually, by rewriting equation (16), the spin injection ratio can be derived as

$$\frac{I_s}{I_c} = \frac{\Delta V/I}{P_m R_{s,m}} = \frac{P_m R_{s,m}}{2R_{s,m}\exp(L_{ch}/\lambda_{ch}) + R_{s,ch}\sinh(L_{ch}/\lambda_{ch})} \tag{17}$$

As can be seen in (17), the spin injection ratio depends strongly on the material parameters, as well as the device geometry.

Using this analytical model, we can predict the spin injection ratio for ASL gate with different dimensions. The dimensions of the magnet are estimated based on the thermal stability requirement for a chip failure rate of 1 FIT as described in previous section. The local channel length is assumed as 10nm considering minimum space between two magnets, which is also short enough that additional buffers are not necessary. The optimal channel thickness is then decided for high spin injection ratio. Note that a thinner channel lowers resistance of input current path (i.e. magnet and channel stack on the input side) but a narrow channel results in a large spin signal loss due to spin scattering. Based on device dimensions and material parameters listed in Table 2.3, the $\Delta V/I$ and the spin injection ratio of the PMA-based ASL are estimated as 8Ω and 22.1% at room temperature, respectively. Finally, the critical charge current ($I_{c,critical}$) applied to the input magnet can be estimated by $I_{c,critical}=I_{s,critical}\cdot(I_c/I_s)$. The minimum value of $V_{supply}$ is also calculated based on the resistance of input current path. For a switching time of 2ns, the switching energy of a single ASL gate with FO=4 can be estimated as 3.5fJ using $E=V_{supply}\cdot I_{s,critical}\cdot(I_c/I_s)\cdot t_{sw}$.

## 2.3 ASL Interconnect Considerations

One critical issue of ASL devices is that the spin signal quickly attenuates over interconnects as spin torque has an exp(-d/λ) dependency on traveling distance *d*, and the characteristic spin diffusion length *λ*. Fig. 2.10 shows a steep decrease in spin injection ratio for a copper channel ($\lambda_{Cu}$=400nm). As such, an all spin-based interconnect scheme necessitates a large number of ASL buffers to transfer the spin signal, not only degrading system performance, but also resulting in a prohibitively high power overhead. To investigate this issue further, this section analyzes the power overhead of interconnect buffers and explores practical solutions for mitigating this issue.

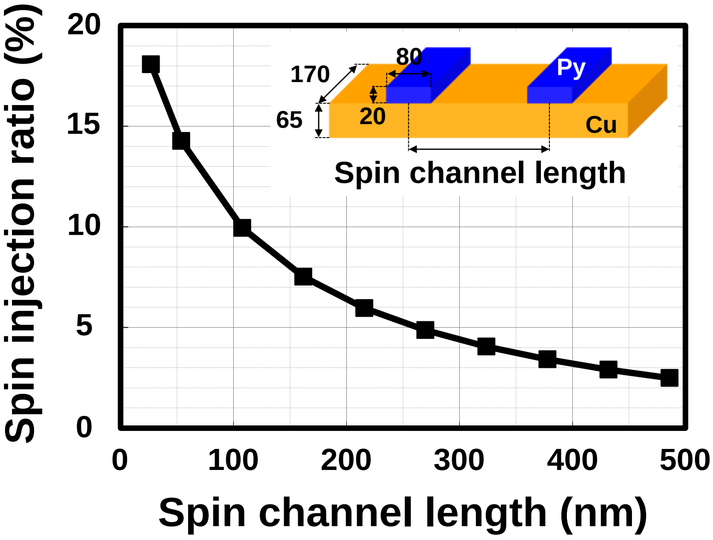


**Fig. 2.10 Spin injection ratio as a function of spin channel length for a Py+Cu device (i.e. Py based magnet with a Cu channel with device dimensions given in [39], inset).**

### *2.3.1 Power Overhead of Spin-Based Interconnect*

In order to measure the overhead of spin-based interconnect in ASL Core i7, it is necessary to count the number of ASL buffers needed. Interconnect density function based on Rent's rule is used to model the statistical distribution of wire lengths in a random logic block [80]

Region I: $1 \le l \le \sqrt{N}$ ,

$$i(l) = \frac{\alpha k}{2}\Gamma(\frac{l^3}{3} - 2\sqrt{N}l^2 + 2Nl)l^{2p-4} \tag{18}$$

Region II: $\sqrt{N} \le l \le 2\sqrt{N}$ ,

$$i(l) = \frac{\alpha k}{6}\Gamma(2\sqrt{N} - l)^3 l^{2p-4} \quad . \tag{19}$$

Here, $l$ is the interconnect length normalized to the gate pitch and $\alpha$ is defined as

$$\alpha = \frac{FO}{FO+1}, \tag{20}$$

where $FO$ is the average fanout of a logic gate. $k$ and $N$ were defined earlier as the average number of terminals per gate and the total number of gates in the processor, respectively. The $\Gamma$ parameter used in (18) and (19) is the normalization factor. We assume $k$=3.2 and $p$=0.6 as suggested in [91] for typical logic blocks. The number of gates $N$ can be estimated as we did in previous section. Since $k$ is approximately 3 (i.e. 3-terminal gate), we can assume that the representative logic gate to be a 2-input NAND gate comprising 4 CMOS transistors. From the specification that the logic part of a single core has 135 million transistors (0.54 billion of transistors for logic/4 cores=135 million), the number of its equivalent logic gates is calculated as 33.8 million (i.e. $N$=135 million transistors for logic of 1 core/4 transistors for an equivalent gate=33.8 million).With an ASL gate pitch of 10nm and an average $FO$ of 4, the wire length distribution for the random logic portion of the Core i7 processor can be plotted as shown in Fig. 2.11(a). The ASL buffer distribution, *buffer_count(l)*, that gives the expected number of ASL buffers for a wire with a length of $l$ is simply expressed as

$$buffer\_count(l) = quotient(l, L_{ch}) \cdot i(l) \tag{21}$$

where $L_{ch}$ is the buffer channel length, and *quotient(l,* $L_{ch}$*)* is the number of buffers for a wire length of $l$. Fig. 2.11(b) displays the cumulative distribution of ASL buffer count for a single processor core as a function of spin channel length.

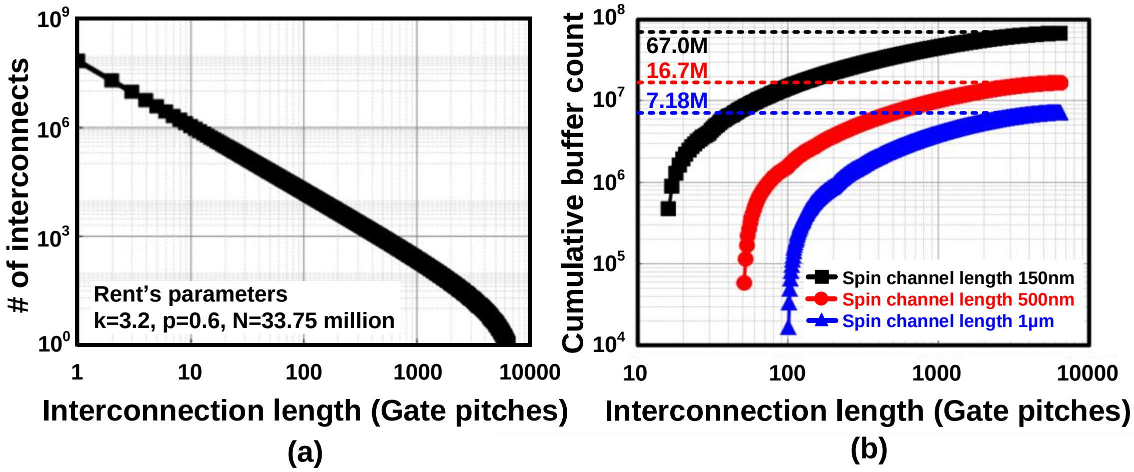


**Fig. 2.11 (a) Estimated wire length distribution for a Core i7 processor (single core, logic part only). (b) Cumulative buffer count distribution for different spin channel lengths.**

### *2.3.2 Optimization of Spin-Based Interconnect*

Spin channel length directly impacts interconnect power. For longer channel lengths, total number of ASL buffers is reduced but each buffer requires a higher input current to counteract the loss in spin current. Due to these two conflicting effects, an optimum spin channel length exists where the interconnect power is minimum. Fig. 2.12(a) shows the dependency of buffer count and critical charge current ($I_{c,critical}$) on Cu spin channel length indicating that the optimal spin channel length of Cu is 150nm. However, as estimated in Fig. 2.11, corresponding ASL buffer count is about 67 millions/core which amounts to the total logic device count of a single core. This simple back-of-the-envelope analysis reveals that interconnect buffers consume much more than the total chip power at the Cu channel based ASL Core i7 necessitating further optimization. Detailed analysis for power calculation is presented in the following section.

Novel channel materials with longer spin lifetime have been developed to overcome the loss in spin current and realize full potential of the ASL device. As described in Fig. 2.12(b), longer spin diffusion channel supports longer optimal channel length thereby reducing number of buffers and total power consumption. Single graphene layer (SLG) is the leading candidate among the materials which show exceptional spin transport characteristics. It has a spin diffusion length of 2μm at room temperature [41]. However, graphene-based spin valve devices require a tunnel barrier such as MgO inserted between SLG and ferromagnet in order to mitigate the impedance mismatch between graphene and ferromagnet.

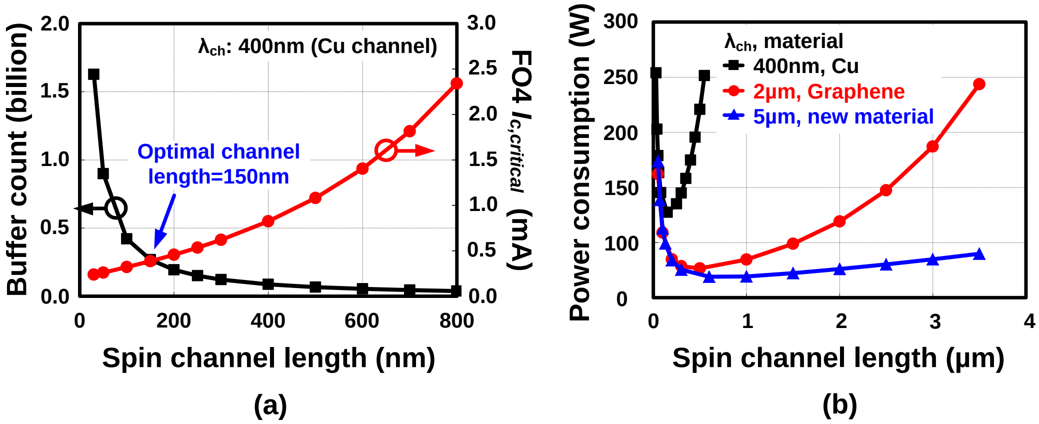


**Fig. 2.12 Spin diffusion length impact on power consumption. (a) Dependence of buffer count and critical input charge current on spin channel length for copper (4 cores, 25MHz). (b) New channel material such as graphene with a longer spin diffusion length enables longer spin channels thereby reducing the number of buffers required.**

## 2.4 System Level Power Comparison

Despite promising benefits and recent advances in spin-based devices, a system-level analysis is still necessary since ASL based system will have many unique features (e.g. majority logic, cascading effect of gates, spin-based interconnects, etc.) compared to conventional CMOS based designs. In order to verify the promises of ASL at the system level, this section presents a power comparison with CMOS using aforementioned realistic test vehicle, Intel's Core i7, considering various combinations of device parameters and power reduction schemes. This comparison study suggests the direction of optimization for ASL system to compete with CMOS-based system in terms of power consumption.

### *2.4.1 Power Calculation of ASL Based Processor*

The following equation was used to estimate the logic and interconnect power of ASL: (switching energy of a single device)×(clock frequency)×(device count). We assume that each ASL gate in the pipeline stage is sequentially pulsed by the clock to save unnecessary power consumption. Switching energy for core logic devices and interconnect buffers are estimated while considering the different spin injection ratio and $V_{supply}$ since core logic devices use 10nm spin channel length whereas interconnect buffers are placed at optimal intervals calculated in previous section. We chose 25MHz as the operating frequency for power comparison since frequencies much higher than this would make the comparison meaningless due to the extremely high ASL power. Although we did not perform a full energy-delay optimization for CMOS, the supply voltage was reduced to account for the lower operating frequency. An industrial 32nm process design kit was utilized for the schematic design and HSPICE simulation of CMOS gates. The switching time of an ASL device is 2ns by assuming 20 logic gates in a single pipeline stage (i.e. 40ns/20=2ns) [92]. The device count for the core logic can cut down by half using ASL, while the number of ASL interconnect buffers required was calculated depending on the channel material and optimal buffering interval.

In order to assess the advantage of ASL while considering both static and active power, we consider various power management schemes depending on the activity levels of the processor cores. Modern microprocessors such as Intel's Core i7 are capable of adjusting the voltage and frequency, gating off clocks, and shutting down cores all together, depending on the computation demand [93]. According to Intel's Core i7's datasheet, C0 state represents the highest power consumption mode where all four cores are switching, while C1 state is used for clock gating mode which draws static leakage power only. C6 state represents a power gating mode which can achieve the lowest static power consumption [94]. In this work, various idle power states of ASL Core i7 are considered to show the power savings under different ratios between active and static power consumption.

### *2.4.2 Activity Factor Between ASL and CMOS*

Our analysis so far shows that the static power savings of ASL is offset by the high switching energy due to the low spin injection ratio and the large number of buffers for interconnects. Another critical obstacle that has been largely overlooked by the community is ASL's 100% activity factor. As can be seen in Table 2.4, in a CMOS logic gate, the output only switches when the input changes. In other words, if the input remains constant, CMOS logic gates do not consume any dynamic power. Typically, CMOS gates in complex blocks have an activity factor much less than 10%. ASL on the other hand, has to evaluate every cycle regardless of whether the input data changes or not. This is equivalent to an activity factor of 100%. As shown in Table 2.4, ASL consumes $I_c$ when the clocked $V_{supply}$ is on. This is an inherent drawback of ASL independent of the material or device choices that will have to be mitigated with the help of auxiliary CMOS circuits.

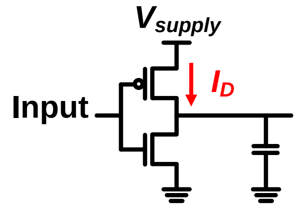
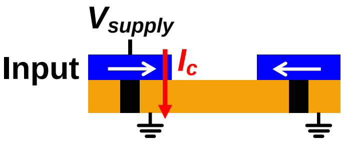
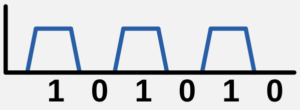
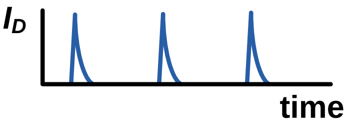
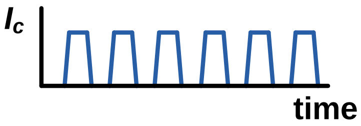
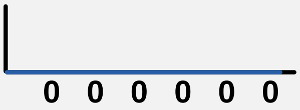
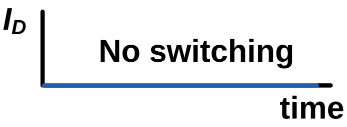
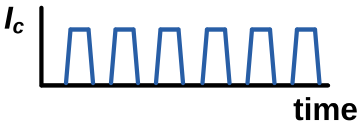
| | CMOS Inverter | ASL Inverter |
|---|---|---|
| Input sequence |  |  |
|  <br>High activity |  |  |
|  <br>Low activity |  |  |

**Table 2.4 Activity factor comparison between CMOS and ASL. CMOS gates only consume power when the input signal switches while ASL gates consume power every cycle irrespective of the input pattern. The inherently high activity factor of ASL is a critical issue.**

### *2.4.3 Power Comparison: ASL vs. CMOS*

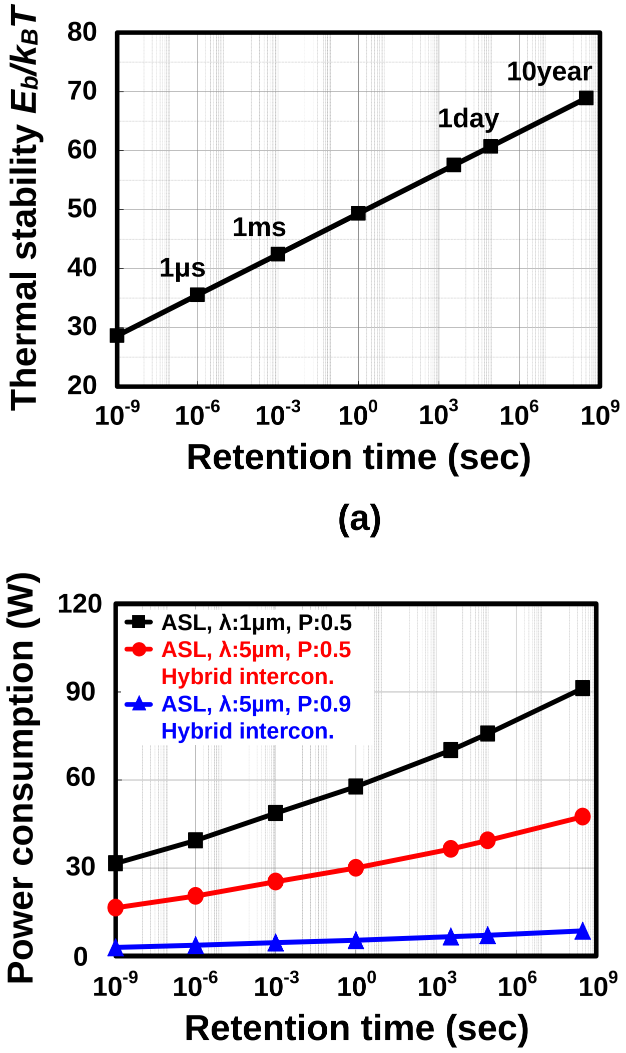
Table 2.5 presents the power comparison between CMOS based and various ASL based Core i7 implementations. We estimate the power consumption of future ASL technologies considering improvement in the magnet and channel properties to provide a design guideline for the material community. All ASL devices are assumed to have a minimum feature size of 5nm. That is, the minimum magnet width and the minimum gate-to-gate distance are both 5nm. Ideally, the comparison between CMOS and ASL should be done at the same technology node (i.e. 5nm CMOS versus 5nm ASL). However, there were several practical issues that prevented us from doing this. For instance, the supply voltage, transistor parameters, threshold voltage, operating frequency for 5nm CMOS are largely unknown at this point so the CMOS results will not be very accurate. As a compromise, we chose to compare 5nm ASL to 32nm CMOS, hoping that this will give readers at least a sense of how the power consumption of ASL compares to that of today's microprocessors. Note that all ASL implementations have 270 million devices for the core logic which is half the number of devices compared to an equivalent CMOS implementation.

Charge-based interconnects can be utilized for long wires to mitigate the high power consumption and performance limitation of spin-based interconnects. So we also considered a hybrid spin-charge interconnect scheme in which interconnects longer than a certain length (e.g. 5µm) are replaced with charge based interconnect. The minimum wire length for switching to charge will depend on the conversion overhead as well as the performance and power benefits. The total number of interconnect buffers were estimated based on the specific type of channel material and interconnect scheme (e.g. spin-only or hybrid). Another possible method for reducing ASL power is to deliberately lower the thermal stability to the point of guaranteeing non-volatility for a single logic operation cycle (e.g. 1/25 microsecond). This can be achieved by either shrinking the volume of the magnet or using a lower $K_u$ material. Fig. 2.13(a) plots the minimum thermal stability required for an ASL based core i7 processor to satisfy a chip failure rate of <0.01% for different target retention times. Results in Fig. 2.13(b) show the clear tradeoff between retention time and power consumption for different spin diffusion lengths, polarization factors, and interconnect schemes.

| Parameter | | 32nm CMOS | ASL, λ:1μm | ASL, λ:5μm | | | |
|---|---|---|---|---|---|---|---|
| | | | | - | Hybrid interconnect (>5μm) | | |
| | | | | | - | High polarization | |
| | | | | | | - | Short retention |
| Technology node | | 32nm | 5nm | 5nm | 5nm | 5nm | 5nm |
| Device count | | 540 million | 357 million | 299 million | 280 million | 280 million | 280 million |
| Activity factor | | 5% | 100% | 100% | 100% | 100% | 100% |
| Channel | $\lambda_{ch}$ | - | 1μm | 5μm | 5μm | 5μm | 5μm |
| | $L_{ch}$ Core/Buffer | - | 10nm/400nm | 10nm/1μm | 10nm/1μm | 10nm/1μm | 10nm/1μm |
| Magnet | $P_{FM}$ | - | 0.5 | 0.5 | 0.5 | 0.8 | 0.9 |
| | Retention time/$K_u$ | - | 10year/3.15 | 10year/3.15 | 10year/3.15 | 10year/3.15 | 1μs/1.75 |
| | Critical $I_s$ (FO4) | - | 51μA | 51μA | 51μA | 51μA | 33μA |
| Device | $I_s/I_c$ Core/Buffer | - | 24.3%/11.2% | 24.5%/8.2% | 24.5%/8.2% | 39.2%/12.6% | 44.3%/16.5% |
| | $V_{supply}$ Core/Buffer | - | 12.8mV/27.8mV | 12.7mV/37.8mV | 12.7mV/37.8mV | 5.5mV/17.1mV | 2.84mV/7.61mV |
| Power (@25MHz) | C0 Active/Static/Total | 0.05/3.70/3.75W | 91.2/0.00/91.2W | 69.3/0.00/69.3W | 47.4/0.00/47.4W | 13.2/0.00/13.2W | 3.68/0.00/3.68W |
| | C1 Active/Static/Total | 0.01/3.70/3.71W | 22.8/0.00/22.8W | 17.3/0.00/17.3W | 11.9/0.00/11.9W | 3.30/0.00/3.30W | 0.92/0.00/0.92W |
| | C6 Active/Static/Total | 0.01/1.00/1.01W | 22.8/0.00/22.8W | 17.3/0.00/17.3W | 11.9/0.00/11.9W | 3.30/0.00/3.30W | 0.92/0.00/0.92W |

*λ: spin diffusion length, *L*: channel length, *P*: spin polarization, $K_u$: crystal anisotropy ($10^6 J/m^3$), $I_s$: spin current, $I_c$:charge current

**Table 2.5 ASL vs. CMOS power comparison under an operating frequency of 25MHz. (C0: all four cores active, C1: one core active while three cores are clock gated, C6: one core active while three cores are power gated)**



(a)

(b)

**Fig. 2.13 Trade-off between ASL retention time and switching power. (a) Thermal stability versus retention time (0.01% chip failure rate assumed). (b) Power consumption versus retention time for various ASL devices.**

Material and device parameters of ASL to meet a system requirement of 10 years of retention and 25MHz of operating frequency are listed separately for the core devices and interconnect buffers in Table 2.5. Total system power for ASL is estimated for different power down modes (i.e. C0, C1, and C6) to evaluate the power saving benefits under different active to static power ratios. The power consumption values are listed in the bottom part of Table 2.5 for different operating mode, but to provide better intuition, we also provide a bar chart version of the same data in Fig. 2.14 showing the core logic power and interconnect buffer power separately.

For C0 state where all four cores are actively switching, ASL with $\lambda$=1μm consumes excessively high core and interconnect power compared to its CMOS counterpart as shown in Fig. 2.14 (above). The interconnect power can be reduced by employing a longer spin diffusion channel material ($\lambda$=5μm) and a hybrid interconnect scheme where signals are converted to charge domain for interconnects longer than 5μm. Note that the impact of longer spin diffusion channel on core logic power is negligible since the interconnect length between local ASL gates is too short to benefit from the longer spin diffusion. Meanwhile, a high polarization factor ($P$=0.8~0.9) material is considered to enhance the spin injection ratio resulting in significant power savings in both core logic and interconnect. Finally, we show another future scenario in which the retention time is traded off (down to 1μs) which eventually makes the ASL power comparable to CMOS.

In Fig. 2.14 (below), power consumption numbers are shown for a C1 operating mode where only a single core is active while the other three cores are in a clock gated mode and hence dissipating leakage power (i.e. idle mode). ASL is expected to show more promise in this case since the portion of leakage power has gone up for the CMOS implementation. Indeed, our estimation results show that ASL can achieve a power level comparable to CMOS even without sacrificing retention time or requiring a very high polarization factor of 0.9.

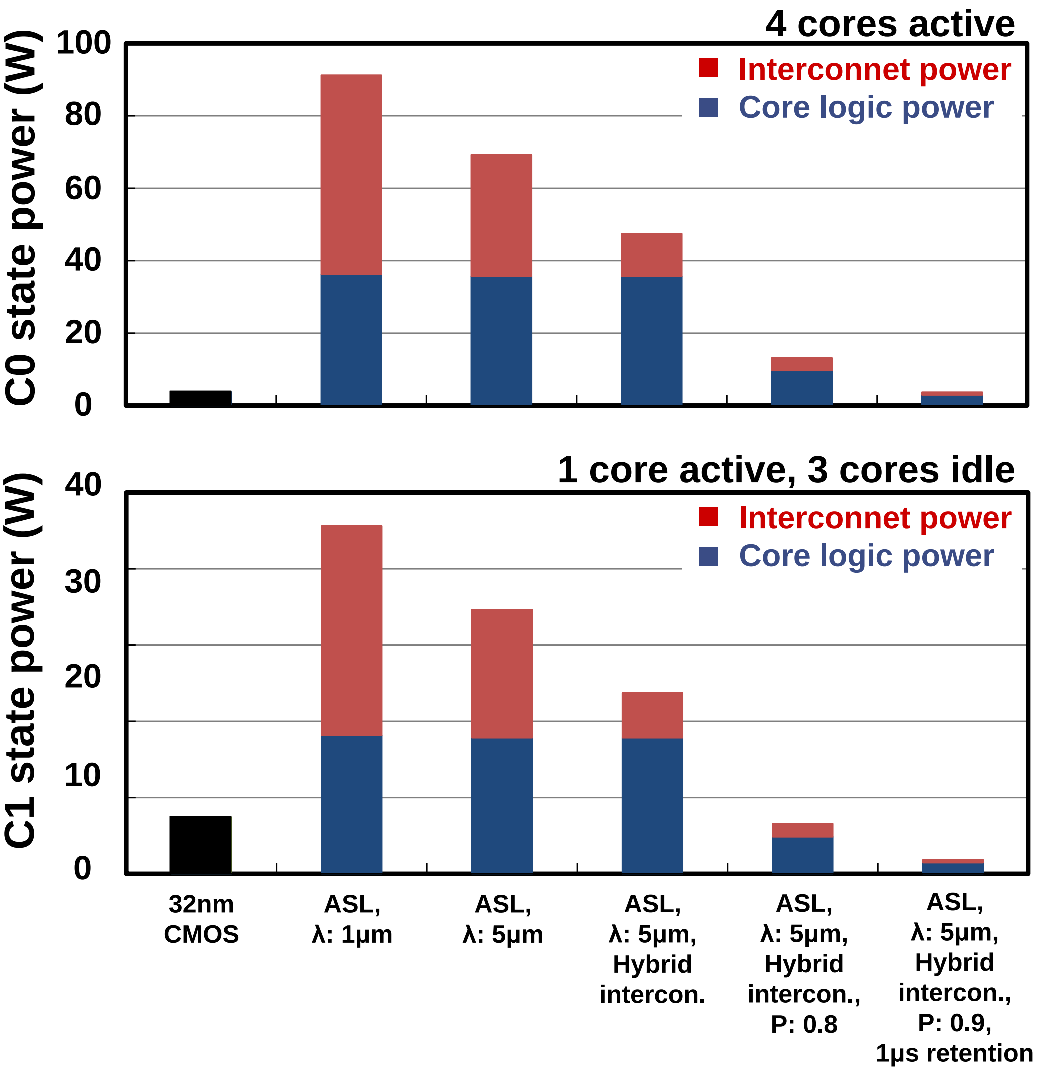


**Fig. 2.14 Power comparison between CMOS and future ASLs. (Above) C0 state power (all four cores active). (Below) C1 state power (one core active, other 3 cores clock gated).**

## 2.5 Limitations of This Work and Future Directions

Due to limited experimental data available and the speculative nature of this type of research, our benchmarking study had to rely on many assumptions and workarounds. Here, we summarize some of the known limitations of this work which could be addressed in future work.

