January 2015

BOOLEAN AND BRAIN-INSPIRED COMPUTING USING SPIN-TRANSFER TORQUE DEVICES

Deliang Fan
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By DELIANG FAN

Entitled

BOOLEAN AND BRAIN-INSPIRED COMPUTING USING SPIN-TRANSFER TORQUE DEVICES

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

KAUSHIK ROY
Chair

ANAND RAGHUNATHAN

BYUNGHOO JUNG

VIJAY RAGHUNATHAN

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): KAUSHIK ROY

Approved by: MICHAEL R. MELLOCH 7/17/2015

Head of the Departmental Graduate Program Date

BOOLEAN AND BRAIN-INSPIRED COMPUTING USING SPIN-TRANSFER TORQUE DEVICES

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Deliang Fan

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2015

Purdue University

West Lafayette, Indiana

Dedicated to my family and friends for their endless love and selfless support

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my Ph.D. advisor, Prof. Kaushik Roy. His exceptional guidance, constant support and encouragement are pivotal in making this dissertation possible. I would never forget the words he told me-"You will never know if you don't try", which inspired me during my hard times. I sincerely thank him, not only for teaching me to be a qualified research scientist, but also for guiding me to be a man who can face the challenges with courage and confidence.

I would also like to thank my advisory committee members, Prof. Anand Raghunathan, Prof. Byunghoo Jung, and Prof. Vijay Raghunathan, for their constructive advice and feedback to improve the quality of my doctoral research and dissertation.

I am thankful to the former Nanoelectronic Research Lab (NRL) members for their valuable discussion and inspiration: Dr. Mrigank Sharad, Dr. Xuanyao Fong, Dr. Chao Lu, Dr. Harsha Choday, Dr. Chih-Hsiang Ho, Dr. Charles Augustine and Dr. Sumeet Gupta. Special thanks to Dr. Mrigank Sharad for being an exceptional mentor, collaborators and good friend since the time I started my doctoral program.

I sincerely thank my other colleagues and collaborators from NRL and Embedded System Lab (ESL) for their technical discussion, contribution and support: Karthik Yogendra, Supriyo Maji, Yong Shim, Abhronil Sengupta, Yusung Kim, Minsuk Koo, Yeongkyo Seo and all other NRL & ESL members.

Last but not the least, I would like to express my gratitude to my parents - Tianlu Fan & Tuansi Liu, my relatives and my girlfriend - Zhao Cui for their endless love and selfless support.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

ABSTRACT

Fan, Deliang. Ph.D., Purdue University, August 2015. Boolean and Brain-Inspired Computing Using Spin-Transfer Torque Devices. Major Professor: Kaushik Roy.

Several completely new approaches (such as spintronic, carbon nanotube, graphene, TFETs, etc.) to information processing and data storage technologies are emerging to address the time frame beyond current Complementary Metal-Oxide-Semiconductor (CMOS) roadmap. The high speed magnetization switching of a nano-magnet due to current induced spin-transfer torque (STT) have been demonstrated in recent experiments. Such STT devices can be explored in compact, low power memory and logic design. In order to truly leverage STT devices based computing, researchers require a re-think of circuit, architecture, and computing model, since the STT devices are unlikely to be drop-in replacements for CMOS. The potential of STT devices based computing will be best realized by considering new computing models that are inherently suited to the characteristics of STT devices, and new applications that are enabled by their unique capabilities, thereby attaining performance that CMOS cannot achieve. The goal of this research is to conduct synergistic exploration in architecture, circuit and device levels for Boolean and brain-inspired computing using nanoscale STT devices. Specifically, we first show that the non-volatile STT devices can be used in designing configurable Boolean logic blocks. We propose a spin-memristor threshold logic (SMTL) gate design, where memristive cross-bar array is used to perform current mode summation of binary inputs and the low power current mode spintronic threshold device carries out the energy efficient threshold operation. Next, for brain-inspired computing, we have exploited different spin-transfer torque device structures that can implement the hard-limiting and soft-limiting artificial neuron transfer functions respectively. We apply

such STT based neuron (or ‘spin-neuron’) in various neural network architectures, such as hierarchical temporal memory and feed-forward neural network, for performing “human-like” cognitive computing, which show more than two orders of lower energy consumption compared to state of the art CMOS implementation. Finally, we show the dynamics of injection locked Spin Hall Effect Spin-Torque Oscillator (SHE-STO) cluster can be exploited as a robust multi-dimensional distance metric for associative computing, image/ video analysis, etc. Our simulation results show that the proposed system architecture with injection locked SHE-STOs and the associated CMOS interface circuits can be suitable for robust and energy efficient associative computing and pattern matching.

1. INTRODUCTION

1.1. Introduction

The scaling of Complementary Metal-Oxide Semiconductor (CMOS) transistors brings a lot of issues, such as short channel effect, large leakage current and so on. Considerable research efforts has started in earnest to explore new devices that can potentially replace CMOS. Several completely new approaches (such as spintronic [1]-[7], TFETs [8][9], etc.) to information processing and data storage technologies are emerging to address the time frame beyond