- Our estimation takes into account only the logic part of the processor to evaluate the power consumption for logic operation.
- We assume a 5nm technology for ASL, which is beyond the limit of today's lithography tools. Recently developed gas phase synthesis methods could enhance the patterning resolution by direct placement of nanoparticles [95]. Variation in the magnet dimensions in extremely scaled technologies will have significant impact on the thermal stability and critical switching current of spin based devices. Further studies are necessary to assess device performance in the presence of dimensional variability and material imperfections.
- Physical parameters in this work are set based on room temperature. However, the worst case operating temperature in many integrated circuits (ICs) is generally higher [96]. This may result in a higher magnet resistivity, lower spin polarization and shorter spin diffusion length [89].
- We assume that each ASL gate receives a pulsed clock that is delayed from one logic stage to the next. By doing so, we can assume that static power is consumed only during the short computation period.

- Our power estimation is based on the device count for core and interconnect circuits, and their individual energy dissipation. The power and area overhead associated with clocking the ASL gates is not considered in this work for a clearer picture. Several recent studies have shown that clocking power can be made negligible compared to the intrinsic ASL power by utilizing clocking transistors operating in deep triode mode shared between multiple ASL devices.
- A tunneling barrier may be required for good impedance matching between a metallic magnet and graphene channel. This may necessitate a higher voltage that could result in a higher overall energy consumption for ASL.
- In the hybrid interconnect scheme, we assume that the overhead for spin-to-charge and charge-to-spin conversion is negligible compared to the buffer power overhead for long wires. Additional work will be required for an accurate estimation of the spin-to-charge and charge-to-spin conversion overhead.
- Our device count methodology for ASL system is based on a drop-in replacement scenario; that is, each CMOS gate is substituted with an equivalent ASL gate. However, certain logic functions may be able to take advantage of the inherent majority function of ASL which could open up new design methodologies for ASL [43], [75].

## 2.6 Chapter Summary

A case study of ASL technology on a hypothetical Intel Core i7 processor was presented in this chapter where the key advantages of ASL based computing such as zero static power, lower device count and lower supply voltage were highlighted. Technical barriers associated with spin devices such as low spin injection, limited spin diffusion length, and intrinsically high activity factor, were also extensively discussed. It is our sincere hope that this work will provide the general engineering community with a more honest and clearer picture of spintronics technology from a system level perspective.

# Chapter 3 Scaling Behavior of In-Plane and Perpendicular MTJ Based STT-MRAMs

Spin transfer torque magnetoresistive random access memory (STT-MRAM) has been gaining interest as an alternative cache memory due to unique properties such as nonvolatility (i.e. zero static power), compact bit-cell size, and high endurance [13], [53]. Moreover, STT-MRAM can outperform SRAM for large caches (e.g. >64Mb) since its higher bit-cell density reduces the global interconnect delay which is the critical performance bottleneck in last level caches [97]. With the successful integration of magnetic tunnel junctions (MTJs) into advanced CMOS processes, STT-MRAM is being accepted as a viable embedded nonvolatile working memory for next-generation microprocessor systems [98]-[100].

One of the key objectives of STT-MRAM design has been to reduce the switching current of the MTJ without compromising nonvolatility. A STT-MRAM bit-cell consists of an access transistor and an MTJ. Typically, a lower switching current for the MTJ allows a smaller access transistor, which translates into a faster, denser, and cheaper memory. Recent efforts for lowering switching current have resulted in MTJs with different anisotropy sources: namely, shape anisotropy-based in-plane MTJ (IMTJ), crystal anisotropy-based perpendicular MTJ (c-PMTJ), and interface anisotropy-based perpendicular MTJ (i-PMTJ).

Typically, IMTJ has to overcome a large out-of-plane demagnetizing field ($H_{dz}$), which attempts to keep the magnetization within the plane, giving rise to a large switching current. As an alternative to IMTJ, PMTJ can provide a lower switching current since $H_{dz}$ assists STT switching by partially canceling out the perpendicular anisotropy field ($H_{k\perp}$). However, $H_{k\perp}$ must be greater than $H_{dz}$ in order to maintain magnetization in the perpendicular direction [101]. This $H_{k\perp}$ can be achieved using $L1_0$-phase alloys such as FePt and FePd with high crystal anisotropy (e.g. $K_u > 10^6$ J/m$^3$) or using an ultra-thin CoFeB layer with interface anisotropy ($K_i$). We refer to these two perpendicular anisotropy devices as c-PMTJ and i-PMTJ, respectively [34], [84].

In case these MTJs are compared only with intrinsic anisotropy behavior, PMTJs are supposed to have smaller $I_c$ than IMTJ for a given thermal stability since IMTJ always has to overcome additional $H_{dz}$. However, when the practical considerations such as material parameters and dimensional constraints are taken into account, it is still unclear which of the three MTJ technologies will prevail at the extremely scaled technology nodes such as 7nm [102].

To answer this open ended question, this work presents a comprehensive study on the scalability of various MTJ devices using a scalable MTJ SPICE model and ITRS predicted transistor parameters. Our model is specifically designed for performing scaling studies as it incorporates dimension dependent effective anisotropy field ($H_{keff}$) into the LLG equation, which instantly reflects dimensional changes into MTJ parameters [103]. To our knowledge, this was not achieved in prior works [97], [104], [105]. Then, we provide the detailed scaling scenarios based on realistic MTJ dimensions and material parameters under the same degree of data retention failure and read disturbance. A scaling roadmap for critical performance metrics such as write and read delays is projected down to the 7nm node by combining the SPICE MTJ model with ITRS predicted transistors. Finally, possible limitations and requirements for each MTJ technology are presented.

## 3.1 A Technology-Agnostic MTJ SPICE Model

A key aspect of evaluating STT-MRAM technology is the development of a scalable MTJ compact model which can be used to incorporate realistic variability effects across different technology nodes. Several SPICE compatible MTJ models have been reported in the past [106]-[108] to fulfil this goal. In [106], MTJ behaviors were emulated with SPICE subcircuits (e.g. bistable circuit, curve fitting circuit) based on empirical input parameters such as thermal stability (Δ), parallel and antiparallel resistance ($R_P$, $R_{AP}$), and the critical switching current ($I_c$). In order to capture realistic spin dynamics, the models in [107], [108] implemented the LLG equation using built-in SPICE elements such as resistors, capacitors, and voltage-/current-dependent voltage/current sources, considering physical parameters such as $H_{keff}$, MTJ dimensions, and material parameters. However, these models lack the flexibility for studying the scalability of various STT-MRAM designs since they still relied on a pre-calculated $H_{keff}$ value which is a function of the MTJ width, length, and thickness dimensions. For a compact model to be useful in evaluating STT-MRAM circuits across different technologies, it has to be scalable for future nodes and at the same time be fully-compatible with SPICE. To satisfy these requirements, we propose a scalable physics-based SPICE MTJ model with user-defined dimensional and material parameters. The main improvements of our model compared to previous SPICE MTJ models are summarized as follows.

- We provide a self-contained MTJ model that comprehends anisotropy, STT switching, TMR, and temperature effect, which is reconfigurable using user-defined input parameters.
- Our model generates $H_{keff}$ for all types of anisotropy sources such as shape, crystal, and interface, which makes it possible to simulate both in-plane and perpendicular MTJs in *any* given technology node.
- Spin dynamics computed by incorporating the dimension-dependent $H_{keff}$ into the LLG equation, instantly reflects dimensional changes in the MTJ performance, which enables more accurate scalability and variability analyses.
- The stochastic nature of the magnetization switching is captured by changing the initial magnetization angle. The switching distribution can be easily obtained using this simple approach.
- The temperature dependency of various material parameters is included considering internal Joule heating during the switching process.

### *3.1.1 MTJ Physics to Be Modeled*

To reproduce a realistic MTJ behavior, the model has to incorporate key physics such as magnetic anisotropy, STT switching, temperature effect, and TMR. This subsection provides a brief explanation on those concepts to be included in our compact model.

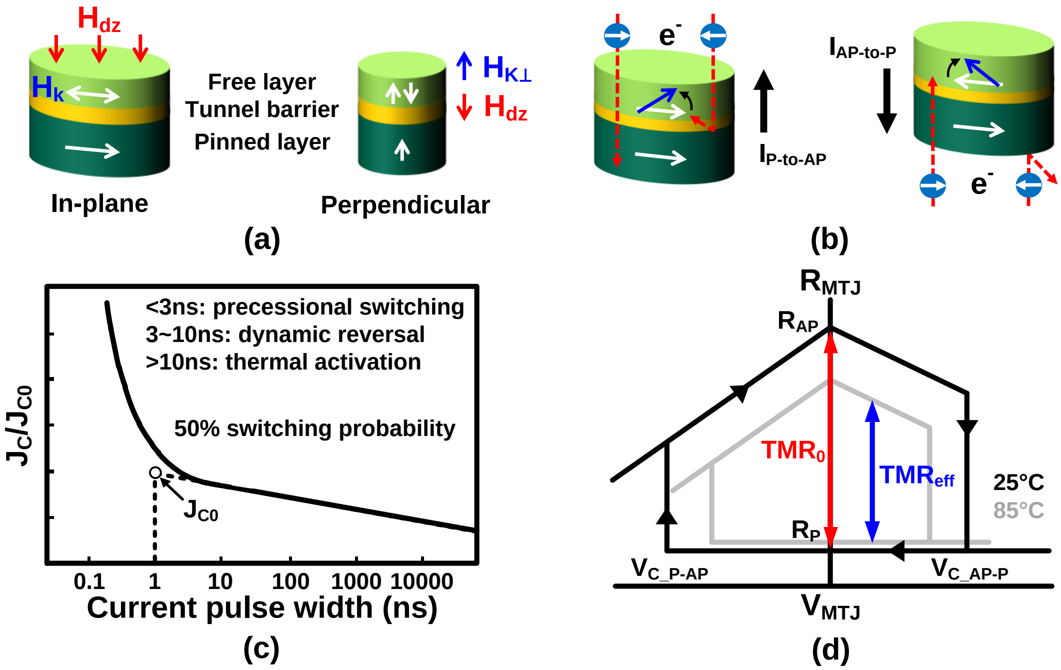


**Fig. 3.1 Basic MTJ characteristics (a) In-plane and perpendicular magnetic anisotropy. (b) Bi-directional STT switching. (c) Critical switching current as a function of pulse width. (d) Temperature-dependent R-V hysteresis curve.**

Several MTJ options exist depending on the physical origin of magnetic anisotropy (MA): IMTJ, c-PMTJ, and i-PMTJ. For IMTJ, the origin of shape anisotropy is the demagnetizing field ($H_d$) which is stronger along the axis with a shorter dimension. As a result, the magnetization ($M$) has a tendency to align with the longest axis giving rise to shape-dependent magnetic anisotropy. The free layer of the IMTJ can be regarded as an elongated thin film with the shortest axis being in the z-direction. Here, *M* stays in the x-y plane as shown in Fig. 3.1 (a). The shape anisotropy field ($H_{k,shape}$) can be expressed as:

$$H_{k,shape} = 4\pi(N_{dx} - N_{dy})M_s. \tag{1}$$

where $M_s$ is the saturation magnetization and $N_d$ is the geometry-dependent demagnetizing factor. However, the IMTJ has to overcome a large $H_{dz}$ resulting in a large switching current. PMTJ on the other hand can provide a lower switching current compared to IMTJ since $H_{dz}$ assists the magnetization switching by partially canceling out the perpendicular anisotropy field ($H_{k\perp}$) as shown in Fig. 3.1(a). However, in order to maintain the proper orientation of the *M*, $H_{k\perp}$ must overpower $H_{dz}$. This can be achieved by using either a high crystal anisotropy ($K_u$) using L1$_0$-phase alloys (e.g. CoPt, FePd) or interface anisotropy ($K_i$) using a CoFeB layer thinner than its critical thickness ($t_c$). The effective perpendicular anisotropy field ($H_{k\perp eff}$) is given by:

$$H_{k\perp eff} = H_{k\perp} - H_{dz} = 2K_\perp / M_s - 4\pi N_{dz} M_s. \tag{2}$$

For c-PMTJ, $K_\perp$ is equivalent to $K_u$ of the specific material. For i-PMTJ, $K_\perp$ is replaced with $K_i/t_F$ ($=2\pi M_s^2 t_c/t_F$) where $t_F$ is the free layer thickness.

During STT switching, bi-directional spin-polarized electrons exert spin torque to the free layer and induce magnetization switching in the desired direction as shown in Fig 3.1(b). If we treat the free layer as a single magnetic domain, the spin dynamics can be characterized with a time-varying unit magnetization vector given as $\bar{M}(t)$=[$M_x(t)$, $M_y(t)$, $M_z(t)$] [29]. When a switching current density (*J*) is applied to the MTJ, the spin-polarized current exerts spin torque to flip $\bar{M}$ against $H_{keff}$. Here, the spin direction of polarized current depends on the magnetization of the pinned layer $\bar{M}_p$. Since $H_{keff}$ of an IMTJ is mainly governed by $H_d$, its magnetization vector can be denoted as follows [83]:

$$\overline{H}_{k=eff}(t) = -4\pi M_s[N_{dx}M_x(t), N_{dy}M_y(t), N_{dz}M_z(t)]. \tag{3}$$

On the other hand, $H_{keff}$ of a PMTJ is the combination of $H_{k\perp}$ and $H_d$ so its vector notation is given by

$$\overline{H}_{k\perp eff}(t) = [0,0,(\frac{2K_\perp}{M_S})M_z(t)] - 4\pi M_s[N_{dx}M_x(t), N_{dy}M_y(t), N_{dz}M_z(t)]. \tag{4}$$

The dynamics of $\bar{M}(t)$ is generally described by the LLG equation with $H_{Keff}$ incorporated as follows:

$$\frac{1+\alpha^2}{\gamma}\cdot\frac{d\overline{M}}{dt} = -\overline{M}\times\overline{H}_{keff} - \alpha\cdot\overline{M}\times(\overline{M}\times\overline{H}_{keff}) + \frac{\hbar PJ}{2et_F M_s}\cdot\overline{M}\times(\overline{M}\times\overline{M}_p). \tag{5}$$

where $\gamma$ is the gyromagnetic ratio, $\alpha$ is the damping constant, $\hbar$ is the reduced Planck's constant, $P$ is the spin polarization factor, and $e$ is the electron charge. When the switching current exceeds the critical value, the dynamic precession motion overcomes $H_{keff}$ and flips the magnetization to the opposite stable state.

The transient behavior of an MTJ is non-deterministic due to the random thermal field which causes the magnetization vector to deviate from the easy axis by an angle determined by the MTJ temperature. Moreover, at long current pulses (i.e. >10ns), an increase in internal temperature due to Joule heating excites the thermal field thereby reducing the switching current as shown in Fig. 3.1(c). For fast precessional switching, which is more relevant in today's high speed STT-MRAM, the stochastic behavior can be captured using a switching probability ($P_{sw}$) as a function of the critical initial angle ($\theta_c$), which can be described as [109]:

$$P_{sw} = 1 - \int_0^{\theta_c} \frac{\sin\theta \exp(-\Delta \sin^2\theta)}{\int_0^{\frac{\pi}{2}} \sin\theta \exp(-\Delta \sin^2\theta) d\theta} d\theta,\ \Delta = \frac{E_b}{k_B T}. \tag{6}$$

where $T$ is the temperature, $E_b$ is the energy barrier, $k_B$ is the Boltzmann constant. Material parameters related to device performance also have a temperature dependency as follows [110]:

$$M_s(T) = M_{s0}(1 - T/T_c)^{\beta}. \tag{7}$$

$$P(T) = P_0(1 - \alpha_{sp} T^{3/2}). \tag{8}$$

where $M_{s0}$ and $P_0$ are the saturation magnetization and the polarization factor at absolute zero temperature, $T_c$ is the Curie temperature, $\beta$ and $\alpha_{sp}$ are the material-dependent constants.

An MTJ can be considered as a voltage-controlled variable resistance represented by the resistance-voltage (R-V) hysteresis curve shown in Fig. 3.1(d). The resistance ratio at a zero bias ($\mathrm{TMR}_0$) is defined as ($\mathrm{R}_{\mathrm{AP}}$-$\mathrm{R}_{\mathrm{P}}$)/$\mathrm{R}_{\mathrm{P}}$. Since the TMR depends on temperature and bias voltage, $\mathrm{TMR}_{\mathrm{eff}}$ is a better measure to evaluate read/write performances. The voltage and temperature dependency of TMR can be captured using the temperature-dependent polarization equation [111]:

$$TMR(T,V) = \frac{2P_0^2(1-\alpha_{sp}T^{3/2})^2}{1-P_0^2(1-\alpha_{sp}T^{3/2})^2} \cdot \frac{1}{1+(V/V_0)^2}. \tag{9}$$

Here, $V_0$ is a fitting parameter. Once RA value is determined, $\mathrm{R}_{\mathrm{P}}$ and $\mathrm{R}_{\mathrm{AP}}$ can be calculated by considering MTJ area and TMR.

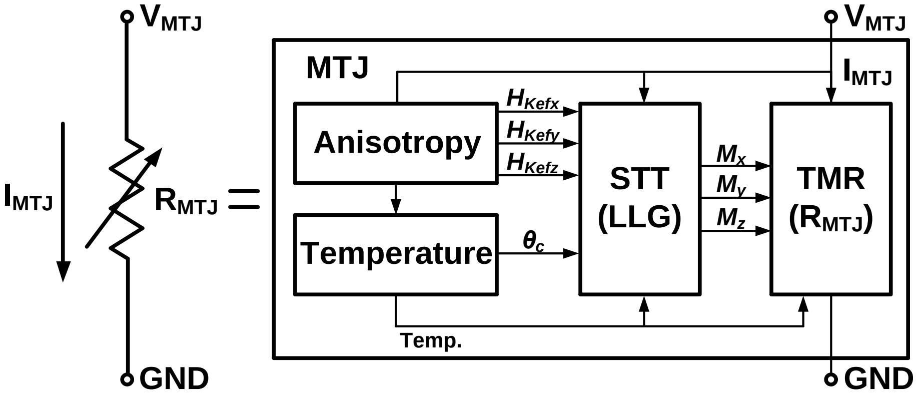


**Fig. 3.2 Simulation framework of the proposed MTJ model.**

### *3.1.2 Model Framework and Implementation*

The physical behaviors of a real MTJ were recreated using four dedicated SPICE subcircuits: namely, anisotropy, STT, TMR, and temperature subcircuits as shown in Fig. 3.2 [103]. Once the type of MTJ is selected, the anisotropy circuit generates $H_{keff}$ as derived in (3) and (4) for the given MTJ dimensions and material parameters. Meanwhile, the temperature circuit estimates $\theta_c$ at a given temperature as well as the switching probability given in (6), which sets the initial position of $\bar{M}$.

When a bias voltage ($V_{MTJ}$) is applied to the MTJ, a charge current ($I_{MTJ}$) passes through the MTJ. The $I_{MTJ}$ fed to the STT circuit generates a spin-polarized current and triggers dynamic spin motion returning x, y, and z coordinates of the time-varying vector $\bar{M}$. The Cartesian coordinates are converted to spherical coordinates by the TMR circuit generating a relative angle between the free and pinned layers to determine the MTJ resistance ($R_{MTJ}$). The $I_{MTJ}$ is also an input to the temperature circuit to estimate the increase in internal temperature due to Joule heating. The updated temperature is fed back to the STT and TMR circuits modifying material parameters that have a temperature dependency as given in (7) and (8). Table 3.1 shows input knobs which provide sufficient flexibility to explore various aspects of MTJ switching behavior.

| Input | Description | Remark |
|---|---|---|
| $W$ | Free layer width | Δ dependent |
| $L$ | Free layer length | Δ dependent |
| $t_F$ | Free layer thickness | Δ dependent |
| $\alpha$ | Magnetic damping factor | Material related |
| $M_{s0}$ | Saturation magnetization, 0K | Material related |
| $P_0$ | Polarization factor, 0K | Material related |
| $K_u$ | Crystal anisotropy constant | for c-PMTJ |
| $t_c$ | Critical thickness | for i-PMTJ |
| $T_0$ | Initial temperature | Ambient |
| $P_{sw}$ | Switching probability | by initial angle |
| *RA* | Resistance-area product | Measured data |
| *asym* | Bidirectional $I_c$ asymmetry | Measured data |
| *MA* | In-plane/Perpendicular selection | 0/1 |
| *State* | Parallel/Anti-parallel selection | 0/1 |

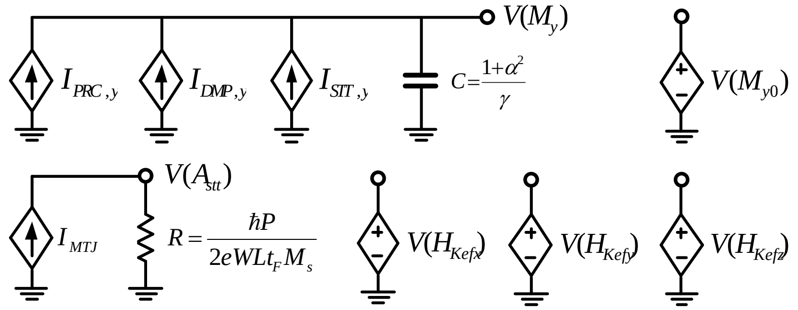
**Table 3.1 User-defined input parameters of the proposed MTJ model.**

In terms of circuit implementation, the differential behavior of $\bar{M}$ can be captured using a capacitor with voltage-dependent current sources connected in parallel, which emulates an incremental charge build-up over time in the capacitor: $I=C\cdot dV/dt$. In Fig. 3.3, three current sources represent the precession, damping, and spin torque terms in the LLG equation, and their vector cross product can be rewritten into linear forms as described in the SPICE script. $M_{y0}$ is the additional node to set the initial angle in case of consecutive switching. To solve a three-dimensional LLG equation, separate circuits for x, y, and z coordinates are implemented in the same way. The anisotropy and TMR circuits are simply implemented with SPICE parameters and voltage sources, while the temperature circuit uses a distributed RC line model which emulates the heat diffusion equation as suggested in [108].

**Numerical form:**

$$\frac{1+\alpha^2}{\gamma}\cdot\frac{d\overline{M}}{dt} = \underset{\text{Precession}}{-\overline{M}\times\overline{H}_{Keff}} \underset{\text{Damping}}{-\alpha\cdot\overline{M}\times(\overline{M}\times\overline{H}_{Keff})} + \underset{\text{Spin torque}}{A_{stt}\cdot\overline{M}\times(\overline{M}\times\overline{M}_P)},\ A_{stt}=\frac{\hbar PJ}{2et_F M_s}$$

**Circuit implementation (y-coordinate):**



**HSPICE script (y-coordinate):**

```
C_My        My  0   '(1+α^2)/γ'
G_dMy_prc   0   My  cur='-(v(Mz)·v(HKefx)-v(HKefz)·v(Mx))'
G_dMy_dmp   0   My  cur='-α·(v(Mz)·(v(My0)·v(HKefz)-v(HKefy)·v(Mz))-(v(Mx)·v(HKefy)-v(HKefx)·v(My0))·v(Mx))'
G_dMy_stt   0   My  cur='v(Astt)·(v(Mz)·(v(My0)·Mpz-Mpy·v(Mz))-(v(Mx)·Mpy-Mpx·v(My0))·v(Mx))'
E_My0       My0 0   vol='v(My)'  max='cos(v(θc))' min='-cos(v(θc))'
```

**Fig. 3.3 SPICE implementation of LLG equation (only y-coordinate shown here for simplicity).**

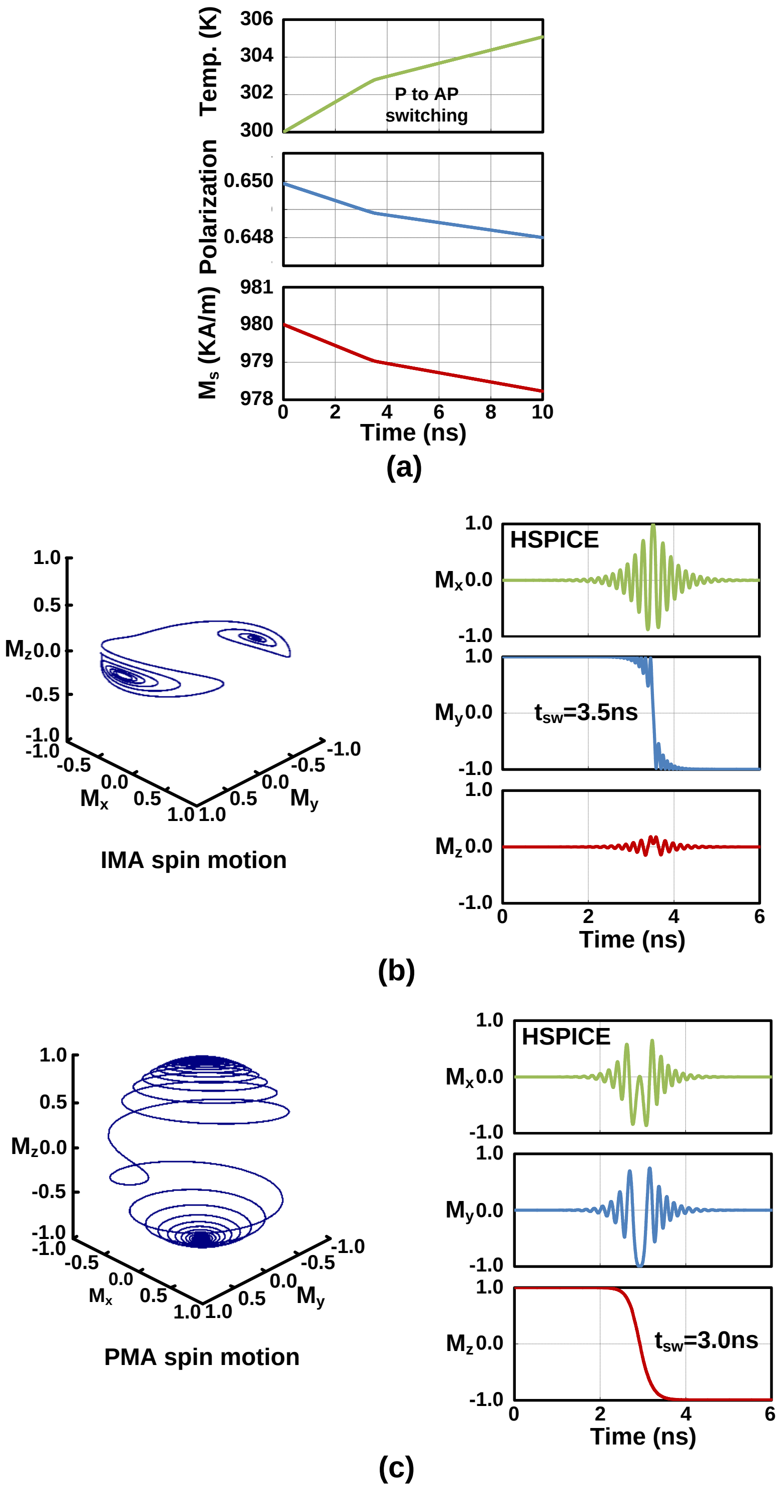


**Fig. 3.4 MTJ switching dynamics verification. (a) Temperature dependency of material parameters during a switching event. (b) Dynamic spin motion for in-plane MTJ. (c) Dynamic spin motion for perpendicular MTJ.**

### *3.1.3 Model Verification*

In order to verify the accuracy of the model, simulation results were compared with experiment data. In Fig. 3.4, the simulated dynamic spin motions of IMTJ and PMTJ are presented alongside the temperature dependency of several material parameters. The results show a clear contrast between the magnetization trajectories of IMTJ and PMTJ, which was not observed using the previous models [106]-[108].

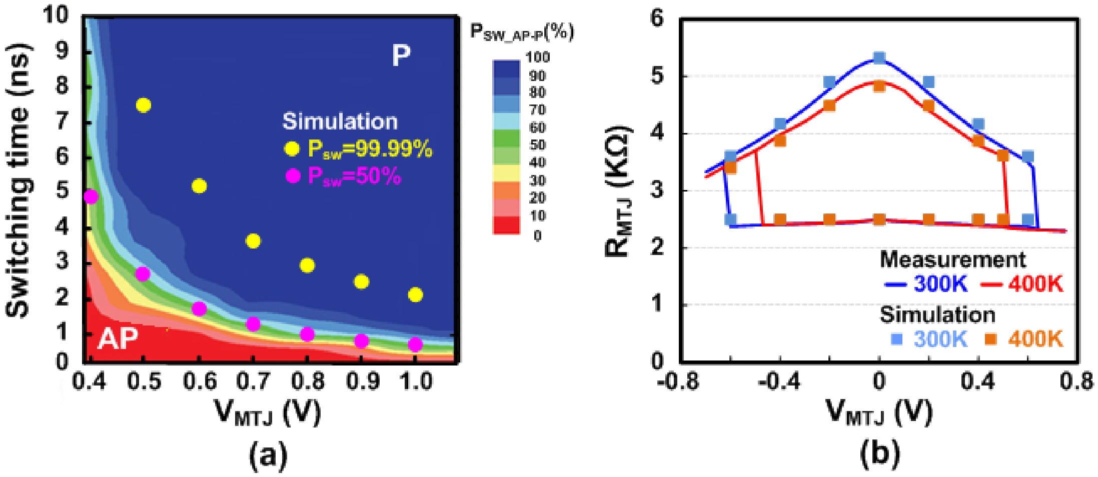


**Fig. 3.5 Simulation results compared to experimental data. (a) Switching time as a function of bias voltage across the MTJ. (b) Temperature-dependent R-V hysteresis curve.**

In Fig. 3.5(a), simulation results show good agreement with the 50% switching probability contour where switching time was measured as a function of bias voltage [112]. The model can be extended to track a higher percentile contour such as 99.99%. In Fig. 3.5(b), the simulated R-V hysteresis curves reproduce the MTJ resistance and the critical switching voltage ($V_c$) at different temperatures [113]. The flexibility of our model makes it easy to incorporate additional MTJ characteristics. For instance, asymmetry in the I-V hysteresis curve shown in Fig. 3.6(a) can be included using a user-defined asymmetry factor. Fig. 3.6(b) shows the switching probability according to the critical initial angle, providing detailed info regarding the switching current requirement.

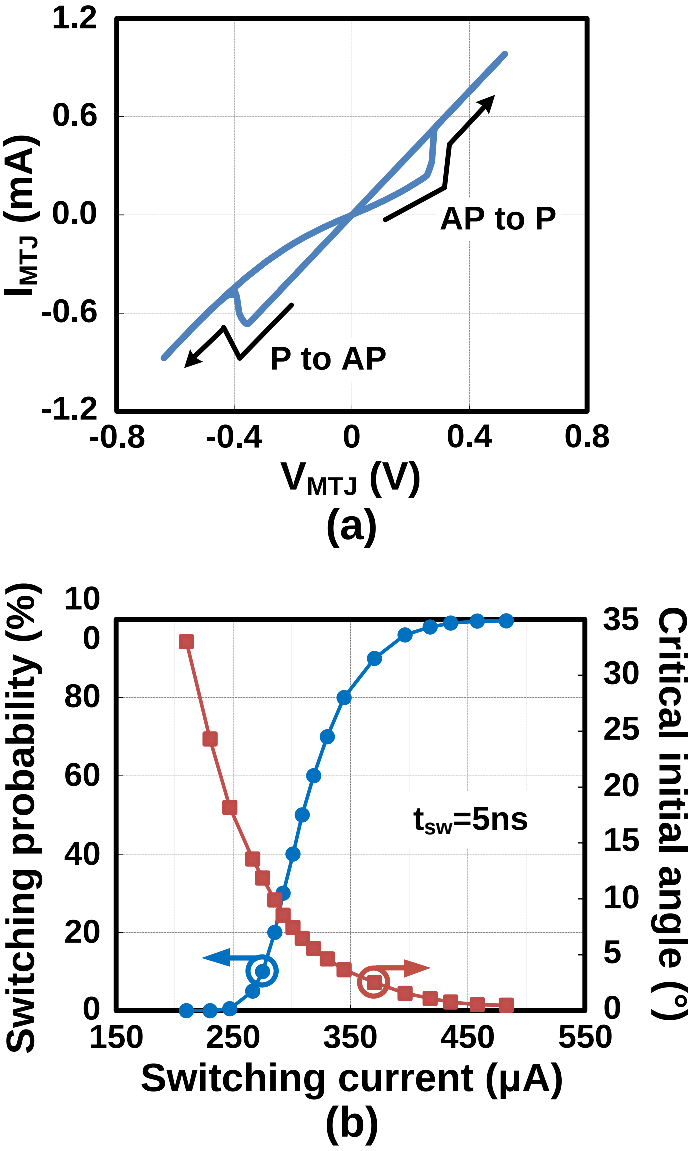
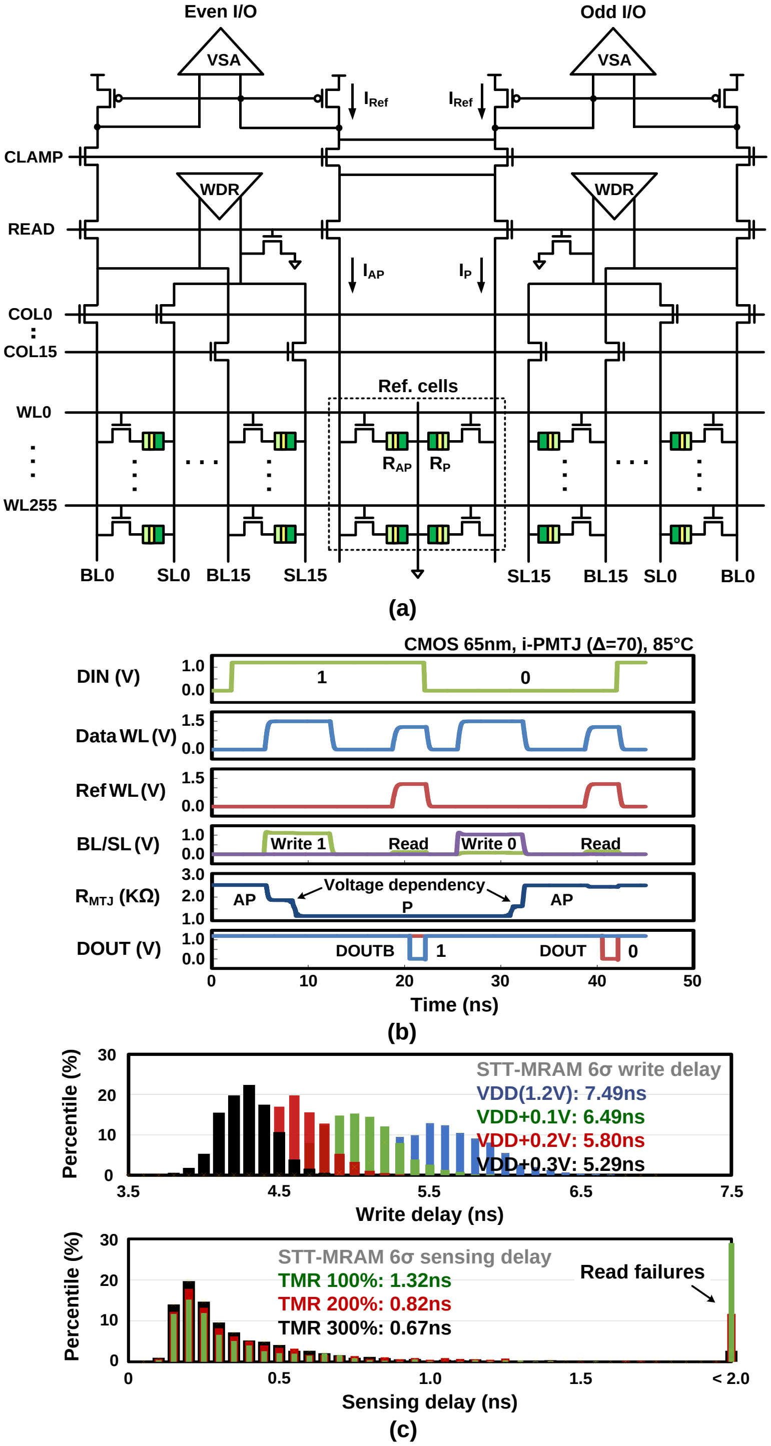


**Fig. 3.6 MTJ switching behavior characterization. (a) I-V hysteresis curve with asymmetric $V_c$ for bi-directional switching. (b) Switching probability as a function of switching current using an initial angle dependency.**

### *3.1.4 Model Application to Circuit Simulation*

The proposed model was used to study the statistical behavior of STT-MRAM read and write delays considering geometrical variation in both MTJ and CMOS. Fig. 3.7(a) shows the simplified read/write circuit schematic used for this experiment. MTJs were assumed to be connected in the reverse direction (i.e. so called top-pinned MTJ structure) to balance bi-directional switching [113]. Note that the current for parallel to anti-parallel switching is typically larger than that for anti-parallel to parallel switching so that a free layer contact is connected to access transistor and a pinned layer is connected to SL to provide larger current for parallel to anti-parallel switching. Bi-directional write current drivers and dual-voltage WL drivers are used to ensure sufficient write margin [99], [114]. Self-referencing MTJ cells and a mid-point reference circuit generating $I_{Ref}=(I_{AP}+I_P)/2$ are incorporated for good readability [115]. As shown in Fig. 3.7(b), the STT-MRAM critical path comprising of the MTJ and CMOS shows good read/write operations. Using this simulation setup, realistic variation is introduced to MTJ input parameters (i.e. $W$, $L$, $t_F$, RA) as well as CMOS input parameters (i.e. transistor $W$, $L$, $V_{th}$, $T_{ox}$). Fig. 3.7(c) shows write and sensing delay distributions with 6σ values denoted in the figure legends. The write delay is measured from WL activation to the time when the $\bar{M}$ flips while the sensing delay is measured from WL activation to the point when the bitline voltage difference reaches 25mV. As the write voltage increases, the switching delay distribution becomes narrower due to the faster precession at the higher bias voltage. For the sensing delay, the mismatch of read current paths between data and reference cells directly affects sensing voltages so a higher TMR is required for better read margin.



(a)

(b)

(c)

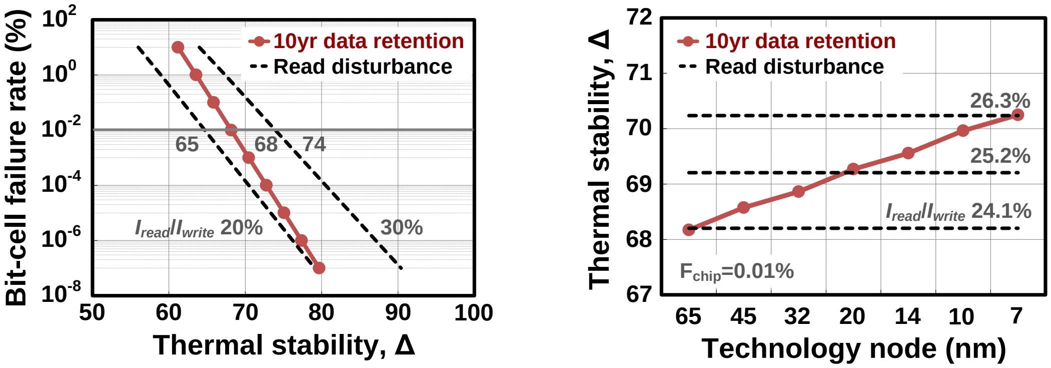
**Fig. 3.7 STT-MRAM statistical simulation results. (a) Column circuit for read/write circuitry. (b) Waveforms for STT-MRAM read/write operation. (c) Write and sensing delay distributions under process variation ($10^3$ MC runs).**

## 3.2 Thermal Stability Criteria

In order for a fair comparison between the various MTJs, our scaling analysis is performed under the same thermal stability ($\Delta = E_b/k_BT$) which represents the energy required to flip the magnetization direction under thermal fluctuation. The bit-cell retention failure rate ($F_{bit\text{-}cell}$) can be derived using the following equation [82]:

$$F_{bit-cell} = 1 - \exp[-m\frac{t}{\tau_0}\exp\{-\Delta(1-\frac{I_{cell}}{I_{write}})\}] \tag{6}$$

where *m* is the memory size, *t* is the retention time, $\tau_0$ is the attempt cycle time (typically, 1ns), $I_{cell}$ is the applied current flowing through the MTJ, and $I_{write}$ is the write current which is equivalent to critical switching current ($I_c$). Here, $F_{bit\text{-}cell}$ indicates the probability of bit cells experiencing unwanted magnetization flip for a given total memory capacity. Typically, the bit-cell failure could possibly occur during retention mode and read mode. When no current flows through the MTJ, which corresponds to the retention mode, the probability of unwanted switching depends on thermal fluctuation of the magnetization in the free layer. Therefore, the required thermal stability for the retention mode can be estimated by making *m* and *t* the memory density and the retention period (e.g. 10years) while setting $I_{cell}=0$ in the equation (6). During read mode, a current smaller than critical switching current flows through the MTJ and excites the magnetization in the free layer, which increases the probability of unwanted switching than that in the retention mode by reducing effective energy barrier. Under this read disturbance, the required thermal stability can be expected by making *m* and *t* the number of bits accessed in parallel and the worst case read time (e.g. 10years×$t_{read}/t_{cycle}$) while setting $I_{cell}=I_{read}$. Note that thermal fluctuation increases along with temperature so that the energy barrier ($E_b$) between two stable states (i.e. parallel and anti-parallel states) has to be large enough to maintain the same thermal stability at a higher temperature.



(a) (b)

**Fig. 3.8 Thermal stability criteria under data retention failure and read disturbance. (a) Thermal stability requirement as a function of data failure rate (memory size: 16MB, 32bit parallel read, $t_{read}/t_{cycle}$: 30%). (b) Thermal stability for 10yr data retention and tolerable read current under read-disturbance criterion over the technology scaling (memory size at 65nm:16MB).**

Fig. 3.8(a) shows the bit-cell failure rate as a function of thermal stability for 10-year retention and read modes. As the thermal stability become larger, the bit-cell failure rate reduces for both cases. Moreover, high thermal stability allows larger read current for a given bit-cell failure rate. In Fig. 3.8(b), the thermal stability requirements for various technology nodes are estimated targeting 0.01% of bit-cell failure rate for 10year retention. Here, we choose Intel’s server class processor as a benchmark target assuming its L3 cache memory density doubles every two nodes [116], [117]. As memory density increases with technology scaling, higher thermal stability is needed. Note that these thermal stability criteria will be used to decide MTJ dimensions in section 3.3. Moreover, the read current constraint for each technology node is also estimated based on the thermal stability criteria, which will be used to evaluate the read performance of MTJ technologies in section 3.4.

## 3.3 MTJ Scaling Scenarios for Low Switching Current

Once the thermal stability requirement to maintain necessary level of nonvolatility is specified, the next step involves a composition of detailed scaling scenario for three MTJ options: IMTJ, c-PMTJ, and i-PMTJ. In this section, we first discuss detailed scaling methods for each MTJ technology under iso-retention condition. Then, the possible scenarios are suggested to find a way to reduce switching current along the MTJ scaling while considering realistic MTJ dimensions and material parameters.

### *3.3.1 MTJ Scaling Methods under Iso-Retention Condition*

Typically, thermal stability is a strong function of MTJ volume (*V*) and anisotropy field ($H_k$) as follows:

$$\Delta = \frac{E_b}{k_B T} = \frac{H_k M_s V}{2k_B T} = \frac{H_k M_s W L t_F}{2k_B T} \tag{7}$$

where *W*, *L*, $t_F$ are the MTJ width, length, and thickness, respectively. Since our scaling analysis is based on iso-retention condition (i.e. 10years), a reduction in *W* and *L* during the MTJ scaling, which leads to decrease in thermal stability, has to be compensated to meet the target retention by adjusting $H_k$ or $t_F$. Moreover, $H_k$ is defined by MTJ dimensions in different ways according to anisotropy sources so that the scaling method is also different for each MTJ technology.

For IMTJ, $H_k$ is primarily determined by $H_{k,shape}$, which depends on aspect ratio (AR) of magnet as expressed in equation (1). Therefore, Δ of the IMTJ can be expressed as

$$\Delta_{IMTJ} = \frac{H_{k,shape} M_s V}{2k_B T} = \frac{2\pi(N_{dx} - N_{dy}) M_s^2 W L t_F}{k_B T} \qquad (8)$$

When $W$ and $L$ become smaller during the scaling, either the aspect ratio or $t_F$ needs to be increased to achieve necessary Δ. For PMTJs, $H_k$ is governed by a difference between $H_{k\perp}$ and $H_{dz}$ as derived in equation (2). Thus, Δ of the PMA can be expressed as

$$\Delta_{PMTJ} = \frac{H_{k\perp eff} M_s V}{2k_B T} = \frac{(K_\perp - 2\pi N_{dz} M_s^2) W L t_F}{k_B T} \qquad (9)$$

As mentioned earlier, in case of c-PMTJ, $K_\perp$ becomes $K_u$ of specific material so larger $K_u$ or $t_F$ increases Δ. For i-PMTJ, $K_\perp$ is proportional to $t_c/t_F$, and thus, decreasing $t_F$ provides higher Δ. Overall MTJ scaling methods for three MTJ technologies to meet iso-retention condition are illustrated in Fig. 3.9.

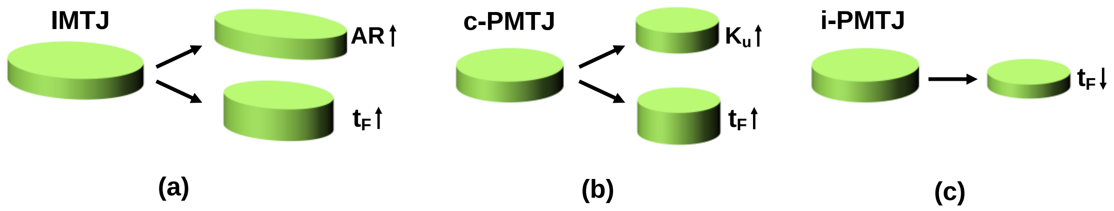
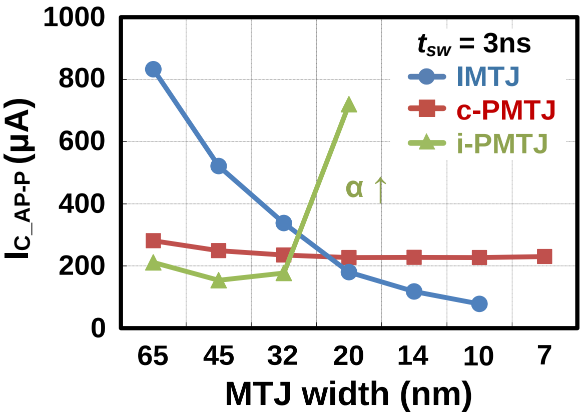


**Fig. 3.9 MTJ scaling methods under iso-retention condition for (a) IMTJ, (b) c-PMTJ, and (c) i-PMTJ.**

| Technology node (nm) | | | 65 | 45 | 32 | 20 | 14 | 10 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| L3 cache memory size (MB) | | | 16 | 24 | 32 | 48 | 64 | 96 | 128 |
| Thermal stability (10yrs, 85C) | | | 68.2 | 68.6 | 69.0 | 69.3 | 69.7 | 70.0 | 70.2 |
| MTJ width (nm) | | | 65 | 45 | 32 | 20 | 14 | 10 | 7 |
| IMTJ (CoFeB) | $M_s$=1077, P=0.6 | AR | 2.2 | 2.9 | 3 | 3 | 3 | 3 | N/A |
| | | $t_F$ (nm) | 2 | 2 | 2.44 | 3.51 | 4.73 | 6.54 | N/A |
| | | α | 0.0068 | 0.0068 | 0.0062 | 0.0055 | 0.0051 | 0.0048 | N/A |
| | | Remark | AR ↑ | | $t_F$ ↑ , $t_F$ dependent α | | | | No IMA |
| c-PMTJ (FePtX) | $M_s$=1116, P=0.51, α=0.055 | $K_u$ | 0.94 | 1.13 | 1.49 | 2.65 | 4.63 | 6.6 | 6.6 |
| | | $t_F$ (nm) | 0.45 | 0.45 | 0.45 | 0.45 | 0.45 | 0.58 | 1.17 |
| | | Remark | constant $t_F$, $K_u$ ↑ | | | | | $t_F$ ↑ | |
| i-PMTJ (CoFeB) | $M_s$=1077, P=0.6, $t_c$=1.5nm | $t_F$ (nm) | 1.49 | 1.38 | 1.15 | 0.35 | N/A | N/A | N/A |
| | | α | 0.013 | 0.018 | 0.04 | 0.22 | N/A | N/A | N/A |
| | | Remark | $t_F$ ↓, $t_F$ dependent α | | | | Insufficient Δ | | |

*$M_s$: Saturation magnetization ($10^3$A/m), $K_u$: Crystal anisotropy ($10^6$J/m$^3$)

(a)



(b)

**Fig. 3.10 (a) MTJ scaling scenario with minimum MTJ width (= minimum feature size). (b) Discontinued Ic scaling of MTJs due to severe dimensional scaling.**

### *3.3.2 MTJ Scaling with Minimum MTJ Width*

In order to explore the possible issues during the scaling, we first severely scales MTJ dimensions with minimum MTJ width for a given technology node (i.e. minimum feature size). Fig. 3.10(a) shows a detailed MTJ scaling scenario when the MTJ width ($W$) is fixed as the minimum feature size. Note that the MTJ length ($L$) is set as $W$×AR for IMTJ while AR of 1 is assumed for PMTJs since its thermal stability no longer depends on the shape anisotropy. To meet the thermal stability criteria given in Fig. 3.8(b), MTJ dimensions for each technology node are adjusted using scaling methods discussed in previous section. A typical free layer material for each MTJ technology is also considered with critical material parameters such as $M_s$, $P$, and $\alpha$.

As for IMTJ, Δ of CoFeB free layer is achieved by increasing AR up to 3 with scaling down to 45nm node to ensure a single-domain switching and then increasing $t_F$ for further scaling [105]. Note that if we continue to increase $t_F$ down to the 7nm node, IMTJ will lose the in-plane anisotropy (IMA) since $t_F$ is longer compared to the in-plane dimensions (i.e. it's no longer an IMTJ). For c-PMTJ, to provide enough Δ down to 7nm node, FePtX, which can provide a very high $K_u$ value, is chosen as a target material even though its switching efficiency is relatively low due to high damping and low polarization. For iso-retention scaling, $K_u$ is increased up to $6.6\times10^6$J/m$^3$ while maintaining 0.45nm of $t_F$ to reduce the switching current by magnifying out-of-plane demagnetizing field ($H_{dz}$) [118], [119]. Beyond 14nm node, $t_F$ is increased. In case of i-PMTJ, $t_F$ of CoFeB free layer is decreased to increase interface anisotropy [34]. However, i-PMTJ still shows insufficient Δ below 20nm node since a smaller $t_F$

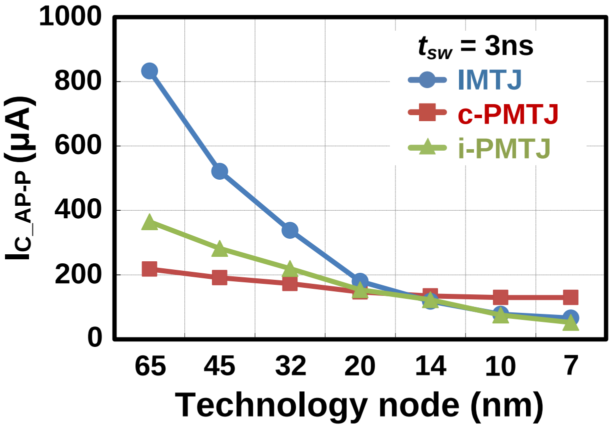
decreases MTJ volume also. Thickness dependency of $\alpha$ in CoFeB is also considered for both IMTJ and i-PMTJ [34], [120].

Based on the scaling scenario, Fig. 3.10(b) shows critical switching current ($I_c$) for the three types of MTJs under a constant switching time of 3ns by using our MTJ model. For IMTJ, $I_c$ rapidly decreases with scaling and eventually becomes lower than that of PMTJs from the 20nm node. This is because $H_{dz}$ greatly reduces with increasingly large thickness so that magnetization switching becomes easier. Note that $H_{dz}$ contributes to an increase in switching current but thermal stability in in-plane devices. For c-PMTJ, $I_c$ is relatively constant during the scaling but still large due to low $P$ and high $\alpha$ of FePtX. Moreover, i-PMTJ shows a sharp increase in $I_c$ since $\alpha$ of CoFeB free layer exponentially increases when its thickness is thinner than 2nm [34].

Unfortunately, with severely scaled MTJ dimensions, $I_c$ scaling down to the 7nm node is not readily achievable for all three MTJ technologies. However, one interesting finding is that $I_c$ of c-PMTJs is less sensitive to dimensional scaling. This characteristic is still valid for i-PMTJ if we assume an increase in $\alpha$ is not considerable. Therefore, we can consider a relaxed MTJ width, which is larger than minimum feature size, to mitigate the scaling issues described in this section. As expressed in (9), an increase in MTJ volume reduces necessary anisotropy field for a given target thermal stability, which, in turn, provides more freedom to select magnet materials and dimensions for lower $I_c$. Note that this approach is not applicable to IMTJ since its switching current reduces along with a dimensional scaling.

| Technology node (nm) | | | 65 | 45 | 32 | 20 | 14 | 10 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| IMTJ (CoFeB) | $M_s$=1077, P=0.6 | MTJ width | 65 | 45 | 32 | 20 | 14 | 10 | 8.3 |
| | | AR | 2.2 | 2.9 | 3 | 3 | 3 | 3 | 3 |
| | | $t_F$ (nm) | 2 | 2 | 2.44 | 3.51 | 4.73 | 6.54 | 8 |
| | | Remark | Area shrinkage needed for continued $I_c$ scaling | | | | | | |
| c-PMTJ (FePdX) | $M_s$=1077, P=0.51, α=0.03 | MTJ width | 80 | 70 | 60 | 45 | 32 | 24 | 20 |
| | | $K_u$ | 0.83 | 0.87 | 0.92 | 1.08 | 1.45 | 2 | 2 |
| | | $t_F$ (nm) | 0.45 | 0.45 | 0.45 | 0.45 | 0.45 | 0.46 | 0.65 |
| | | Remark | Moderate $K_u$ material with low damping | | | | | | |
| i-PMTJ (CoFeB) | Ms=1077, P=0.6, $t_c$=1.5nm | MTJ width | 80 | 70 | 60 | 45 | 32 | 24 | 20 |
| | | $t_F$ (nm) | 1.49 | 1.49 | 1.47 | 1.38 | 2.99 | 2.8 | 2.31 |
| | | α | 0.012 | 0.012 | 0.013 | 0.018 | 0.0058 | 0.006 | 0.0062 |
| | | Remark | Single interface | | | | Double interface | | |

(a)



(b)

**Fig. 3.11 (a) MTJ scaling scenario with relaxed MTJ width. (b) Continued Ic scaling of MTJs with relieved dimensional scaling.**

### *3.3.3 MTJ Scaling with Relaxed MTJ Width*

Fig. 3.11(a) shows a MTJ scaling scenario with relaxed MTJ width to continue lowering $I_c$. Since an MTJ volume increases with the relaxed dimensions contributing thermal stability, the necessary anisotropy field becomes smaller. Since the anisotropy field is a strong function of MTJ dimensions and material parameters, we can optimize those parameters for lowering switching current rather than increasing thermal stability. For example, a relaxed c-PMTJ can meet Δ requirement down to 7nm even with moderate-$K_u$ materials such as FePdX ($K_u$=2×10$^6$J/m$^3$) which has a lower $\alpha$ than FePtX (i.e. high-$K_u$ material) [118], [121]. For i-PMTJ, a relaxed MTJ width makes it easy to meet target thermal stability even without aggressive thinning of the free layer, avoiding an abrupt increase in $\alpha$. From 14nm node, a double interface MTJ, which has two MgO layers on the top and bottom of the CoFeB free layer, is introduced to improve both thermal stability and switching efficiency [122]. Two interfaces between CoFeB and MgO doubles interface anisotropy so that an MTJ can have sufficient thermal stability even with relaxed free layer thickness while mitigating an increase in $\alpha$. For IMTJ, a larger width is used at 7nm to meet Δ requirement while keeping in-plane magnetization. Based on this scenario, the scaling trend of $I_c$ is presented in Fig. 3.11(b). Compared to Fig. 3.10(b), $I_c$ of three MTJs continue to scale down to 7nm showing better switching efficiency. One additional advantage of relaxing MTJ size is that $R_{MTJ}$ becomes smaller allowing larger driving current for a given voltage across the MTJ. All these considerations are integrated to define the input parameters of the MTJ SPICE model and a circuit-level performance of MTJ technologies are evaluated in the following section.

## 3.4 Circuit-Level Scaling Trend of In-plane and Perpendicular MTJ Based STT-MRAMs

In order to compare the circuit-level scaling trend of in-plane and perpendicular MTJ based STT-MRAMs, we first implement STT-MRAM bit-cells using the proposed MTJ SPICE model. The MTJ parameters from the scaling scenario, which ensure the requirements for fixed retention failure and read disturbance, are incorporated. Along with scalable bit-cell design, peripheral circuits for STT-MRAM operations are also implemented specifically for scaling analysis by using the predictive technology model (PTM) which provides electrical properties of advanced transistors based on ITRS roadmap. Using this simulation setup, the scalability of read/write performance is evaluated to find the most viable MTJ technology for future STT-MRAMs.

### *3.4.1 Simulation Setup for Circuit-Level Scaling Analysis*

Fig. 3.12(a) shows a schematic for STT-MRAM read/write circuit which is simplified version of the STT-MRAM column circuit. The details on the circuit implementation are presented in section 3.1.4. Fig. 3.12(b) and (c) show the waveforms for write and read operation indicating delay criteria used in this work. A write delay measures the amount of time it took from WL activation to the point when the magnetization switches. Here, we assume that switching is completed when the magnetization pass through a hard axis (i.e. magnetization=0). For a read delay, it is defined from WL activation to the time when the voltage difference between sensing nodes reaches 25 mV, which is half the sensing margin of SRAM. Due to the single-ended sensing nature and the limited TMR, it is not practical for STT-MRAM to enforce the same BL voltage difference requirement as SRAM [97].

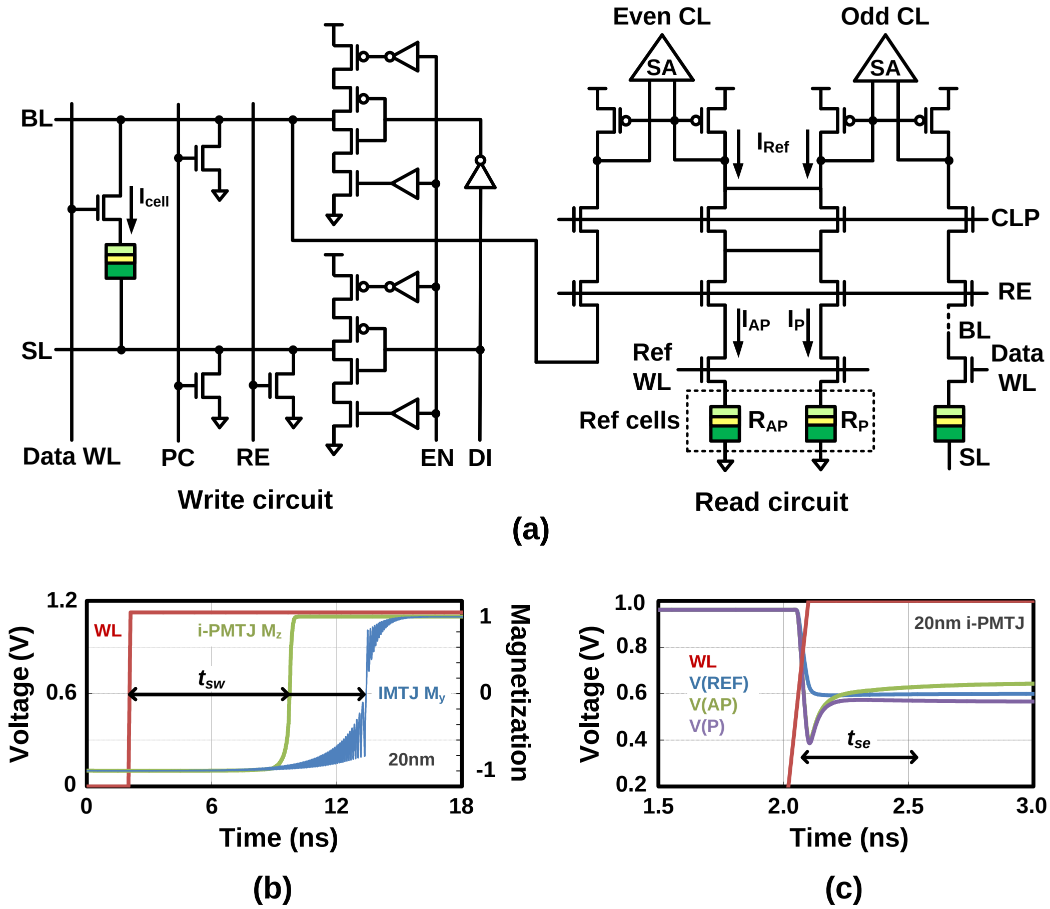


**Fig. 3.12 STT-MRAM circuit-level simulation. (a) Schematic for write/read circuitry. (b) Write delay criterion. (c) Read delay criterion.**

In order to implement peripheral circuits of STT-MRAM for various technology nodes, the predictive technology (PTM) model based on the high performance logic transistor roadmap from ITRS, is used considering advanced logic technologies as shown in Fig. 3.13 [124]. The RA values extrapolated from ITRS projection are also taken into account. For double interface MTJ, 10% larger RA is assumed considering two MgO layers. As shown in Fig. 3.13(a), $R_{AP}$ for three MTJs becomes increasingly large throughout the scaling due to a rapid decrease in MTJ area. Fig. 3.13(b) shows a cell current ($I_{cell}$) from access transistor during AP-to-P switching when the width of the access device is chosen as 18F. As technology scales, $I_{cell}$ rapidly reduces showing that an increase in resistance can degrade the current drivability of access transistor during the scaling.

| Technology node (nm) | 65 | 45 | 32 | 20 | 14 | 10 | 7 |
|---|---|---|---|---|---|---|---|
| Improvement of Transistor performance | Planar bulk | | | Multi-gate | | | |
| | Strained-Si | | | | | | |
| | Poly | High-K metal gate | | | | | |
| VDD (V) | 1.2 | 1.1 | 1 | 0.9 | 0.8 | 0.75 | 0.7 |
| *$R_pA$ ($\Omega\cdot\mu m^2$ ) | 11 | 9 | 7 | 5 | 4 | 3 | 2.5 |

* 10% larger RA assumed for double interface PMTJ

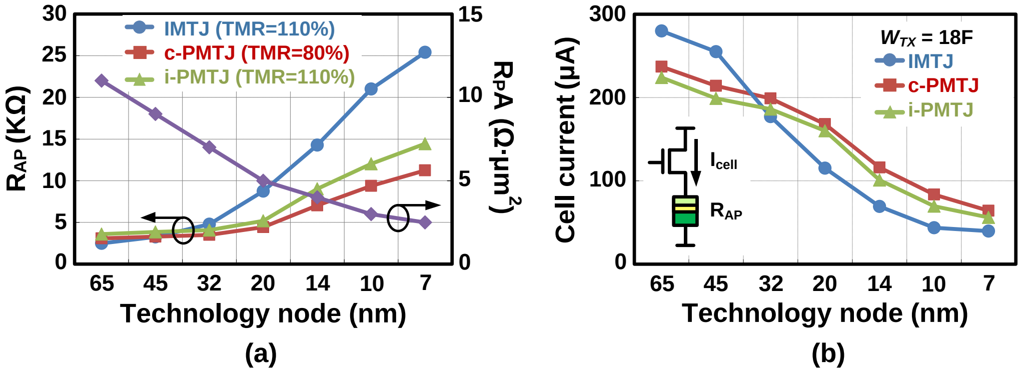


**Fig. 3.13 Scaling trends of (a) MTJ resistance in anti-parallel state and (b) Drivability of cell current when the anti-parallel state MTJ is connected.**

### *3.4.2 Circuit-Level Scalability of MTJ Technologies*

Fig. 3.14(a) shows the scaling trend of write delay for in-plane and perpendicular MTJ based STT-MRAMs. Interestingly, three STT-MRAMs show different delay trends. This can be explained by different $I_c$ scalability of each MTJ with respect to continuously decreasing $I_{cell}$. In Fig. 3.11(b), $I_c$ of three MTJ show different slopes during the scaling. However, $I_{cell}$ of three MTJs reduce with a similar slope as shown in Fig. 3.13(b). Therefore, IMTJ shows faster switching at the scaled nodes since $I_c$ scales faster than $I_{cell}$. On the other hand, c-PMTJ shows an exponential increase in the write delay due to slower $I_c$ scaling compared to $I_{cell}$. For i-PMTJ, a similar scalability between $I_c$ and $I_{cell}$ leads to a relatively constant delay trend showing the shortest delay among three options. It is noteworthy that $I_c$ needs to scale down faster than $I_{cell}$ to keep reducing the write delay over the scaling.

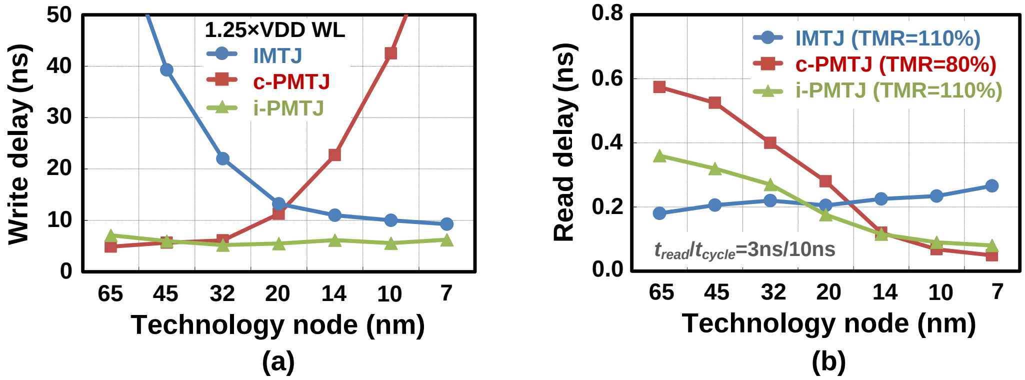


**Fig. 3.14 Scaling trends of (a) write delay and (b) read delay for in-plane and perpendicular MTJ based STT-MRAMs (128WL x 128 BL assumed).**

In order to fairly assess the read performance, $I_{read}$ is set to maximum percentile of write current that is tolerable under read-disturbance criterion (i.e. $F_{bit\text{-}cell}$=0.01%) as discussed in section 3.2. In this work, $I_{read}$ is determined based on write current for 3ns of switching time, which is equivalent to $I_c$ values in Fig. 3.11(b). Moreover, the line capacitances assuming 16Kbit cell array (i.e. 128WLs x 128BLs) and different TMR values corresponding to polarization factor of each MTJ are considered. As shown in Fig. 3.14(b), the simulated read delay shows different scaling trends for three STT-MRAMs. Typically, a read delay can decrease along with a scaled line capacitance or increase with reduced $I_{read}$ and larger $R_{MTJ}$ at the scaled nodes. As a result, IMTJ shows increasingly large read delay due to a rapid decrease in $I_{read}$ and an increase in $R_{MTJ}$ while PMTJs shows a faster read along with scaled BL capacitance. Based on the scaling trend of circuit performance, the i-PMTJ based STT-MRAM shows a good balance between read and write margins compared to other options.

| | IMTJ | c-PMTJ | i-PMTJ |
|---|---|---|---|
| Anisotropy source | shape | crystal | interface |
| Adjustable options for Δ | AR ↑ or $t_F$ ↑ | $t_F$ ↑ | $t_F$ ↓ |
| Applicable technologies | - | Moderate $K_u$ material | Double interface |
| | | Relaxed MTJ width | |
| Write delay[1] | 1X | 2.06X | 0.56X |
| Read delay[1] | 1X | 0.51X | 0.55X |
| Possible issues | - Higher $R_{MTJ}$ than enlarged PMTJs<br>- large Δ variation | - High damping<br>- Low TMR<br>- Difficult process | - High RA with double interface<br>- Thickness variation |
| | | - Likely to cause magnetic coupling issues | |

[1] All metrics are compared at 14nm node.

**Table 3.2 Summary of this work in consideration of possible issues with scaling.**

## 3.5 Chapter Summary

New MTJ materials are actively being pursued by the magnetics community aiming at low switching current and high thermal stability. This work explores the scalability of STT-MRAM memory cells based on various MTJ technologies: namely, in-plane MTJ, crystal perpendicular MTJ, and interface perpendicular MTJs. A dedicated SPICE MTJ model was developed and calibrated to perform an extensive scaling analysis. Using the scaling scenario with relaxed MTJ width, we compare the key performance metrics such as write and read delay down to the 7nm node under the same degree of retention failure and read disturbance. Our results show that i-PMTJ is a promising option for future STT-MRAM achieving a balanced performance in write and read operations. Table 3.2 provides an overall summary of this work in consideration of possible issues with scaling.

# Chapter 4 Spin-Hall Effect MRAM Based Cache Memory: A Feasibility Study

One of the key objectives of STT-MRAM research has been on minimizing switching current while maintaining the required nonvolatility. To address this challenge, non-traditional MRAMs based on novel switching mechanisms have been proposed. In particular, spin-Hall effect (SHE) which utilizes large spin currents generated in the direction transverse to the charge current have been recently drawing attention [64]. Despite early promises such as lower switching current by means of efficient spin generation (i.e. $I_{spin}/I_{charge}$>100%) and longer device lifetime owing to the decoupled read and write paths, there is still a lack of a comprehensive study for benchmarking SHE-MRAM against other memory technologies. In this work, we explore the trade-off points across different levels of design abstraction (i.e. device, circuit, and architecture) to evaluate the feasibility of SHE-MRAM for large on-die cache memory [125].

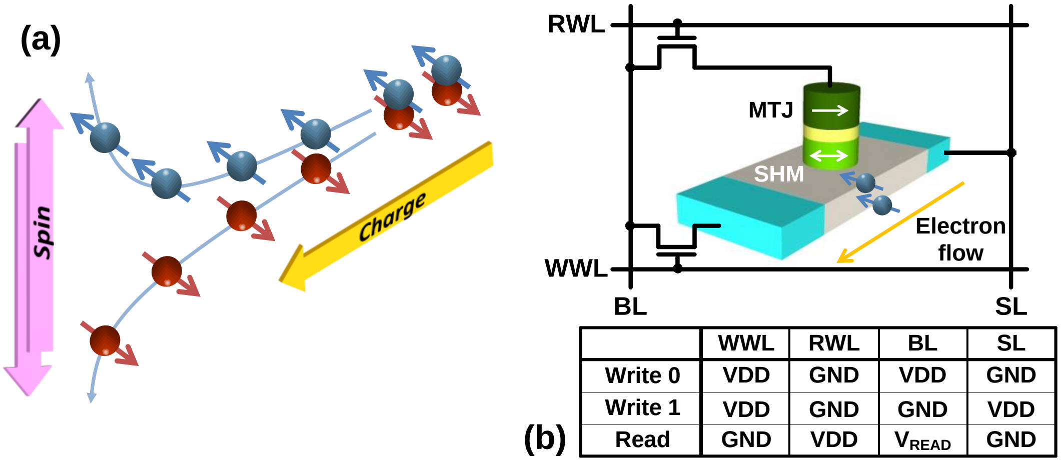
## 4.1 SHE-MRAM Device Design

### *4.1.1 Device Concepts of SHE-MRAM*

Recently, low-energy STT–MTJ switching based on the spin-Hall effect (SHE) has been proposed demonstrating that a charge current through a spin-Hall metal (SHM such as Pt, β-Ta, β-W, and others) on the CoFeB layer generates a spin current in the traverse direction with large spin–orbit coupling. Since this spin current generation efficiency is much higher than that of standard STT-MRAM, it has been drawing a lot of interests from research communities. Fig. 4.1 illustrates the generation of spin current by SHE, along with the cell structure of a SHE-MRAM which requires two transistors for separate read and bidirectional write operation. During write operation, the write transistor is on and the read transistor is off while a charge current ($I_c$) flowing between BL and SL through SHM induces a spin current ($I_s$) through a free layer of the MTJ device as follows [65]:

$$I_s = \frac{A_{MTJ}}{A_{SHM}} \theta_{SH} (1 - \sec h(\frac{t_{SHM}}{\lambda_{SHM}})) \cdot I_c \ , \qquad (1)$$

where $A_{SHM}$ and $A_{MTJ}$ are the cross-sectional area of the underlying SHM layer and the MTJ device, respectively. And $\theta_{SH}$ is referred to the spin-Hall angle, which can be as large as 0.15 in β-tantalum (Ta) and 0.3 in tungsten (W). $t_{SHE}$ and $\lambda_{SHE}$ are the thickness and the spin diffusion length of SHM layer. During read operation, the write transistor is off and the read transistor is on so that a read current can flow through SL due to high off-resistance of the write transistor in the BL side. Although this three-terminal device potentially results in an area penalty, it offers several advantages over the traditional 1T-MTJ STT-MRAM. First, the charge-to-spin conversion efficiency ($I_{spin}/I_{charge}$) higher than 100% using optimal SHM dimension enables a significantly low switching current without impacting nonvolatility. Second, the separate read and write paths, allowing for longer device lifetime because only the small read current flows through the tunnel oxide as the write current flows through the SHM itself. Since the upper bound for the voltage across the MTJ (i.e. the maximum write-speed) is determined by the tunnel barrier reliability, SHE-MRAM can have additional improvement in writability [66]. By utilizing these advantages, SHE-MRAM could be a promising option to achieve high speed and low power on-chip memory beyond the limitation of standard STT-MRAM.

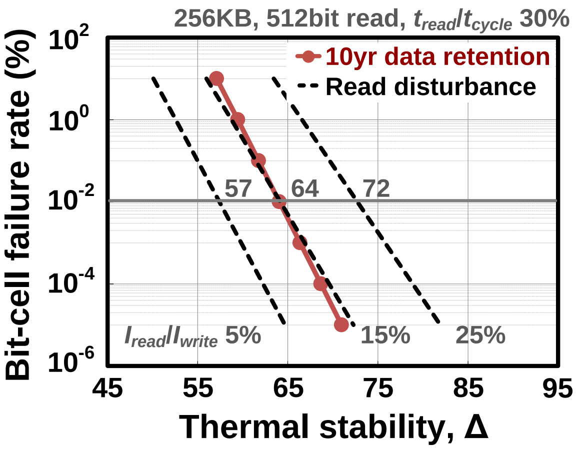


|  | WWL | RWL | BL | SL |
|---|---|---|---|---|
| Write 0 | VDD | GND | VDD | GND |
| Write 1 | VDD | GND | GND | VDD |
| Read | GND | VDD | $V_{READ}$ | GND |

**Fig. 4.1 (a) Transverse spin current generation by SHE. (b) SHE-MRAM cell configuration.**

### *4.1.2 Device Parameter Setup*

We first decide device parameters for SHE-MTJ structure such as device dimensions and material options targeting 10 years of retention time. In Fig. 4.2, the target thermal stability (Δ) criterion is set as 65 by considering 256Kbit of memory size with 0.01% of the bit-cell failure rate [82]. The maximum read current is also determined under the same degree of read disturbance failure rate as presented in chapter 3. To physically obtain this Δ requirement, the dimensions for CoFeB free layer are determined assuming 22nm technology node as shown in Table 4.1. Here, we used an in-plane MTJ as a storage element of SHE-MRAM since the direction of polarized spins from SHM is aligned with the lateral dimension. To the best of our knowledge, a PMA-based SHE-MRAM requires a small magnetic field to flip the magnetization towards intended state [64], which is not practical for high-density and low-power memory application. Since the direction of spin from SHM is parallel to the plane, the perpendicular magnetization of free layer is aligned with its hard axis by spin-Hall torque, and thus an additional stimulus such as magnetic field is needed to determine the final state.



$$F_{bi\text{-}cell} = 1 - \exp[-m\frac{t}{\tau_0}\exp\{-\frac{E}{k_B T}(1-\frac{I_{cell}}{I_{write}})\}]$$

- **Failure mode: 10yr data retention**

  - m: total memory size
  - t=10yrs, $I_{cell}$=0

- **Failure mode: Read disturbance**

  - m: number of bits per read
  - $t = t_{read}/t_{cycle} \times 10\text{yrs}$ $I_{cell} = I_{read}$

**Fig. 4.2 Thermal stability requirement estimation considering 10 year data retention and read disturbance.**

For device sizing, the MTJ width is set to 22nm which is corresponding to minimum feature size of the technology. To create shape anisotropy of in-plane MTJ, an aspect ratio is set to 3.5 ensuring single-domain magnet switching. Then, the thickness of CoFeB is determined as 2.7nm for sufficient energy barrier, which meets the Δ requirement at the temperature of 85°C. Moreover, the thickness dependency of damping and TMR value corresponding to the polarization factor are considered. The resistance-area product (RA) value of SHE-MRAM includes the resistance of SHM underneath the MTJ device. For STT-MRAM, an interface perpendicular MTJ is considered as a storage element. For SHM design, we choose the tungsten (W) for the SHM material, which provides a superior spin-Hall angle (i.e. $\theta_{SH}$=0.3). Its width and the length are selected to be sufficient large to place the MTJ on the top of the SHM layer while aligning the free layer magnetization with the spin direction from the SHM. Moreover, we choose 2.2nm of the SHM thickness where the maximum spin generation takes place using the equation (1) as shown in Fig. 4.3.

| Parameters | STT-MRAM | SHE-MRAM |
|---|---|---|
| MTJ type | Interface perpendicular | In-plane |
| Thermal stability factor | 65 | 65 |
| Free layer material | CoFeB | CoFeB |
| Free layer dimensions, $W_F \times L_F \times t_F$ | 40nm×40nm×1.34nm | 22nm×77nm×2.7nm |
| Saturation magnetization, $M_s$ | $1.077\times10^3$A/m | $1.077\times10^3$A/m |
| Damping factor, $\alpha$ | 0.018 | 0.006 |
| Polarization factor, $P$ | 0.63 | 0.63 |
| Critical thickness, $t_c$ | 1.5nm | - |
| TMR | 130% | 130% |
| RA ($\Omega\cdot\mu m^2$) | 5 | 5.5 |
| SHM dimensions, $W_{SHM} \times L_{SHM} \times t_{SHM}$ | - | 77nm×44nm×2.2nm |
| SHM spin diffusion length, $\lambda_{ch}$ | - | 1.5nm |
| SHM resistivity, $\rho_{ch}$ | - | $200\mu\Omega\cdot cm^2$ (for W) |
| Spin Hall angle, $\theta_{SH}$ | - | 0.3 (for W) |

* Δ and material parameters are extracted based on 85°C.

* RA of SHE-MRAM includes $R_{SHM}$. Thickness dependency of $\alpha$ is also considered.

**Table 4.1 Material parameters and device dimensions used for STT- and SHE-MRAM under the same degree of retention time (10 years).**

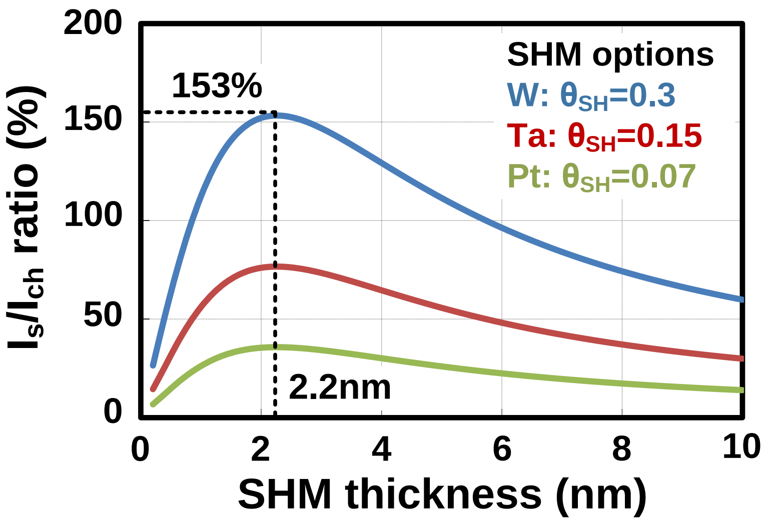
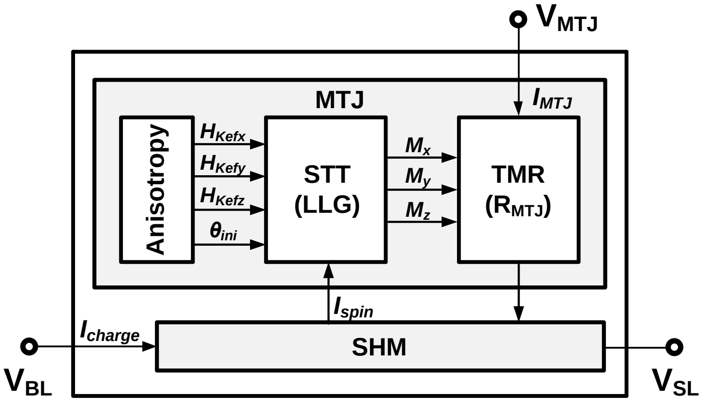


**Fig. 4.3 Spin current generation as a function of spin-Hall metal thickness which provides the optimal thickness for the maximum spin generation rate.**

## 4.2 SHE-MRAM Sub-Array Evaluation

In order to conduct a circuit-level evaluation, a SPICE-compatible SHE-MTJ model was implemented by incorporating the spin current from SHM into LLG equation as shown in Fig. 4.4 [66]. All the parameters listed in Table 4.1 are included as model input parameters. As described in chapter 3, the MTJ model consists of 4 main subcircuits such as anisotropy, STT, TMR, and temperature circuits. Based on this model frame, SHM circuit is added to generate spin current depending on SHM material and dimensions. During write operation when $V_{BL}$ and $V_{SL}$ terminals connected to the SHM are selected, a bi-directional charge current flow through the SHM depending on the bias polarity and a transverse spin current is generated transferring spin torque to the free layer of the MTJ. During read operation, $V_{MTJ}$ terminal is selected to apply a small read current or bias to the MTJ stack.



**LLG:**
$$\frac{1+\alpha^2}{\gamma}\cdot\frac{d\overline{M}}{dt} = -\overline{M}\times\overline{H}_{Keff} - \alpha\cdot\overline{M}\times(\overline{M}\times\overline{H}_{Keff}) + \frac{\hbar I_s}{2eWLt_F M_s}\cdot\overline{M}\times(\overline{M}\times\overline{M}_p)$$

**SHE:**
$$I_s = \frac{A_{MTJ}}{A_{SHM}}\theta_{SH}(1-\mathrm{sec}h(\frac{t_{SHM}}{\lambda_{sf}}))I_{ch}$$

**Fig. 4.4 SPICE simulation framework for SHE-MTJ device.**

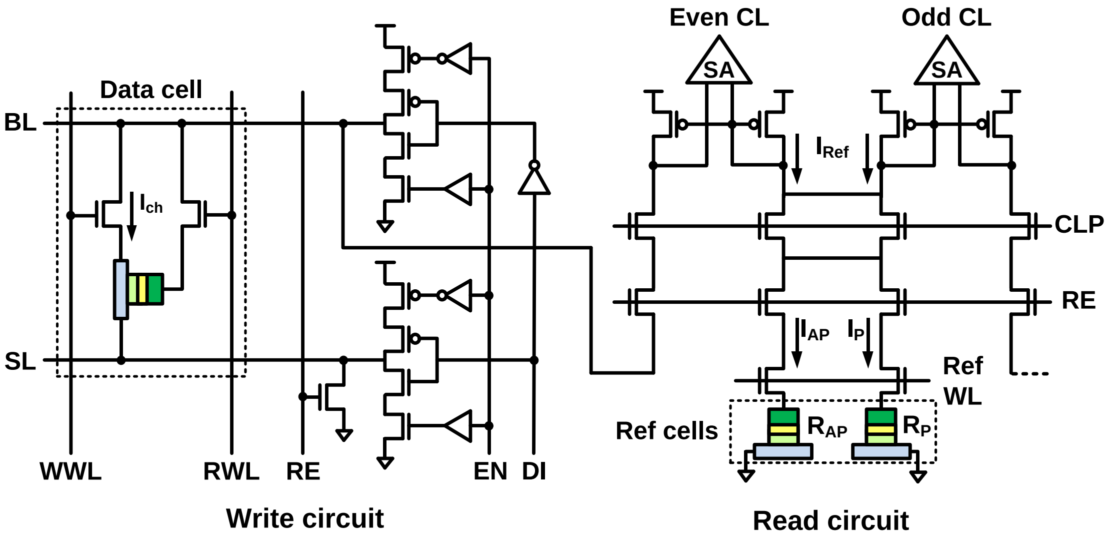
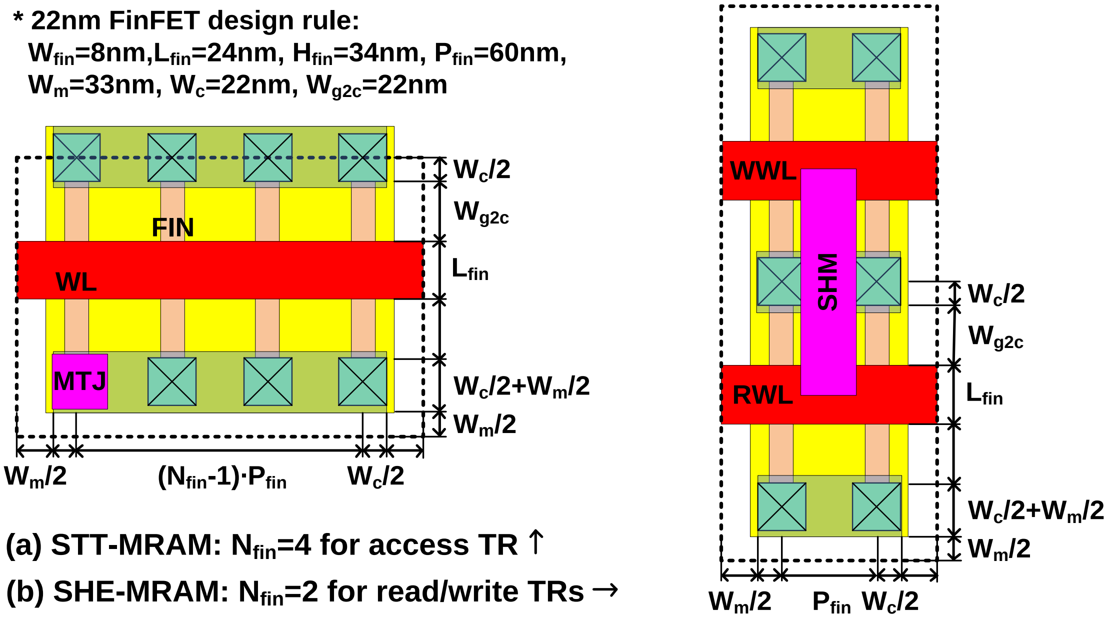


**Fig. 4.5 Read and write circuitry for SHE-MRAM featuring a bi-directional write current driver and a reference current ($I_{REF}$) generation circuit using an average cell current ($I_{AP}$ and $I_P$).**

Then, we use the model to simulate the read and write circuitry in Fig. 4.5. Here, 22nm high performance (HP) CMOS transistors from a publically available predictive technology (PTM) model were used for the circuit simulation [124]. The critical electrical properties from advanced technologies such as high-K metal gate, strained-Si channel, and multi-gate structure are included. Moreover, we constructed a realistic memory macro including bi-directional write current drivers and dual-voltage WL drivers to ensure a robust write operation [99], [114]. To maximize the read sensing margin, we adopted dummy MTJ cells for generating a reference current corresponding to the average value of the parallel and anti-parallel stage currents (i.e. $I_{Ref} = (I_{AP}+I_P)/2$) [115]. Note that read operation of STT-MRAM is typically done by current-forcing and voltage-measuring. The read current is applied in the antiparallel to parallel direction to minimize read disturbance issues [123].



* 22nm FinFET design rule:
$W_{fin}$=8nm,$L_{fin}$=24nm, $H_{fin}$=34nm, $P_{fin}$=60nm,
$W_m$=33nm, $W_c$=22nm, $W_{g2c}$=22nm
FIN
WL
MTJ
$W_c/2$
$W_{g2c}$
$L_{fin}$
$W_c/2+W_m/2$
$W_m/2$
$W_m/2$
$(N_{fin}-1)\cdot P_{fin}$
$W_c/2$
(a) STT-MRAM: $N_{fin}$=4 for access TR ↑
(b) SHE-MRAM: $N_{fin}$=2 for read/write TRs →
WWL
SHM
RWL
$W_c/2$
$W_{g2c}$
$L_{fin}$
$W_c/2+W_m/2$
$W_m/2$
$W_m/2$
$P_{fin}$
$W_c/2$

**Fig. 4.6 22nm FinFET-based layout for (a) STT-/ (b) SHE-MRAMs (3x denser than SRAM and equivalent to eDRAM).**

For estimating the bit-cell area, a FinFET based layout is considered as shown in Fig. 4.6 [126]. Here, we use 2 fins for read and write transistors which makes the cell area of SHE-MRAM comparable to that of a standard STT-MRAM cell. For STT-MRAM, 4 fins are used for single access transistor based on 1T-1MTJ layout style. Compared to an SRAM cell, both SHE-MRAM and STT-MRAM are roughly 3x denser, which are equivalent to an eDRAM cell in 22nm technology (i.e. 0.029μm$^2$) [127].

Using the proposed simulation setup, we first compare the performance of a single 256Kbit subarray in Table 4.2, which shows that SHE-MRAM has a 4.7x shorter write time and 1.3x shorter read delay as compared to a standard STT-MRAM with the same cell size. These results indicate that SHE-MRAM will always outperform STT-MRAM regardless of the cache size. Note that, unlike early concerns on area penalty of SHE-MRAM, SHE-MRAM shows better performance than STT-MRAM even with the same cell size. This is due to reduced charge current requirement during write operation so that SHE-MRAM bit-cell can be downsized with a smaller write transistor, which is half the size of the access transistor of STT-MRAM in this analysis.

| Metrics | STT-MRAM | SHE-MRAM |
|---|---|---|
| $I_{spin}$ at $t_{sw}$=3ns (μA) | 73 | 134 |
| Read delay (ns) | 0.42 | 0.33 |
| Read energy (fJ) | 4.5 | 7.1 |
| Write delay (ns) | 6.6 | 1.4 |
| Write energy (fJ) | 720 | 208 |
| Area (μm$^2$) | 0.029 | 0.029 |

**Table 4.2 Single 256Kbit sub-array performance (Δ=65 @ 85°C, 512 cells/BL, 512 cells/WL).**

## 4.3 SHE-MRAM as a L2 Cache Memory

### *4.3.1 Read Performance Boosting in SHE-MRAM*

For L3 or L4 cache, we can expect that SHE-MRAM, which has the same cell size as STT-MRAM, to have shorter access latency than SRAM since the access time of these larges caches is dominated by the global interconnect delay rather than the single subarray delay [97]. So the more interesting question is whether SHE-MRAM can outperform SRAM for smaller L2 caches with densities in the order of 1Mbit. Standard STT-MRAM could not outperform SRAM for smaller caches due to the long write delay, but the faster write coupled with the shorter global interconnect delay could potentially make SHE-MRAM a viable option for L2.

One unique advantage of SHE-MRAM is that the read delay can be reduced without a commensurate increase in write delay by simply increasing the thermal stability Δ. As shown in Fig. 4.7(a), SHE-MRAM shows less increase in switching current requirement than that of STT-MRAM along with the increased thermal stability. Since SHE-MRAM provides more efficient spin current generation, a switching overhead with higher thermal stability, which can be translated to an increase in write delay, is not significant compared to STT-MRAM in Fig. 4.7(b). Typically, high thermal stability allows a larger read current for a given read disturbance failure rate so that SHE-MRAM with over-designed thermal stability can boost read speed while mitigating an increase in the write delay as shown in Fig. 4.8. When the thermal stability increases from 65 to 85, SHE-MRAM can have 2.4-time faster read operation with a small increase in write delay less than 1ns.

This characteristic is contrary to standard STT-MRAM which has an inherent conflict between read and write delays.

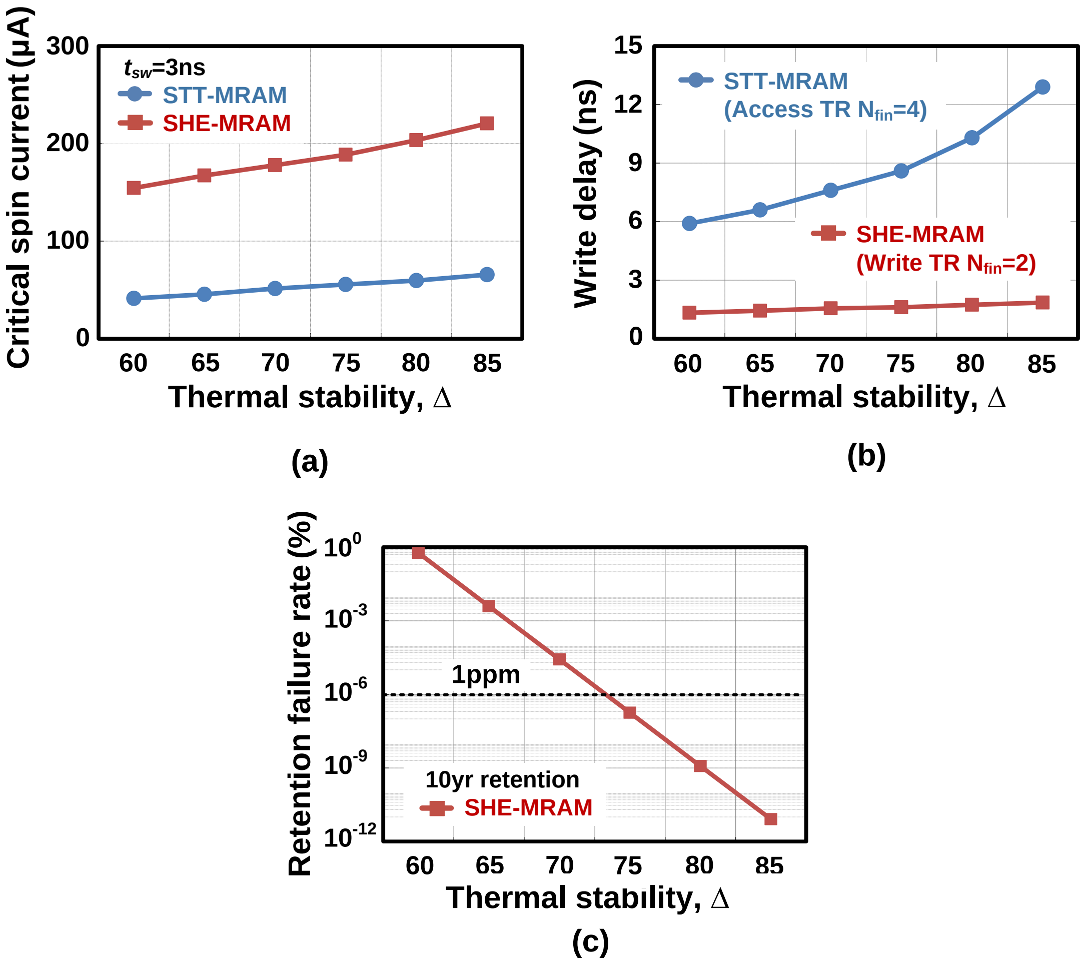
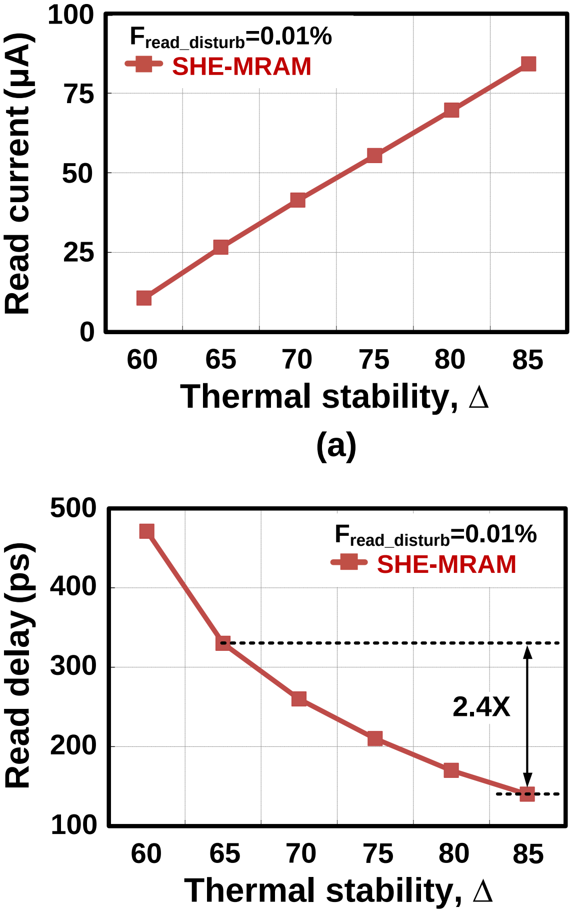


**Fig. 4.7 Impact of thermal stability factor on (a) critical spin current, (b) write delay, and (c) retention failure rate.**



(a)

(b)

**Fig. 4.8 Impact of thermal stability factor on (a) read current and (b) read delay of SHE-MRAM for a 0.01% read disturbance failure rate.**

| Metrics | SRAM | STT-MRAM | SHE-MRAM |
|---|---|---|---|
| Thermal stability (@85°C) | - | 65 | 85 |
| Bit-cell failure rate (%) | - | $10^{-2}$ | $10^{-11}$ |
| Read latency (ns) | 0.42 | 0.71 | 0.43 |
| Read energy (nJ) | 0.07 | 0.22 | 0.44 |
| Write latency (ns) | 0.42 | 6.77 | 1.95 |
| Write energy (nJ) | 0.10 | 0.41 | 0.21 |
| Leakage power (mW) | 39.5 | 4.96 | 4.96 |
| Area ($mm^2$) | 0.55 | 0.16 | 0.16 |

**Table 4.3 L2 cache performance summary (1Mbit, 8-way associativity, private bank, CACTI simulator [128]).**

### *4.3.2 Performance Benchmark for L2 Cache Memory*

In order to explore the feasibility of SHE-MRAM as a L2 cache memory, we compare the cache-level performance of SHE-MRAM with other embedded memory technologies such as SRAM and STT-MRAM. We consider a realistic cache configuration for 22nm technology node such as 1Mbit memory size, 8-way associativity, and private bank. We used CACTI, a widely accepted architecture simulator, for extracting the power and performance numbers of cache memory [128]. Based on the proposed read boosting scheme, we assume higher thermal stability for SHE-MRAM than that of STT-MRAM. As shown in Table 4.3, a SHE-MRAM based L2 cache with higher thermal stability has a read latency comparable to that of SRAM while maintaining a lower leakage power and denser area, which shows its promises as a L2 cache alternative. It should be noted however that a higher TMR and efficient sensing circuits are necessary to reduce the high read energy incurred by the current-forcing read of SHE-MRAM. When compared to standard STT-MRAM, SHE-MRAM shows better read and write performance even for L2 cache application ensuring much smaller bit-cell failure rate.

## 4.4 Chapter Summary

Since future scalability of STT-RAM is limited by its current based mechanism, novel switching mechanisms have been demonstrated with the common goal of reducing the switching current while maintaining sufficient nonvolatility. In particular, spin-Hall effect (SHE), which provides a significant reduction in the switching current with highly efficient spin current generation, has been recently drawing attention. In this chapter, we explore the feasibility and capability of SHE-MRAM for an on-chip memory, especially focusing on L2 cache application. Based on the realistic consideration from device to circuit and architecture levels, the performance of SHE-MRAM is compared with other embedded memory technologies such as SRAM and STT-MRAM. With the proposed read speed boosting scheme using trade-off points in design parameters such as thermal stability and read/write currents, SHE-MRAM shows a SRAM-competitive read latency for L2 cache maintaining the existing advantages such as low static power and compact macro size.

# Chapter 5 Neuromorphic Core Design with Multi-level Synapses Using Embedded Flash Memory

Recently, a neuromorphic computing mimicking human brain has been gaining lots of interests since it provides unique functionality such as perception, action, and recognition with high efficiency compared to traditional Boolean computing [129]. The key challenge in neuromorphic architecture is to create a highly efficient hardware design functionally equivalent to software models while achieving ultra-low power consumption, compact size, and high throughput. In this work, a new architecture of neuromorphic core using embedded flash (eflash) memory is designed in 65nm standard CMOS process. The eflash memory cells are used as storage of synaptic weights for a highly efficient implementation of restricted Boltzmann machine (RBM) which is a well-known neural algorithm for digit recognition. It also provides unique features such as zero static power and instant on/off operation without reloading weights due to nonvolatility. Our eflash cells store multi-level weights with different threshold voltages, which results in different levels of output current. In order to simplify neural computation, excitatory and inhibitory weights are separately stored in a pair of bitlines and the wordlines corresponding to input images are simultaneously activated. In this way, the spike generation can be completed by simply comparing two currents (i.e spike-out when the sum of excitatory weights is larger than that of inhibitory weights). Therefore, all the neuron processes such as weight multiplication, weight integration, and threshold comparison can be compressed into one-time current comparison without large ASIC implementation for digital neuron.

## 5.1 Overview of Neuromorphic Core Design

### *5.1.1 Artificial Neural Network (ANN) and Applications*

Artificial neural network (ANN) is a computational model inspired by biological neural networks, which can be used for approximate computing with a large number of unknown inputs. Fig. 5.1 shows the simplified biological neural network. Basically, electrical signals from transmitting neurons are integrated in a receiving neuron after multiplying weights at the synapse, which could be excitatory or inhibitory. Once integrated weights are larger than threshold, the neuron fires the spike to the next stage. An artificial neuron model captures core functions of biological neuron behavior as shown in Fig. 5.2.

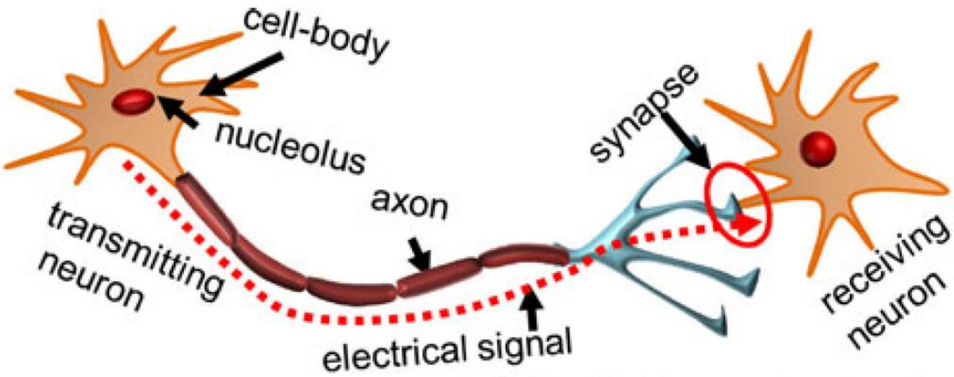


**Fig. 5.1 Simplified biological neural network [15].**

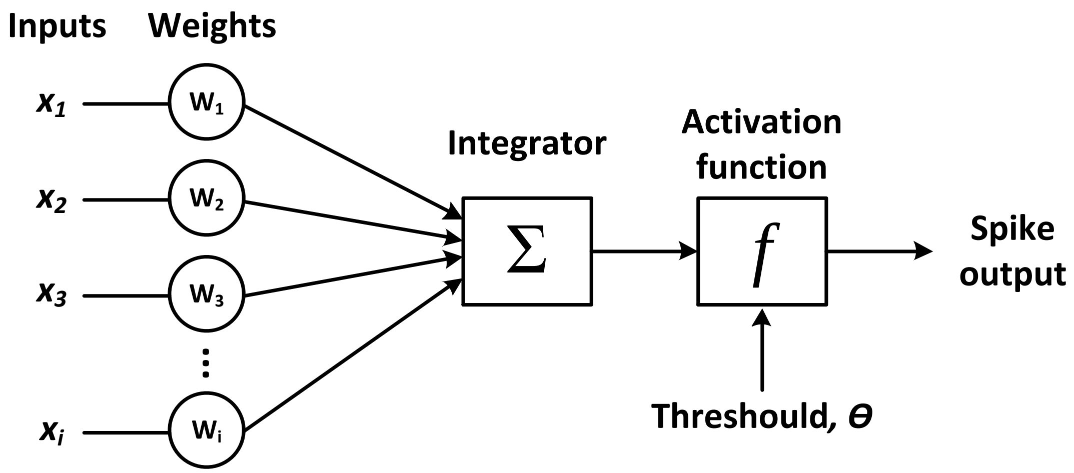


**Fig. 5.2 Artificial neuron model based on biological neuron behavior.**

Based on this simple neuron model, lots of algorithms have been developed for realization of human-like behaviors such as learning, recognition, action, and so on. As an example, the restricted Boltzmann machine so called RBM is a well-known neural algorithm for handwritten digit recognition [130], which is also chosen for our target algorithm in this work. Fig. 5.3 shows the overall flow of handwritten digit recognition, which consists of two main phases: training and validation. During training phase, weights are learnt from 60,000 handwritten digits from MNIST dataset [131] and programed into a neuromorphic core which is an actual hardware performing neuron operation. The various approaches for hardware implementation of neuromorphic core are presented in the following section. During validation, the neuromorphic core generates a spike signal based on the pixel data of test images (i.e. unknown inputs). Then, the classifier compares these spikes with reference, which was set during the learning phase, providing prediction results (i.e. probability values of each digit being a given input).

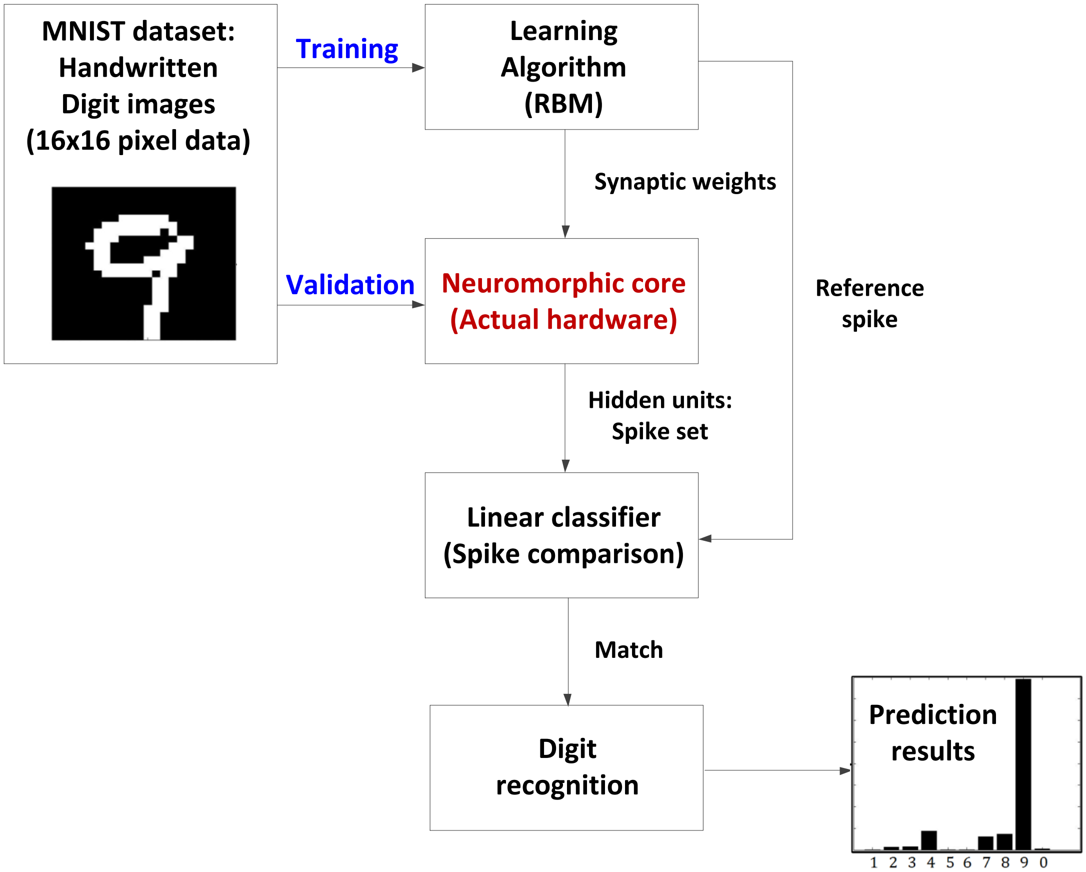


**Fig. 5.3 Overall flow of handwritten digit recognition using RBM algorithm.**

### *5.1.2 Prior Approaches for Neuromorphic Core Implementation*

In terms of neuromorphic core implementation, a challenge is to design highly efficient hardware functionally equivalent to software models while providing low power, compact size and high throughput. In addition, the synaptic weights need to be encoded in analog fashion for high prediction accuracy maintaining weight levels for a long time. Therefore, a choice of devices for a synapse is one of the most critical issues in the neuromorphic core implementation.

In order to accommodate these requirements, analog synapses are demonstrated by using capacitors to store charge [132], [133], which can translate the amount of charge on the capacitor into finely defined weight levels. However, this analog approach shows limited correspondence between software and hardware models due to low immunity to process and temperature variations. Moreover, a large area overhead due to charge capacitor and complex analog circuitry limits the total number of synapse in the core, and a charge leakage in the capacitor increases a chance of misreading weight levels requiring additional circuits to compensate it.

For more efficient implementation of synaptic plasticity, nonvolatile memory (NVM) technologies such as floating-gate (FG) transistor, phase-change memory (PCM), and resistive RAM (RRAM) have been utilized as a storage element for synaptic weights [134]-[136]. Since those memory cells provide gradual resistance modulation, a multi-level synapse can be readily realized achieving a compact bit-cell size. However, those technologies require additional fabrication steps beyond stand logic process to integrate memory cells. Moreover, in terms of practical implementation, it is still an issue to achieve a robust multi-level programming due to low controllability of the heat diffusion and filament formation in PCM and RRAM, respectively, which critically decides the resistance levels.

Recently, fully-digital implementations using SRAM-based synapses and ASIC-based neuron circuits have been demonstrated with advanced CMOS technologies ensuring one-to-one correspondence between hardware and software models [137]-[139]. However, those demonstrations still require a large size of memory and static power along with repeated weight read and integration to generate spike signals (i.e. low throughput). The further details on SRAM-based neuromorphic core design will be presented in section 5.3.1.

## 5.2 Logic-Compatible 5T-Eflash

This section presents the details on logic-compatible single-poly eflash memory cells to provide background for the proposed eflash-based neuromorphic core design. Typically, the dual-poly eflash cells have two stacked gates in a single transistor: floating gate (FG) and Control Gate (CG). However, this device requires additional processes beyond a standard CMOS logic process due to FG formation, thick tunnel oxide and high voltage circuits to support program and erase operations. On the other hand, single-poly eflash, which can be fabricated without any process overhead, is considered as a promising candidate for cost-effective moderate density (e.g. few kilobits) non-volatile storage solution. In this chapter, we choose a logic-compatible 5T-eflash proposed in [140]-[142] as a target device for neuromorphic core application.

### *5.2.1 Device Concepts of 5T-eflash*

Fig. 5.4 shows the schematic of 5T-eflash unit cell structure and the bird's eye view of three core transistors used in [140], [141]. Here, $M_1$ is the coupling device, $M_2$ is the erase device, $M_3$ is the program/read device, and $S_1$ and $S_2$ are the selection devices for array operations such as inhibition operation. Moreover, all these transistors ($M_1$, $M_2$, $M_3$, $S_1$, $S_2$) in the unit cell are implemented using standard 2.5V I/O transistors, which have a tunnel oxide thickness ($T_{OX}$) of 5nm. In order to form FG node, the gate terminals of the three devices $M_1$-$M_3$ are connected in a back-to-back fashion. The PMOS n-well of $M_1$ and $M_2$ functions like the CG of the dual-poly eflash. Note that the high coupling ratio between the n-well of the coupling device and FG can be achieved by upsizing the coupling transistor width, which is 8 times larger than that of both $M_2$ and $M_3$ achieving a high enough coupling ratio for effective erase and program operations.

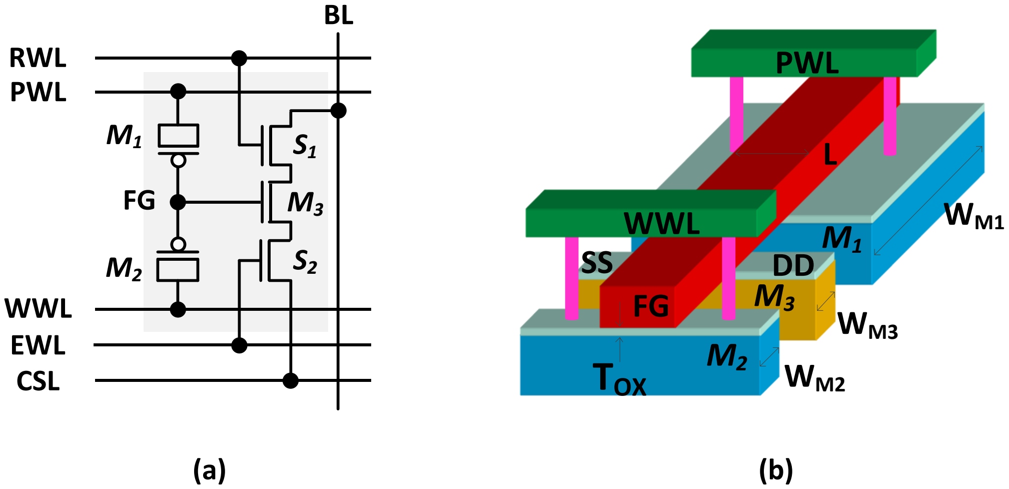


**Fig. 5.4 Logic-compatible 5T-eflash proposed in [141]. (a) Unit cell schematic of 5T-eflash. (b) Bird's eye view of three core transistors (*$M_1$*, *$M_2$*, and *$M_3$*).**

### *5.2.2 Array-Level Operations*

Similarly to dual-poly flash operations, the 5T-eflash requires four operation modes: erase, program, inhibit, and read. Fig. 5.5 shows the bias conditions for erase and program operations of the 5T-eflash cell [141]. During erase operation, as a Write-Word-Line (WWL) is selected, a high voltage pulse is applied to $M_2$ with 0V bias to Program-Word-Line (PWL). Due to high electrical field between $M_1$ and $M_2$, electrons are ejected from FG through Fowler-Nordheim (FN) tunneling. During program operation, a high voltage pulse is applied to both PWL and WWL boosting the FG node voltage to inject electrons from $M_3$ channel to FG node. Here, half-selected cells sharing the same WWL and PWL experiences program disturbance, which leads to unwanted partial programming (i.e. $V_{th}$ shift) in unselected cells due to high WL voltage. In order to inhibit this disturbance issue during program operation, we make the channel of half-selected cells floating by turning off the two selection transistors, which is called self-boosting scheme. In this way, the floating channel voltage is boosted by high WL voltage inducing an insufficient voltage difference between WL and $M_3$ channel for FN tunneling. Fig. 5.6 shows the bias condition and timing diagram for read operation. First, BLs are pre-charged to the supply level (i.e. 1.2V) and PWL and WWL are set to the read reference level (VRD) inducing the channel in the $M_2$. Once the selection transistors ($S_1$, $S_2$) are activated, the pre-charged BL levels are discharged at different rates depending on the cell data (i.e. programmed or erased cell). Then, voltage sense amplifier compares the BL levels to a reference voltage to provide an output signal in logic level.

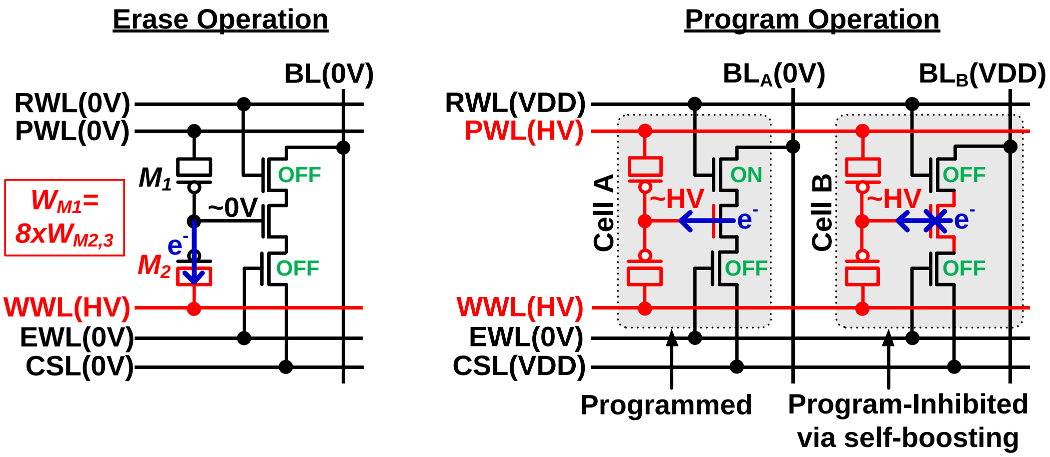


**Fig. 5.5 Bias conditions for erase and program operations of the 5T-eflash cell [141].**

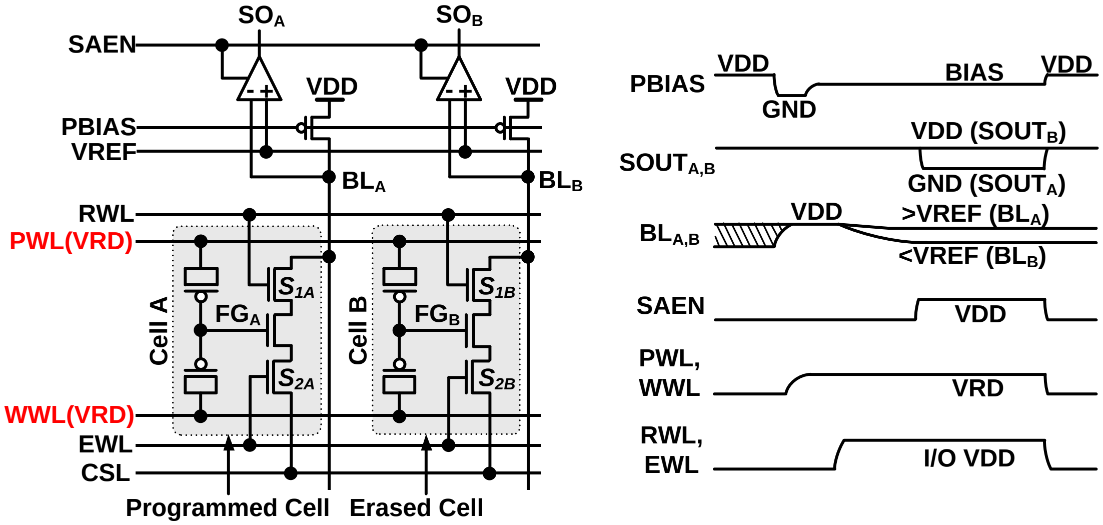


**Fig. 5.6 (left) Bias condition and (right) timing diagram during read operation [141].**

## 5.3 Design Concept of Eflash-Based Neuromorphic Core

### *5.3.1 Prior Art: SRAM-Based Design*

Typically, RBM algorithm includes the weight multiplication along with incoming axon signals, which can be described as follows: $y_i = \sum x_i w_i$ . Here, $x_i$ represents the input information to the neuromorphic core, and $w_i$ represents the synapse weight. When the accumulated output $y$ reaches a predefined threshold, a spike output is generated, which is equivalent to the behavior of biological neurons. To mimic this neuron behavior, the previous neuromorphic designs utilize SRAM-based crossbar array architecture for synapse and digital logic neuron circuits for weight integration, threshold comparison, and spike generation.

As shown in Fig. 5.7, the pixel information of input images is fed to the core in the form of digital sequence (i.e. axon activity in the figure) activating the corresponding wordlines. The single-level weights stored in a SRAM column are read in row-by-row fashion and integrated by a digital neuron which consists of ASIC-based sub-blocks such as multi-bit adder, accumulator, and comparator [138]. However, in this approach, the row-by-row access for weight readout and digital logic based neurons require a large memory and incur significant power and delay overheads due to the repeated weight readout and summation operation for generating spikes.

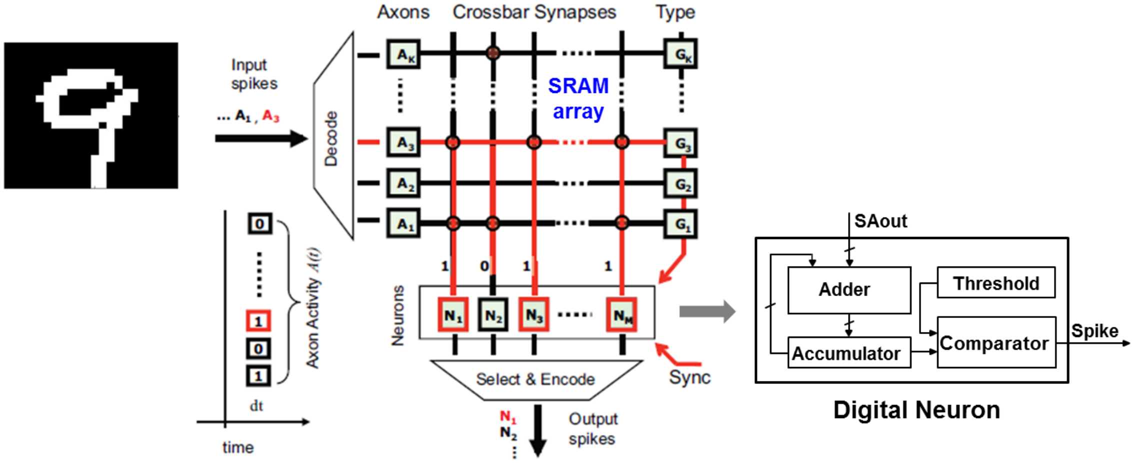


**Fig. 5.7 SRAM-based neuromorphic core design proposed in [138].**

### *5.3.2 Proposed Eflash-Based Design*

In this work, we propose eflash-based neuromorphic core design that can provide a highly efficient implementation of the RBM algorithm. We utilize 5T-eflash cells, which is introduced in chapter 5.2, to store the synaptic weights in a neuromorphic crossbar array. By using this logic-compatible non-volatile memory cells, our neuromorphic core provides unique features such as zero static power and instant on/off operation without the need to reload the synaptic weights, which are not provided in conventional SRAM-based design. Moreover, the eflash cells can store multi-level weights which results in better accuracy of the neural network algorithm. Another notable feature of the proposed design is that all the neural computing such as weight multiplication, weight integration, and spike generation can be performed in a single cycle by activating multiple wordline signals (i.e. axons) for a given input pixel data.

Fig. 5.8 illustrates the design concept of eflash-based neuromorphic core. In most of the previous designs, excitatory and inhibitory weights stored in a single bitline are integrated by row-by-row access and compared with a pre-defined threshold value using large digital logic circuits. On the other hand, our design separately stores excitatory and inhibitory weights in a pair of bitlines. A threshold values are also stored in the additional eflash cells in the form of weights. The cells for negative threshold value are placed on the bitline used for excitatory weights while the cells for positive threshold value are placed on the bitline used for inhibitory weights. Simply, this is for a comparison between positive weight sum and negative weight sum. Based on this cell arrangement, all the selected wordlines are activated together corresponding to the pixel input data, and hence the sum of individual currents from all the activated synaptic cells in a column flows through a bitline. Therefore, a pair of bitlines for excitatory and inhibitory weights generates two currents: excitatory and inhibitory current. The neuron circuit, which compares these two currents, decides whether a spike is generated in the neuron. In this way, a spike can be generated with a single operation by comparing cell currents from a pair of bitlines without large digital logic circuits. This makes reading operation faster than the conventional row-by-row reading schemes by a factor proportional to the number of synaptic weights per bitline (e.g. 64 or 128).

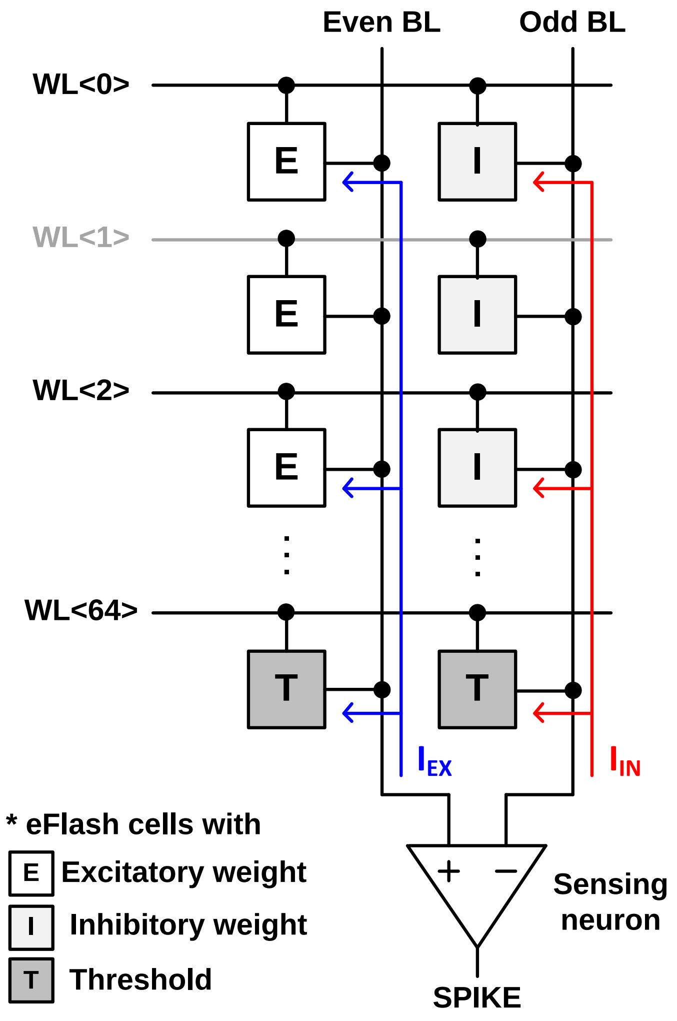


**Fig. 5.8 High throughput eflash-based neuromorphic core design utilizing current integration and current comparison for spike generation.**

## 5.4 Neuromorphic Core Architecture and Circuit Design

### *5.4.1 Overall Architecture and Key Design*

In Fig. 5.9, single core architecture is presented, which features high voltage switch (HVS) for wordline driving, neuron sensing circuits for current verification and spike generation, and multiple scan chains for data input/output.

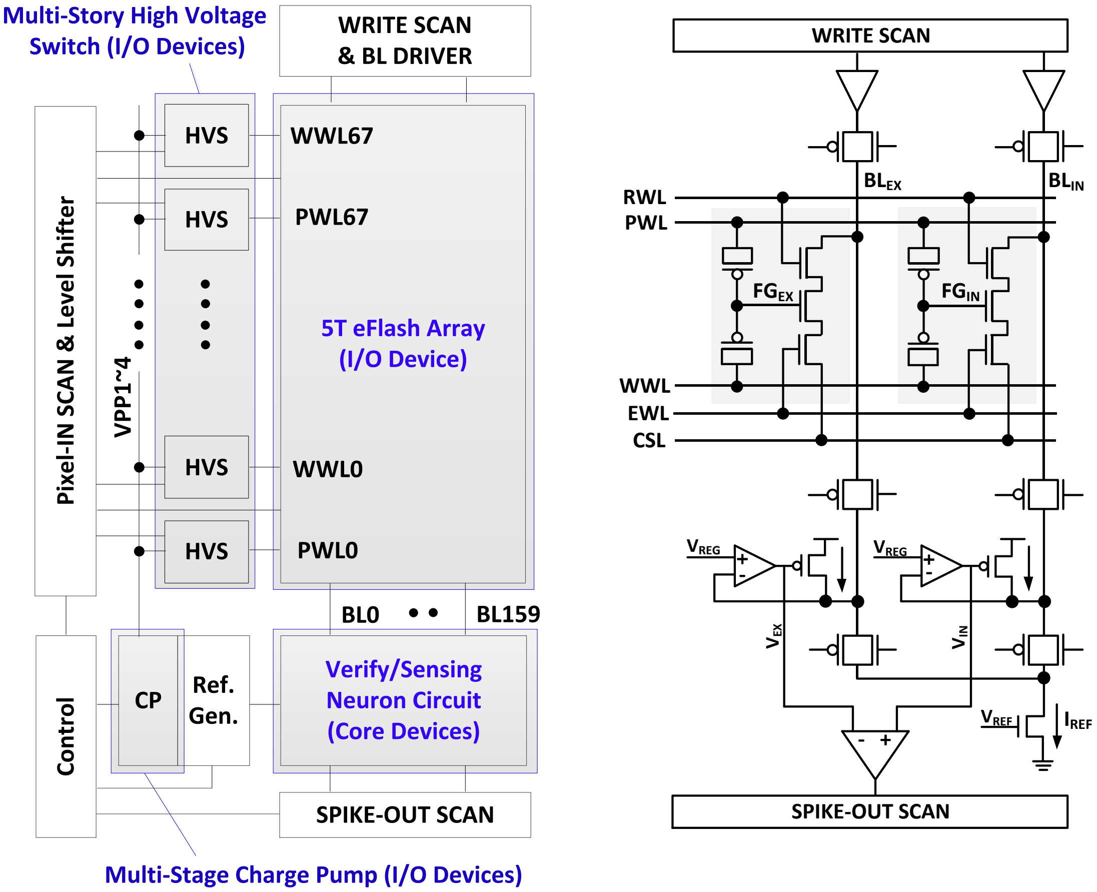


**Fig. 5.9 (left) Overall core architecture featuring logic-compatible high voltage switch and neuron sensing circuit and (right) a simplified schematic for a single column circuit.**

Our design uses 4 cores to accommodate 16x16 pixel input and 256 neurons. Based on this setting, when a pair of excitatory and inhibitory cells store 5-level weights (e.g. 0/10/20μA current levels for both excitatory and inhibitory cells), the recognition accuracy from software model is expected as 93%. During program and erase operation, high wordline voltage around 10V is needed so that critical design point is to implement HVS without overstress within logic technology. In this work, we used HVS using multi-story latches for all transistors to operate within the nominal I/O voltage as presented in [140], [141].

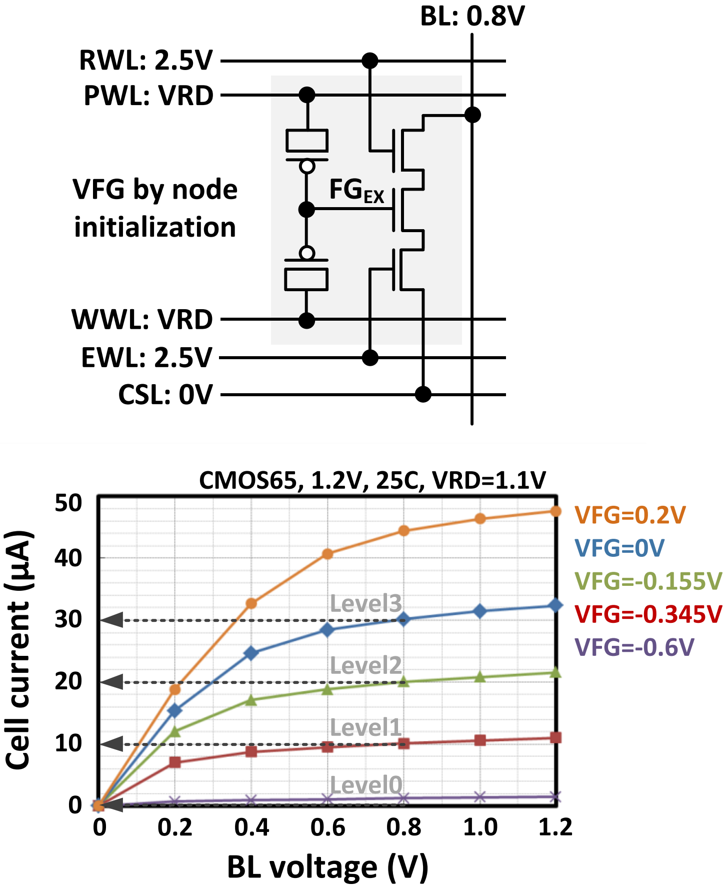


**Fig. 5.10 Eflash output characteristics with different FG node voltages.**

For current-based neuron operation, we first need to program proper current levels into the cells by adjusting the threshold voltage. Based on the output characteristics with different FG node voltages in Fig. 5.10, 10uA of saturation current is chosen as unit current for multi-level operation while fixing bitline and wordline voltages.

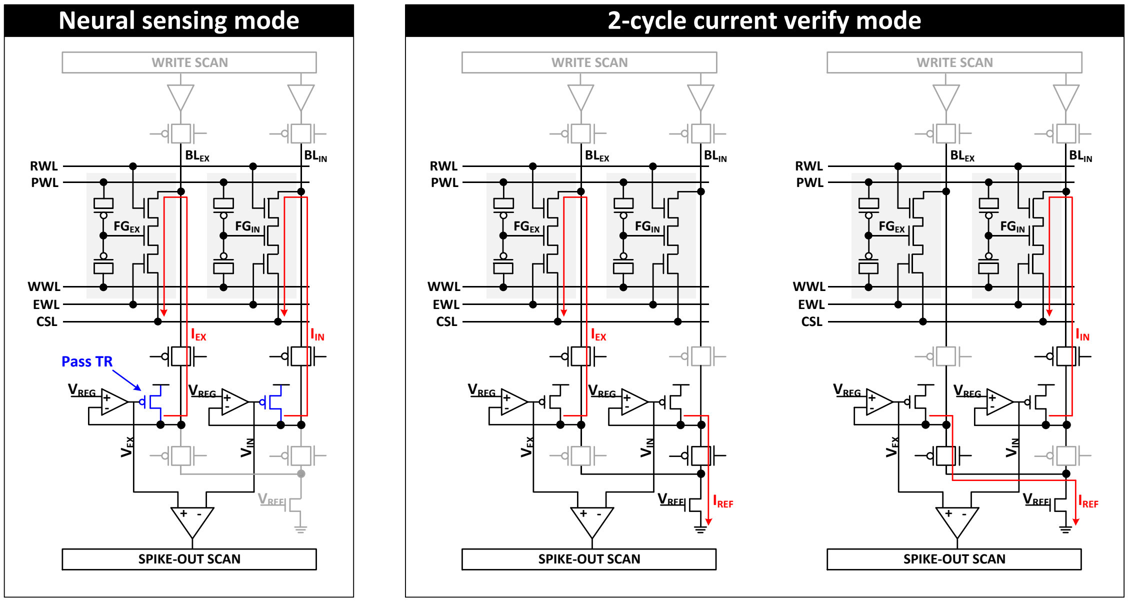


**Fig. 5.11 Neuron circuit operations during (left) current comparison for spike generation and (right) current-verify for weight programming.**

In order to draw precise currents according to weight levels even if each time a different number of cells are activated, a bitline voltage needs to be maintained during the neural sensing operation. Moreover, two currents from excitatory and inhibitory bitlines should be compared in a reliable way such as voltage sensing. To meet these requirements, our neuron circuit utilizes a load regulation to fix a bitline voltage regardless of amount of current drawn in the bitline while sensing the current difference between excitatory and inhibitory bitlines for spike generation. As shown in Fig. 5.11, when the multiple wordlines are activated along with given pixel inputs, the corresponding cell current drops the bitline voltage and a voltage regulator adjusts a pass-transistor gate voltage to match the bitline voltage with regulation voltage ($V_{REG}$). Here, a current-to-voltage conversion is achievable in the pass-transistor gate nodes. Since the pair of pass-transistor gate node voltages change depending on the bitline currents, those node voltages can be indirectly used for current comparison making it possible to use a voltage sense amplifier for robust sensing operation. This neuron circuit is also useful during current-verify operation. In this case, either of two bitline currents is compared to reference current. To provide symmetric reference circuits for two bitlines, 2-cycle verify operation is used by controlling transmission gates.

### *5.4.2 Neuromorphic Core Operations*

Our neuromorphic core mainly operates in two phases: weight loading and neural sensing. The detailed operation sequence is presented in Fig. 5.12. During weight loading phase, excitatory and inhibitory weights generated from RBM algorithm are loaded to the eflash core. Like typical eflash memory operations, after initial erase operation, program-verify operations are repeated to reach the target cell current. Here, cell threshold voltage is increased until cell current is close to the reference current (i.e. lower level weights are encoded with higher threshold voltages). During neural sensing phase, the pixel data of test image is fed to the core and activates multiple wordlines, generating spike signals in the neuron circuits. In Fig. 5.13, the block diagrams for eflash-based neuromorphic core operations are presented. Note that WL_SCAN selects a single wordline during program and erase modes like a decoder while it selects multiple wordlines during neural sensing mode. For the test purpose, external verify and neural sensing modes are considered by adding a BL_DEC_EXT block. When this external measurement option is on, BL_SCAN selects a bitline like a decoder and connects the selected bitline to the external pad, which enables current monitoring.

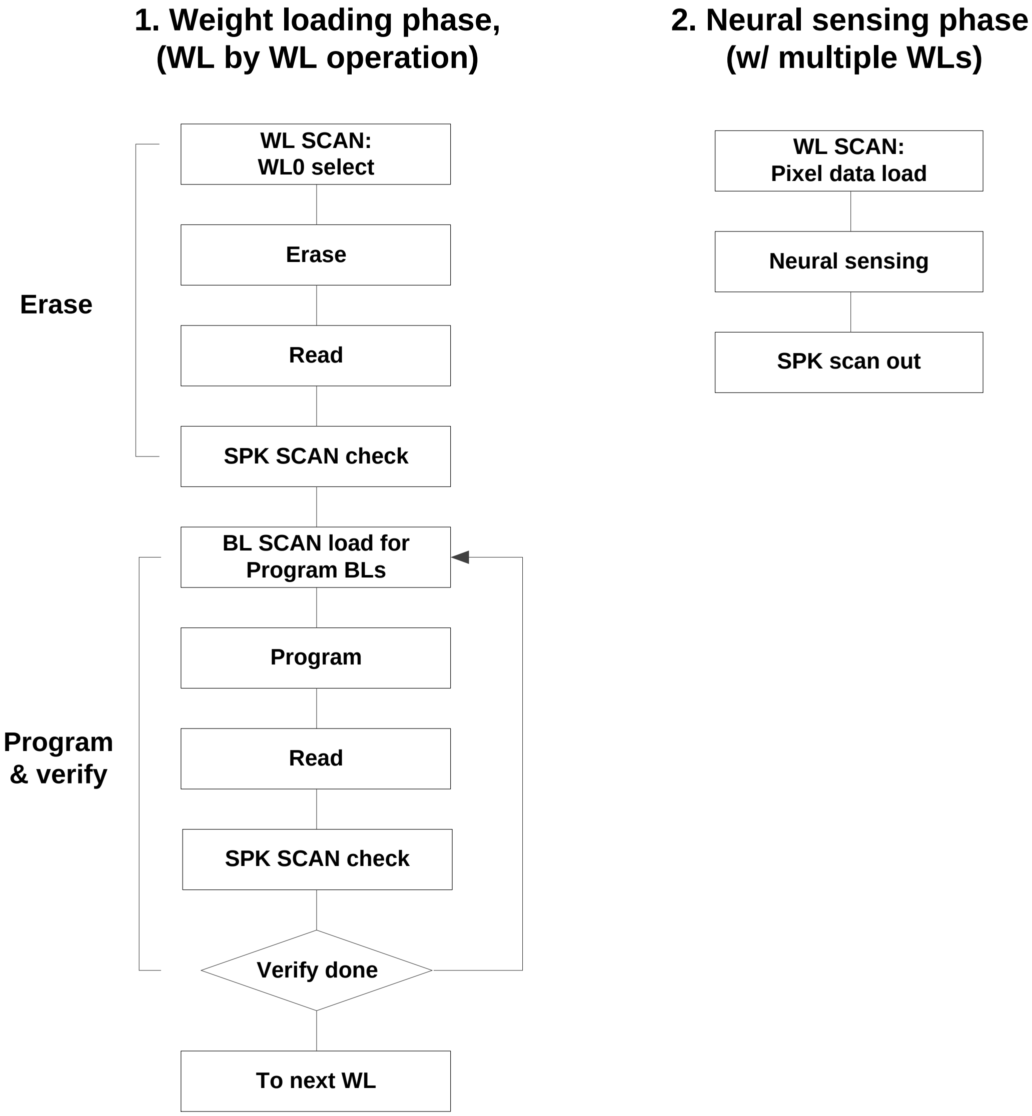
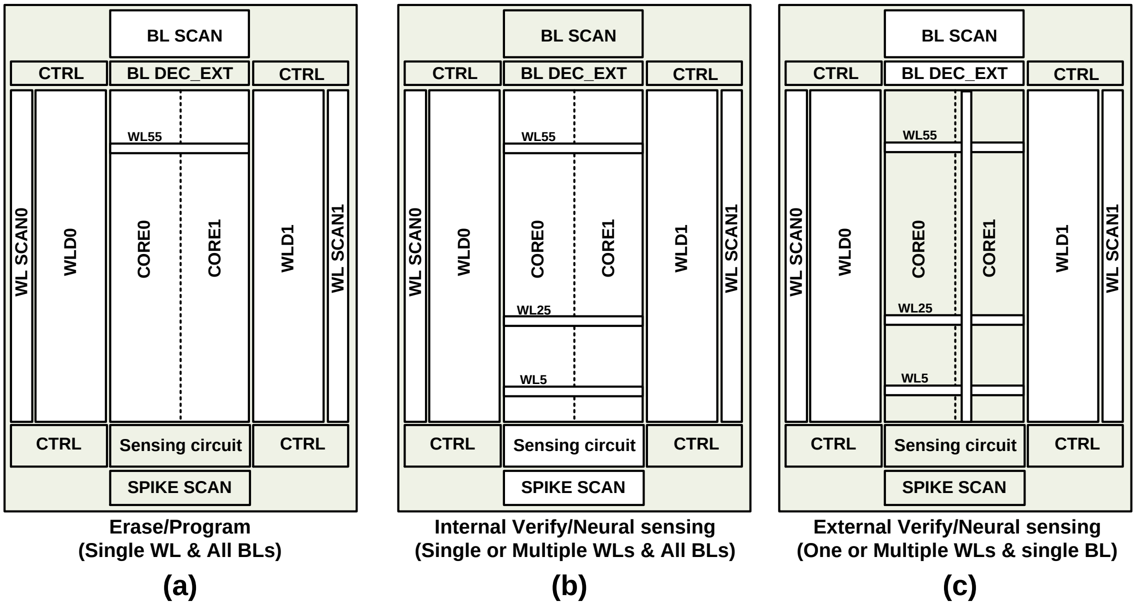


**Fig. 5.12 Neuromorphic core operation sequences for weight loading and neural sensing.**



Erase/Program
(Single WL & All BLs)
(a)

Internal Verify/Neural sensing
(Single or Multiple WLs & All BLs)
(b)

External Verify/Neural sensing
(One or Multiple WLs & single BL)
(c)

**Fig. 5.13 Block diagrams for eflash-based neuromorphic core operation modes. (a) Erase and Program modes. (b) Internal verify and neural sensing modes. (c) External verify and neural sensing modes for test feature.**

### *5.4.3 Simulation Results*

Fig. 5.14 shows the top-level simulation waveforms using 65nm standard CMOS process. For a given operation mode, wordline signals are generated boosting FG node. During verify and neural sensing mode, the bitline voltages are regulated inducing sensing voltages at the pass-gate transistor gate nodes, and then comparison results are returned as a spike signal. Fig. 5.15 shows the full chip layout of our eflash-based neuromorphic core.

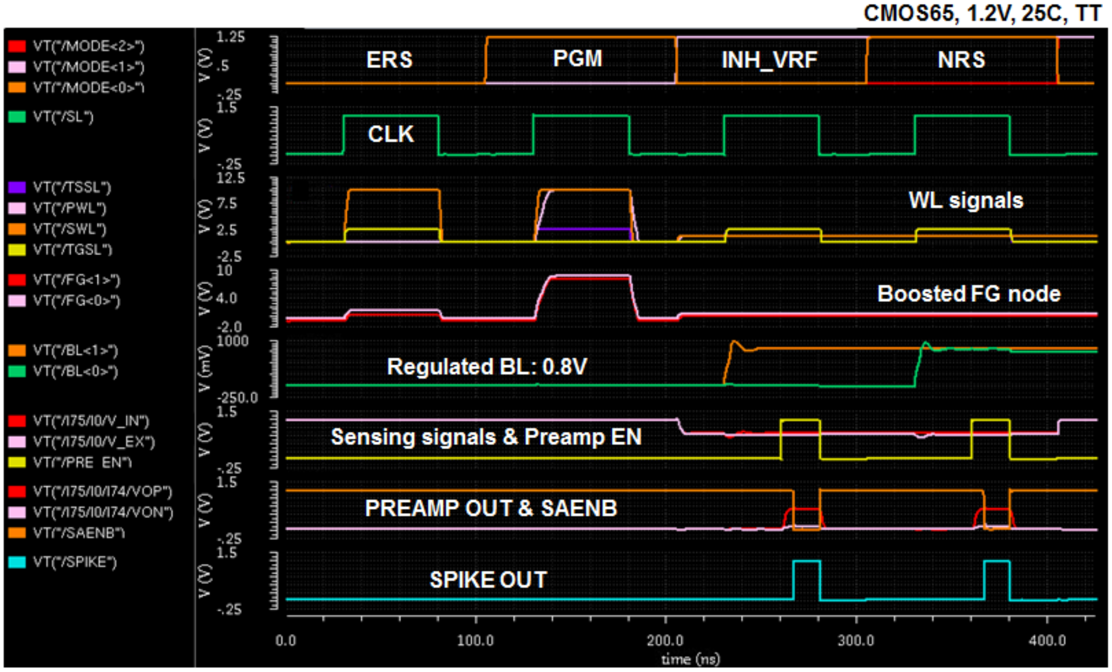


**Fig. 5.14 Top-level simulation waveforms in 65nm standard CMOS technology.**

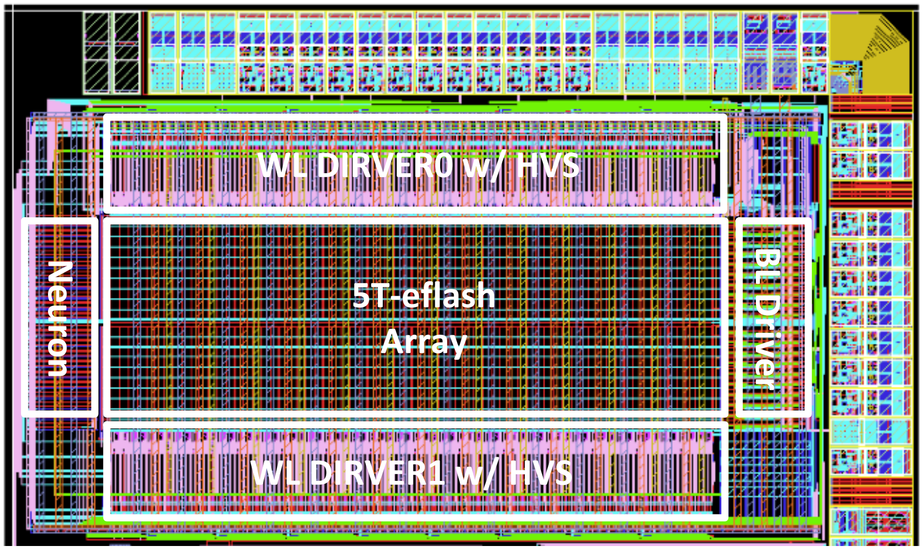


**Fig. 5.15 Full chip layout of eflash based neuromorphic core implemented in 65nm standard logic process (total size: 1100x600µm$^2$).**

## 5.4 Chapter Summary

In this chapter, we propose the eflash-based neuromorphic core, which allows highly efficient implementation of RBM algorithm for handwritten digit recognition. The eflash cell itself provides multi-level weights, zero static power and instant on/off operation without reloading weights, which are unattainable in conventional SRAM-based designs. Moreover, our core provides a high throughput operation by programming the excitatory and inhibitory weights separately in a pair of bitlines, which enables the weight integration with current. Based on this scheme, overall neuron operation can be accomplished by current comparison in a single cycle without large digital neuron circuits used in previous works.

# Chapter 6 Conclusion

Aggressive scaling of CMOS technology towards ultimate physical limit is uncertain due to unsupportable increases in power density and exponentially growing leakage current. These constraints necessitate the search for low-power alternatives to continue functional scaling in a post-CMOS era. The spin-based technology, which utilizes electron spin direction as the state variable for information processing, has been investigated with the promise of low power computing coupled with nonvolatility, excellent scalability and non-Boolean architecture.

In this thesis, we evaluate a system level power and performance of spin-based logic and memory technologies taking realistic design considerations into account, in order to prove their practical potential beyond what is achievable by a single device alone.

In chapter 2, we provided a comprehensive analysis on spin-based logic technology encompassing all levels of design abstractions such as device, circuit and architectural aspects. We chose the ASL as a target device and their system-level power performance was evaluated on a hypothetical Intel Core i7 processor. Even with promising advantages such as zero static power, lower device count and lower supply voltage, the technical barriers such as a large switching current, spin attenuation in the channel, and high activity factor need to be resolved for a practical use. We believe fundamental principles and methodologies established in this work will help pave the way for rapid realization of spin-based logic technology.

In chapter 3, we explored the scalability of MTJ technologies to predict the best option for future STT-MRAM. A dedicated SPICE MTJ model was developed specifically for an extensive scaling analysis. Based on the realistic device dimensions and material parameters, we provided the detailed scaling methods and projection scenarios down to 7nm, and then compared the key performance metrics such as write and read delays. We found that relaxed MTJ width helps continued scaling of switching current and i-PMTJ shows a balanced performance in write and read operations compared to other technologies such as IMTJ and c-PMTJ.

In chapter 4, along with standard STT-MRAM, we also explored the feasibility of non-traditional MRAMs such as SHE-MRAM for on-chip cache application. Since its charge to spin conversion ratio is higher than 100%, it can lower switching current significantly without disturbing nonvolatility. We found that this high switching efficiency can be utilized to boost read speed without area overhead by simply increasing thermal stability. Based on cache-level simulation, our SHE-MRAM achieved a SRAM-comparable read latency showing a promise for L2 cache memory application.

In chapter 5, we proposed high-throughput neuromorphic core using a logic-compatible 5T-eflash as a synapse and the test chip was designed in a 65nm standard logic process. By using nonvolatile eflash cells, our core provides multi-level weights, zero static power, and instant on/off operation without reloading weights, which were not available in previous designs. Our new architecture using current-based weight integration and spike decision provides high-speed neuron operation without ASIC-based large digital neuron circuits.

# Bibliography

[1] G. E. Moore, "Cramming more components onto integrated circuits (Reprinted from Electronics, pg 114-117, April 19, 1965)," *Proc. IEEE*, vol. 86, no. 1, pp. 82-85, Jan. 1998.

[2] R. R. Schaller, "Moore's law: past, present and future," *IEEE Spectrum*, vol. 34, no. 6, pp. 52-591997.

[3] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S. H. Lo, G. A. SaiHalasz, R. G. Viswanathan, H. J. C. Wann, S. J. Wind, and H. S. Wong, "CMOS scaling into the nanometer regime," *Proc. IEEE*, vol. 85, no. 4, pp. 486-504, Apr. 1997.

[4] H. S. P. Wong, D. J. Frank, P. M. Solomon, C. H. J. Wann, and J. J. Welser, "Nanoscale CMOS," *Proc. IEEE*, vol. 87, no. 4, pp. 537-570, Apr. 1999.

[5] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, no. 4, pp. 23-29, Jul-Aug. 1999.

[6] N. S. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J. S. Hu, M. J. Irwin, M. Kandemir, and V. Narayanan, "Leakage current: Moore's law meets static power," *IEEE Computer*, vol. 36, no. 12, pp. 68-752003.

[7] T. N. Theis, and P. M. Solomon, "It's time to reinvent the transistor!," *Science*, vol. 327, no. 5973, pp. 1600-1601, Mar. 2010.

[8] C. Y. Chang, "The highlights in the nano world," *Proc. IEEE*, vol. 91, no. 11, pp. 1756-1764, Nov. 2003.

[9] K. Galatsis, A. Khitun, R. Ostroumov, K. L. Wang, W. R. Dichtel, E. Plummer, J. F. Stoddart, J. I. Zink, J. Y. Lee, Y. H. Xie, and K. W. Kim, "Alternate state

variables for emerging nanoelectronic devices," *IEEE Trans. Nanotechnol.*, vol. 8, no. 1, pp. 66-75, Jan. 2009.

[10] S. A. Wolf, D. D. Awschalom, R. A. Buhrman, J. M. Daughton, S. von Molnar, M. L. Roukes, A. Y. Chtchelkanova, and D. M. Treger, "Spintronics: A spin-based electronics vision for the future," *Science*, vol. 294, no. 5546, pp. 1488-1495, Nov 16. 2001.

[11] I. Zutic, J. Fabian, and S. Das Sarma, "Spintronics: Fundamentals and applications," *Rev. Mod. Phys.*, vol. 76, no. 2, pp. 323-410, Apr. 2004.

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

[13] S. A. Wolf, J. W. Lu, M. R. Stan, E. Chen, and D. M. Treger, "The promise of nanomagnetics and spintronics for future logic and universal memory," *Proc. IEEE*, vol. 98, no. 12, pp. 2155-2168, Dec. 2010.

[14] A. Sarkar, S. Srinivasan, B. Behin-Aein, and S. Datta, "Modeling all spin logic: Multi-magnet networks interacting via spin currents," in Proc. *IEEE Int. Electron Device Meeting*, Dec. 5-7, 2011, pp. 11.11.11-11.11.14.

[15] M. Sharad, C. Augustine, G. Panagopoulos, and K. Roy, "Spin-based neuron model with domain-wall magnets as synapse," *IEEE Trans. Nanotechnol.*, vol. 11, no. 4, pp. 843-8532012.

[16] D. E. Nikonov, and I. A. Young, "Overview of beyond-CMOS devices and a uniform methodology for their benchmarking," *Proc. IEEE*, vol. 101, no. 12, pp. 2498-2533, Dec. 2013.

[17] J. Kim, A. Paul, P. A. Crowell, S. J. Koester, S. S. Sapatnekar, J.-P. Wang, and C. H. Kim, “ Spin-based computing: device concepts, current status, and a case study on a high-performance microprocessor,” *Proc. IEEE*, vol. 103, no. 1, pp. 106-130, Jan. 2015.

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

[19] M. Julliere, "Tunneling between ferromagnetic-films," *Phys. Lett. A*, vol. 54, no. 3, pp. 225-226, Sep. 1975.

[20] M. N. Baibich, J. M. Broto, A. Fert, F. N. Vandau, F. Petroff, P. Eitenne, G. Creuzet, A. Friederich, and J. Chazelas, "Giant magnetoresistance of (001)Fe/(001) Cr magnetic superlattices," *Phys. Rev. Lett.*, vol. 61, no. 21, pp. 2472-2475, Nov. 1988.

[21] B. Dieny, V. S. Speriosu, B. A. Gurney, S. S. P. Parkin, D. R. Wilhoit, K. P. Roche, S. Metin, D. T. Peterson, and S. Nadimi, "Spin-valve effect in soft ferromagnetic sandwiches," *J. Magn.Magn.Mater.*, vol. 93, pp. 101-104, Feb. 1991.

[22] J. S. Moodera, L. R. Kinder, T. M. Wong, and R. Meservey, "Large magnetoresistance at room-temperature in ferromagnetic thin-film tunnel-junctions," *Phys. Rev. Lett.*, vol. 74, no. 16, pp. 3273-3276, Apr. 1995.

[23] T. Ching, R. E. Fontana, T. Lin, D. E. Heim, V. S. Speriosu, B. A. Gurney, and M. L. Williams, "Design, fabrication and testing of spin-valve read heads for high density recording," *IEEE Trans. Magn.*, vol. 30, no. 6, pp. 3801-3806, Nov. 1994.

[24] M. Dax, "The non-volatile memory challenge," *Semicond. Int.*, vol. 20, no. 10, pp. 84-92, Sep. 1997.

[25] R. E. Scheuerlein, "Magneto-resistive IC memory limitations and architecture implications," in Proc. *IEEE Intl. Nonvolatile Memory Technology Conference*, Jun. 22-24, 1998, pp. 47-50.

[26] J. C. Slonczewski, "Current-driven excitation of magnetic multilayers," *J. Magn.Magn.Mater.*, vol. 159, no. 1-2, pp. L1-L7, Jun. 1996.

[27] T. Kawahara, K. Ito, R. Takemura, and H. Ohno, “Spin-transfer torque RAM technology: Review and prospect,” *Microelectronics Reliability*, vol. 52, no. 4, pp. 613-627, Apr. 2012.

[28] D. Apalkov, A. Khvalkovskiy, S. Watts, V. Nikitin, X. Tang, D. Lottis, K. Moon, X. Luo, E. Chen, A. Ong, A. Driskill-Smith, and M. Krounbi, “Spin-transfer torque magnetic random access memory (STT-MRAM),” *ACM Journal on Emerging Technologies in Computing Systems*, vol. 9, no. 2, pp. 13:1-13:35, May 2013.

[29] J. Z. Sun, "Spin-current interaction with a monodomain magnetic body: A model study," *Phys. Rev. B*, vol. 62, no. 1, pp. 570-578, Jul. 2000.

[30] J. A. Katine, F. J. Albert, R. A. Buhrman, E. B. Myers, and D. C. Ralph, "Current-driven magnetization reversal and spin-wave excitations in Co/Cu/Co pillars," *Phys. Rev. Lett.*, vol. 84, no. 14, pp. 3149-3152, Apr. 2000.

[31] W. H. Butler, X. G. Zhang, T. C. Schulthess, and J. M. MacLaren, "Spin-dependent tunneling conductance of Fe/MgO/Fe sandwiches," *Phys. Rev. B*, vol. 63, no. 5, Feb. 2001.

[32] S. S. P. Parkin, C. Kaiser, A. Panchula, P. M. Rice, B. Hughes, M. Samant, and S. H. Yang, "Giant tunnelling magnetoresistance at room temperature with MgO (100) tunnel barriers," *Nature Mater.*, vol. 3, no. 12, pp. 862-867, Dec. 2004.

[33] N. Nishimura, T. Hirai, A. Koganei, T. Ikeda, K. Okano, Y. Sekiguchi, and Y. Osada, "Magnetic tunnel junction device with perpendicular magnetization films for high-density magnetic random access memory," *J. Appl. Phys.*, vol. 91, no. 8, pp. 5246-5249, Apr. 2002.

[34] S. Ikeda, K. Miura, H. Yamamoto, K. Mizunuma, H. D. Gan, M. Endo, S. Kanai, J. Hayakawa, F. Matsukura, and H. Ohno, "A perpendicular-anisotropy CoFeB-

MgO magnetic tunnel junction," *Nature Mater.*, vol. 9, no. 9, pp. 721-724, Sep. 2010.

[35] M. Weisheit, S. Fahler, A. Marty, Y. Souche, C. Poinsignon, and D. Givord, "Electric field-induced modification of magnetism in thin-film ferromagnets," *Science*, vol. 315, no. 5810, pp. 349-351, Jan. 2007.

[36] M. I. Dyakonov, and V. I. Perel, "Current-induced spin orientation of electrons in semiconductors," *Phys. Lett. A*, vol. 35, no. 6, pp. 459-460, Jul. 1971.

[37] M. Johnson, and R. H. Silsbee, "Interfacial charge-spin coupling: Injection and detection of spin magnetization in metals," *Phys. Rev. Lett.*, vol. 55, no. 17, pp. 1790-1793, Oct. 1985.

[38] F. J. Jedema, A. T. Filip, and B. J. van Wees, "Electrical spin injection and accumulation at room temperature in an all-metal mesoscopic spin valve," *Nature*, vol. 410, no. 6826, pp. 345-348, Mar. 2001.

[39] T. Yang, T. Kimura, and Y. Otani, "Giant spin-accumulation signal and pure spin-current-induced reversible magnetization switching," *Nature Phys.*, vol. 4, no. 11, pp. 851-854, Nov. 2008.

[40] X. H. Lou, C. Adelmann, S. A. Crooker, E. S. Garlid, J. Zhang, K. S. M. Reddy, S. D. Flexner, C. J. Palmstrom, and P. A. Crowell, "Electrical detection of spin transport in lateral ferromagnet-semiconductor devices," *Nature Phys.*, vol. 3, no. 3, pp. 197-202, Mar. 2007.

[41] N. Tombros, C. Jozsa, M. Popinciuc, H. T. Jonkman, and B. J. van Wees, "Electronic spin transport and spin precession in single graphene layers at room temperature," *Nature*, vol. 448, no. 7153, pp. 571-574, Aug. 2007.

[42] W. Han, K. Pi, K. M. McCreary, Y. Li, J. J. I. Wong, A. G. Swartz, and R. K. Kawakami, "Tunneling spin injection into single layer graphene," *Phys. Rev. Lett.*, vol. 105, no. 16, Oct. 2010.

[43] 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 Proc. *IEEE/ACM int. Symp. Nano. Arch.*, Jun. 8-9 2011, pp. 129-136.

[44] J. A. Currivan, J. Youngman, M. D. Mascaro, M. A. Baldo, and C. A. Ross, "Low energy magnetic domain wall logic in short, narrow, ferromagnetic wires," *IEEE Mag. Lett.*, vol. 3, pp. 3000104-3000104, Apr. 2012.

[45] R. P. Cowburn, and M. E. Welland, "Room temperature magnetic quantum cellular automata," *Science*, vol. 287, no. 5457, pp. 1466-1468, Feb. 2000.

[46] G. H. Bernstein, A. Imre, V. Metlushko, A. Orlov, L. Zhou, L. Ji, G. Csaba, and W. Porod, "Magnetic QCA systems," *J. Microelectron.*, vol. 36, no. 7, pp. 619-624, Jul. 2005.

[47] A. Imre, G. Csaba, L. Ji, A. Orlov, G. H. Bernstein, and W. Porod, "Majority logic gate for magnetic quantum-dot cellular automata," *Science*, vol. 311, no. 5758, pp. 205-208, Jan. 2006.

[48] D. B. Carlton, N. C. Emley, E. Tuchfeld, and J. Bokor, "Simulation studies of nanomagnet-based logic architecture," *Nano Lett.*, vol. 8, no. 12, pp. 4173-4178, Dec. 2008.

[49] S. Datta, and B. Das, "Electronic analog of the electrooptic modulator," *Appl. Phys. Lett.*, vol. 56, no. 7, pp. 665-667, Feb. 1990.

[50] M. Johnson, "Bipolar spin switch," *Science,* vol. 260, no. 5106, pp. 320-323, Apr. 1993.

[51] T. Marukame, T. Inokuchi, M. Ishikawa, H. Sugiyama, and Y. Saito, "Read/write operation of spin-based MOSFET using highly spin-polarized ferromagnet/MgO tunnel barrier for reconfigurable logic devices," in Proc. *IEEE Int. Electron Device Meeting*, Dec. 7-9, 2009, pp. 1-4.

[52] S. Sugahara, and J. Nitta, "Spin-transistor electronics: an overview and outlook," *Proc. IEEE*, vol. 98, no. 12, pp. 2124-2154, Dec. 2010.

[53] J. G. Zhu, "Magnetoresistive random access memory: the path to competitiveness and scalability," *Proc. IEEE*, vol. 96, no. 11, pp. 1786-1798, Nov. 2008.

[54] J. Barth *et al.*, "A 45 nm SOI embedded DRAM macro for the power™ processor 32 MByte on-chip L3 cache", *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp.64 -75 2011

[55] K. C. Chun *et al.* “A scaling roadmap and performance evaluation of in-plane and perpendicular MTJ based STT-MRAMs for high-density cache memory,” *IEEE J. Solid-State Circuits*, vol. 48, no. 2, pp. 598-610, Feb. 2013.

[56] M. Durlam, P. Naji, A. Omair, M. DeHerrera, J. Calder, J. M. Slaughter, B. Engel, N. Rizzo, G. Grynkewich, B. Butcher, C. Tracy, K. Smith, K. Kyler, J. Ren, J. Molla, B. Feil, R. Williams, and S. Tehrani, "A low power 1 Mbit MRAM based on 1T1MTJ bit cell integrated with copper interconnects," in Proc. *IEEE Symp. Very Large Scale Integr. Circuits,* Jun. 13-15, 2002, pp. 158-161.

[57] B. N. Engel, J. Akerman, B. Butcher, R. W. Dave, M. DeHerrera, M. Durlam, G. Grynkewich, J. Janesky, S. V. Pietambaram, N. D. Rizzo, J. M. Slaughter, K. Smith, J. J. Sun, and S. Tehrani, "A 4-Mb toggle MRAM based on a novel bit and switching method," *IEEE Trans. Magn.*, vol. 41, no. 1, pp. 132-136, Jan. 2005.

[58] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano, "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM," in Proc. *IEEE Int. Electron Device Meeting,* Dec. 5-5, 2005, pp. 459-462.

[59] C. J. Lin, S. H. Kang, Y. J. Wang, K. Lee, X. Zhu, W. C. Chen, X. Li, W. N. Hsu, Y. C. Kao, M. T. Liu, W. C. Chen, L. YiChing, M. Nowak, N. Yu, and T. Luan, "45nm low power CMOS logic compatible embedded STT MRAM utilizing a

reverse-connection 1T/1MTJ cell," in Proc. *IEEE Int. Electron Device Meeting*, Dec. 7-9, 2009, pp. 1-4.

[60] T. Ishigaki, T. Kawahara, R. Takemura, K. Ono, K. Ito, H. Matsuoka, and H. Ohno, "A multi-level-cell spin-transfer torque memory with series-stacked magnetotunnel junctions," in Proc. *IEEE Symp. Very Large Scale Integr. Technol.*, Jun. 15-17, 2010, pp. 47-48.

[61] H. Meng, and W. Jian-Ping, "Spin transfer in nanomagnetic devices with perpendicular anisotropy," *Appl. Phys. Lett.*, vol. 88, no. 17, pp. 172506, Apr. 2006.

[62] T. Nozaki, Y. Shiota, M. Shiraishi, T. Shinjo, and Y. Suzuki, "Voltage-induced perpendicular magnetic anisotropy change in magnetic tunnel junctions," *Appl. Phys. Lett.*, vol. 96, no. 2, pp. 022506, Jan. 2010.

[63] W. G. Wang, M. G. Li, S. Hageman, and C. L. Chien, "Electric-field-assisted switching in magnetic tunnel junctions," *Nature Mater.*, vol. 11, no. 1, pp. 64-68, Jan. 2012.

[64] L. Q. Liu, C. F. Pai, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman, "Spin-torque switching with the giant spin Hall effect of tantalum," *Science*, vol. 336, no. 6081, pp. 555-558, May. 2012.

[65] S. Manipatruni, D. E. Nikonov, and I. A. Young, "Voltage and energy-delay performance of giant spin Hall effect switching for magnetic memory and logic," *Cornell Univ. Press*, pp. 1-16, Jan. 2013.

[66] K. Yusung, S. H. Choday, and K. Roy, "DSH-MRAM: differential spin Hall MRAM for on-chip memories," *IEEE Electron Device Lett.*, vol. 34, no. 10, pp. 1259-1261, Oct. 2013.

[67] S. S. Parkin, M. Hayashi, and L. Thomas, "Magnetic domain-wall racetrack memory," *Science*, vol. 320, no. 5873, pp. 190-194, Apr. 2008.

[68] S. Fukami, T. Suzuki, K. Nagahara, N. Ohshima, Y. Ozaki, S. Saito, R. Nebashi, N. Sakimura, H. Honjo, K. Mori, C. Igarashi, S. Miura, N. Ishiwata, and T. Sugibayashi, "Low-current perpendicular domain wall motion cell for scalable high-speed MRAM," in Proc. *IEEE Symp. Very Large Scale Integr. Technol.*, Jun. 16-18, 2009, pp. 230-231.

[69] R. Venkatesan, M. Sharad, K. Roy, and A. Raghunathan, "DWM-TAPESTRI: an energy efficient all-spin cache using domain wall shift based writes," in Proc. *Conference on Design, Automation and Test in Europe*, Mar. 18-22, 2013, pp. 1825-1830.

[70] S. I. Kiselev, J. C. Sankey, I. N. Krivorotov, N. C. Emley, R. J. Schoelkopf, R. A. Buhrman, and D. C. Ralph, "Microwave oscillations of a nanomagnet driven by a spin-polarized current," *Nature*, vol. 425, no. 6956, pp. 380-383, Sep. 2003.

[71] Z. Zeng, G. Finocchio, B. Zhang, P. K. Amiri, J. A. Katine, I. N. Krivorotov, Y. Huai, J. Langer, B. Azzerboni, K. L. Wang, and H. Jiang, "Ultralow-current-density and bias-field-free spin-transfer nano-oscillator," *Nature Sci. Rep.*, vol. 3, pp. 1-5, Mar. 2013.

[72] P. Villard, U. Ebels, D. Houssameddine, J. Katine, D. Mauri, B. Delaet, P. Vincent, M. C. Cyrille, B. Viala, J. P. Michel, J. Prouvee, and F. Badets, "A GHz spintronic-based RF oscillator," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 214-223, Jan. 2010.

[73] S. Yuasa, A. Fukushima, K. Yakushiji, T. Nozaki, M. Konoto, H. Maehara, H. Kubota, T. Taniguchi, H. Arai, H. Imamura, K. Ando, Y. Shiota, F. Bonell, Y. Suzuki, N. Shimomura, E. Kitagawa, J. Ito, S. Fujita, K. Abe, K. Nomura, H. Noguchi, and H. Yoda, "Future prospects of MRAM technologies," in Proc. *IEEE Int. Electron Devices Meeting*, Dec. 9-10, 2013, pp. 3.1.1-3.1.4.

[74] S. Srinivasan, A. Sarkar, B. Behin-Aein, and S. Datta, "All-spin logic device with inbuilt nonreciprocity," *IEEE Trans. Nanotech.*, vol. 47, no. 10, pp. 4026-4032, Oct. 2011.

[75] M. Sharad, C. Augustine, and K. Roy, "Boolean and non-Boolean computation with spin devices," in Proc. *IEEE Int. Electron Device Meeting*, Dec. 10-13, 2012, pp. 11.16.11-11.16.14.

[76] V. Calayir, D. E. Nikonov, S. Manipatruni, and I. A. Young, "Static and clocked spintronic circuit design and simulation with performance analysis relative to CMOS," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 61, no. 99, pp. 1-14, Feb. 2014.

[77] A. Hartstein, and T. R. Puzak, "Optimum power/performance pipeline depth," in Proc. *IEEE/ACM Int. Symp. Microarch.*, Dec. 3-5, 2003, pp. 117-125.

[78] M. Yuffe, E. Knoll, M. Mehalel, J. Shor, and T. Kurts, "A fully integrated multi-CPU, GPU and memory controller 32nm processor," in Proc. *IEEE Int. Solid-State Circuits Conf.*, Feb. 20-24, 2011, pp. 264-266.

[79] B. S. Landman, and R. L. Russo, "On a pin versus block relationship for partitions of logic graphs," *IEEE Trans. Comput.*, vol. C-20, no. 12, pp. 1469-1479, Dec. 1971.

[80] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI) - Part I: Derivation and validation," *IEEE Trans. Electron Devices*, vol. 45, no. 3, pp. 580-589, Mar. 1998.

[81] W. F. Brown, "Thermal fluctuations of a single-domain particle," *Phy. Rev.*, vol. 130, no. 5, pp. 1677-1686, Jun. 1963.

[82] R. Takemura, T. Kawahara, K. Miura, H. Yamamoto, J. Hayakawa, N. Matsuzaki, K. Ono, M. Yamanouchi, K. Ito, H. Takahashi, S. Ikeda, H. Hasegawa, H. Matsuoka, and H. Ohno, "A 32-Mb SPRAM with 2T1R memory cell, localized

bi-directional write driver and `1'/`0' dual-array equalized reference scheme," *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 869-879, Apr. 2010.

[83] A. Aharoni, "Demagnetizing factors for rectangular ferromagnetic prisms," *J. Appl. Phys.*, vol. 83, no. 6, pp. 3432-3434, Mar. 1998.

[84] M. Yoshikawa, E. Kitagawa, T. Nagase, T. Daibou, M. Nagamine, K. Nishiyama, T. Kishi, and H. Yoda, "Tunnel magnetoresistance over 100% in MgO-based magnetic tunnel junction films with perpendicular magnetic $L1_0$-FePt electrodes," *IEEE Trans. Magn.*, vol. 44, no. 11, pp. 2573-2576, Nov. 2008.

[85] G. Kim, Y. Sakuraba, M. Oogane, Y. Ando, and T. Miyazaki, "Tunneling magnetoresistance of magnetic tunnel junctions using perpendicular magnetization $L1_0$-CoPt electrodes," *Appl. Phys. Lett.*, vol. 92, no. 17, pp. 172502, Apr. 2008.

[86] J. G. Alzate, P. K. Amiri, P. Upadhyaya, S. S. Cherepov, J. Zhu, M. Lewis, R. Dorrance, J. A. Katine, J. Langer, K. Galatsis, D. Markovic, I. Krivorotov, and K. L. Wang, "Voltage-induced switching of nanoscale magnetic tunnel junctions," in Proc. *IEEE Int. Electron Device Meeting*, Dec. 10-13, 2012, pp. 29.25.21-29.25.24.

[87] T. Kimura, Y. Otani, and J. Hamrle, "Switching magnetization of a nanoscale ferromagnetic particle using nonlocal spin injection," *Phys. Rev. Lett.*, vol. 96, no. 3, Jan. 2006.

[88] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L.-C. Wang, and Y. Huai, "Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory," *J. Phys.: Condens. Matter*, vol. 19, no. 16, pp. 165209, Apr. 2007.

[89] T. Kimura, T. Sato, and Y. Otani, "Temperature evolution of spin relaxation in a NiFe/Cu lateral spin valve," *Phys. Rev. Lett.*, vol. 100, no. 6, Feb. 2008.

[90] I. N. Krivorotov, D. V. Berkov, N. L. Gorn, N. C. Emley, C. Sankey, D. C. Ralph, and R. A. Buhrman, "Large-amplitude coherent spin waves excited by spin-polarized current in nanoscale spin valves," *Phys. Rev. B*, vol. 76, no. 2, Jul. 2007.

[91] P. Zarkesh-Ha, and J. D. Meindl, "Stochastic net length distributions for global interconnects in a heterogeneous system-on-a-chip," in Proc. *IEEE Symp. Very Large Scale Integr. Technol.*, Jun. 9-11, 1998, pp. 44-45.

[92] Z. Chishti, and T. N. Vijaykumar, "Optimal power/performance pipeline depth for SMT in scaled technologies," *IEEE Trans. Comput.*, vol. 57, no. 1, pp. 69-81, Jan. 2008.

[93] E. Le Sueur, and G. Heiser, "Slow down or sleep, that is the question.," in Proc. *USENIX Annual Technical Conference*, Apr, 2011.

[94] J. W. Tschanz, S. G. Narendra, Y. B. Ye, B. A. Bloechel, S. Borkar, and V. De, "Dynamic sleep transistor and body bias for active leakage power control of microprocessors," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1838-1845, Nov. 2003.

[95] X. Liu, and W. Jian-Ping, "Fabrication and morphologies of large directly ordered $L1_0$ FePt nanoparticles in gas phase," *J. Appl. Phys.*, vol. 105, no. 7, pp. 07A722, Apr. 2009.

[96] L. Yongpan, R. P. Dick, S. Li, and Y. Huazhong, "Accurate temperature-dependent integrated circuit leakage power estimation is easy," in Proc. *Conference on Design, Automation and Test in Europe*, Apr. 16-20, 2007, pp. 1-6.

[97] K. C. Chun, H. Zhao, J. D. Harms, T. H. Kim, J. P. Wang, and C. H. Kim, “A scaling roadmap and performance evaluation of in-plane and perpendicular MTJ based STT-MRAMs for high-density cache memory,” *IEEE J. Solid-State Circuits*, vol. 48, no. 2, pp. 598-610, Feb. 2013.

[98] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano, “A Novel Nonvolatile Memory with Spin Torque Transfer Magnetization Switching: Spin-RAM,” *in IEDM Tech. Dig.*, pp. 459-462, Dec. 2005.

[99] J. P. Kim, T. Kim, W. Hao, H. M. Rao, K. Lee, X. Zhu, X. Li, W. Hsu, S. H. Kang, N. Matt, and N. Yu, “A 45nm 1Mb Embedded STT-MRAM with design techniques to minimize read-disturbance,” *in Proc. VLSI Circuits Symp.*, pp. 296-297, Jun. 2011.

[100] K. Lee and S. H. Kang, “Development of Embedded STT-MRAM for Mobile System-on-Chips,” *IEEE Trans. Magn.*, vol. 47, no. 1, pp. 131-136, Jan. 2011.

[101] H. Meng and J.P. Wang “Spin transfer in nanomagnetic devices with perpendicular anisotropy," *Appl. Phys. Lett*, vol. 88, no. 17, pp. 172506, Apr. 2006.

[102] J. Kim, H. Zhao, Y. Jiang, A. Klemm, J.-P. Wang, and C. H. Kim, ” Scaling Analysis of In-plane and Perpendicular Anisotropy Magnetic Tunnel Junctions Using a Physics-Based Model,” *in Proc. DRC*, Jun. 2014

[103] J. Kim, A. Chen, B. Behin-Aein, S. Kumar, J. P. Wang, and C. H. Kim, “A technology-agnostic MTJ SPICE model with user-defined dimensions for STT-MRAM scalability studies,” *in Proc. IEEE CICC*, Sep. 2015

[104] U. K. Klostermann, M. Angerbauer, U. Griming, F. Kreupl, M. Ruhrig, F. Dahmani, M. Kund, and G. Muller, “A perpendicular spin torque switching based MRAM for the 28 nm technology node,” *in IEDM Tech. Dig.*, pp. 187-190, Dec. 2007.

[105] D. Apalkov, S. Watts, A. D. Smith, E. Chen, Z. Diao, and V. Nikitin, “Comparison of scaling of in-plane and perpendicular spin transfer switching

technologies by micromagnetic simulation", *IEEE Trans. Magn.*, vol. 46, no. 6, pp. 2240-2243, Jun. 2010.

[106] J. D. Harms, F. Ebrahimi, X. F. Yao, and P. Wang, "SPICE macromodel of spin-torque-transfer-operated magnetic tunnel junctions," *IEEE Trans. Electron Devices*, vol. 57, no. 6, pp. 1425-1430, Jun. 2010

[107] Z. Xu, K. B. Sutaria, C. Yang, C. Chakrabarti, and Y. Cao, "Compact Modeling of STT-MTJ for SPICE Simulation," *in Proc. ESSDERC*, pp. 338-341, Sep. 2013.

[108] G. D. Panagopoulos, C. Augustine, and K. Roy, "Physics-based SPICE-compatible compact model for simulating hybrid MTJ/CMOS circuits," *IEEE Trans. Electron Devices*, vol. 60, no. 9, pp. 2808-2814, Sep. 2013

[109] W. H. Butler, T. Mewes, C. K. A. Mewes, P. B. Visscher, W. H. Rippard, S. E. Russek, and R. Heindl, "Switching distributions for perpendicular spin-torque devices within the macrospin approximation," *IEEE Trans. Magn.*, vol. 48, no. 12, pp. 4684-4700, Dec, 2012.

[110] L. Yuan, S.-H. Liou, and D. Wang, "Temperature dependence of magnetoresistance in magnetic tunnel junctions with different free layer structures," *Phys. Rev. B*, vol. 73, pp. 134403, Apr. 2006.

[111] M. Madec, J.-B. Kammerer, and L. Hebrard, "Compact modeling of a magnetic tunnel junction - Part II: tunneling current model," *IEEE Trans. Electron Devices*, vol. 57, no. 6, pp. 1416-1424, Jun. 2010.

[112] H. Zhao, A. Lyle, Y. Zhang, P. K. Amiri, G. Rowlands, Z. Zeng, J. Katine, H. Jiang, K. Galatsis, K. L. Wang, I. N. Krivorotov, and J.-P. Wang, "Low writing energy and sub nanosecond spin torque transfer switching of in-plane magnetic tunnel junction for spin torque transfer random access memory," *J. Appl. Phys.*, vol. 109, pp. 07C720, Mar. 2011.

[113] C. J. Lin, S. H. Kang, Y. J. Wang, K. Lee, X. Zhu, W. C. Chen, X. Li, W. N. Hsu, Y. C. Kao, M. T. Liu, W. C. Chen, Y. Lin, M. Nowak, N. Yu, and L. Tran, "45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell," *in IEDM Tech. Dig.*, pp. 11.6.1-11.6.4, Dec. 2009.

[114] G. D. Sandre, L. Bettini, A. Pirola, L. Marmonier, M. Pasotti, M. Borghi, P. Mattavelli, P. Zuliani, L. Scotti, G. Mastracchio, F. Bedeschi, R. Gastaldi, R. Bez, "A 90nm 4Mb embedded phase-change memory with 1.2V 12ns read access time and 1MB/s write throughput," *in IEEE ISSCC Dig. Tech. Papers*, pp. 268-269, Feb. 2010.

[115] D. Gogl, C. Arndt, J. C. Barwin, A. Bette, J. DeBrosse, E. Gow, H. Hoenigschmid, S. Lammers, M. Lamorey, Y. Lu, T. Maffitt, K. Maloney, W. Obermaier, A. Sturm, H. Viehmann, D. Willmott, M. Wood, W. J. Gallagher, G. Mueller, and A. R. Sitaram, "A 16-Mb MRAM featuring bootstrapped write drivers," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 902-908, Apr. 2005.

[116] S. Rusu, S. Tam, H. Muljono, J. Stinson, and D. Ayers et al., "A 45 nm 8-Core Enterprise Xeon® Processor", *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 7-14, Jan. 2010.

[117] R. J. Riedlinger, R. Bhatia, L. Biro, B. Bowhill, and E. Fetzer et al., "A 32nm 3.1 Billion Transistor 12-Wide-Issue Itanium® Processor for Mission-Critical Servers", *in IEEE ISSCC Dig. Tech. Papers*, pp. 84-85, Feb. 2011.

[118] D. Weller, A. Moser, L. Folks, M. E. Best, W. Lee, M. F. Toney, M. Schwickert, J. Thiele, and Mary F. Doerner, "High $K_u$ materials approach to 100 Gbits/in$^2$," *IEEE Trans. Magn.*, vol. 36, no. 1, pp. 10-15, Jan. 2000.

[119] S. Mizukami, S. Iihama, N. Inami, T. Hiratsuka, G. Kim, H. Naganuma, M. Oogane, and Y. Ando, "Fast magnetization precession observed in $L1_0$-FePt epitaxial thin film," *Apply. Phys. Lett.*, vol. 98, pp. 052501, Jan. 2011.

[120] X. Liu, W. Zhang, M. J. Carter, and G. Xiao, “Ferromagnetic resonance and damping properties of CoFeB thin films as free layers in MgO-based magnetic tunnel junctions,” *J. Appl. Phys.*, vol. 110, pp. 033910, Aug. 2011.

[121] P. He, X. Ma, J. W. Zhang, H. B. Zhao, G. Lupke, Z. Shi, and S. M. Zhou, “Scaling of intrinsic gilbert damping with spin-orbital coupling strength,” arXiv:1203.0607v1 [cond-mat.mtrl-sci], Mar. 2012.

[122] J.-H. Park, Y. Kim, W. C. Lim, J. H. Kim, S. H. Park, J. H. Kim, W. Kim, K. W. Kim, J. H. Jeong, K. S. Kim, H. Kim, Y. J. Lee, S. C. Oh, J. E. Lee, S. O. Park, S. Watts, D. Apalkov, V. Nikitin, M. Krounbi, S. Jeong, S. Choi, H. K. Kang, and C. Chung, “Enhancement of data retention and write current scaling for sub-20nm STT-MRAM by utilizing dual interfaces for perpendicular magnetic anisotropy,” *in IEDM Tech. Dig.*, pp. 57-58, Dec. 2012.

[123] T. Kawahara, R. Takemura, K. Miura, J. Hayakawa, S. Ikeda, Y. M. Lee, R. Sasaki, Y. Goto, K. Ito, T. Meguro, F. Matsukura, H. Takahashi, H. Matsuoka, and H. Ohno, “2 Mb SPRAM (SPin-Transfer Torque RAM) with bit-by-bit bi-directional current write and parallelizing-direction current read,” *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 109-120, Jan. 2008.

[124] W. Zhao and Y. Cao, “New generation of predictive technology model for sub-45 nm early design exploration,” *IEEE Trans. Electron Devices*, vol. 53, no. 11, pp. 2816-2823, Nov. 2006.

[125] J. Kim, B. Tuohy, C. Ma, W. H. Choi, I. Ahmed, D. Lilja, and C. H. Kim, ” Spin-Hall effect MRAM based cache memory: a feasibility study,” *in Proc. DRC,* Jun. 2015

[126] A. Shafaei, Y. Wang, and M. Pedram, “ Low write-energy STT-MRAMs using FinFET-based access transistors,” *IEEE ICCD,* pp. 374-379, Oct. 2014.

[127] F. Hamzaoglu, U. Arslan, N. Bisnik, S. Ghosh, M. B. Lal, N. Lindert, M. Meterelliyoz, R. B. Osborne, J. Park, S. Tomishima, Y. Wang, and K. Zhang, “ A 1Gb 2GHz embedded DRAM in 22nm tri-gate CMOS technology,” *in IEEE ISSCC Dig. Tech. Papers*, pp. 230-231, Feb. 2014.

[128] N. Muralimanohar, HP Lap. Tech. Rep. HPL-2009-85, 2009

[129] C. Mead, “Neuromorphic electronic systems,” *Proc. IEEE*, vol. 78, no. 10, pp. 1629-1636, Oct. 1990.

[130] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.

[131] The MNIST dataset is available at http://yann.lecun.com/exdb/mnist/index.html

[132] G. Indiveri, E. Chicca, and R. Douglas, “A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity,” *IEEE Trans. Neural Networks*, vol. 17, no. 1, Jan 2006.

[133] C. Bartolozzi, G. Indiveri, “Synaptic dynamics in analog VLSI,” *Neural Computation*, vol. 19, no. 10, pp. 2581-2603, Oct. 2007.

[134] S. Brink, S. Nease, and P. Hasler, “Computing with networks of spiking neurons on a biophysically motivated floating-gate based neuromorphic integrated circuit,” *Neural Networks*, vol. 45, pp. 39-49, Sep. 2013.

[135] D. Kuzum, R. G. D. Jeyasingh, and H.-S. P. Wong, “Energy efficient programming of nanoelectronic synaptic devices for large-scale implementation of associative and temporal sequence learning,” *in IEDM Tech. Dig.*, pp. 693-696, Dec. 2011.

[136] S. Park, A. Sheri, J. Kim, J. Noh, J. Jang, M. Jeon, B. Lee, B. R. Lee, B.H. Lee, and H. Hwang, “Neuromorphic speech systems using advanced ReRAM-based synapse,” *in IEDM Tech. Dig.*, pp. 625-628, Dec. 2013.

[137] J. Seo, et al., “A 45nm CMOS Neuromorphic Chip with a Scalable Architecture for Learning in Networks of Spiking Neurons,” *Proc. IEEE Custom Integrated Circuits Conference,* 2011.

[138] P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar, and D. S. Modha, “A digital neurosynaptic core using embedded crossbar memory with 45 pJ per spike in 45 nm,” *Proc. IEEE Custom Integr. Circuits Conf.*, 2011.

[139] P. Knag, J. Kim, T. Chen, Z. Zhang, “Sparse Coding Neural Network ASIC With On-Chip Learning for Feature Extraction and Encoding,” *IEEE Journal of Solid-State Circuits*, vol. 50, no. 4, April 2015.

[140] S. Song, K. Chun, and C. H. Kim, “A Logic-Compatible Embedded Flash Memory Featuring a Multi-Story High Voltage Switch and a Selective Refresh Scheme,” in *IEEE Symp. on VLSI Circuits Dig.*, 2012, pp. 130-131.

[141] S. Song, K. Chun, and C. H. Kim, “A logic-compatible embedded flash memory for zero-standby power system-on-chips featuring a multi-story high voltage switch and a selective refresh scheme,” *IEEE J. Solid-State Circuits*, vol. 48, no. 5, pp. 1302-1314, May 2013.

[142] S. Song, J. Kim, and C. H. Kim, “Program/erase speed, endurance, retention, and disturbance characteristics of single-poly embedded flash cells,” in *Proc. IEEE Int. Reliability Phys. Symp.(IRPS)*, 2013, pp. MY.4.1-MY.4.6.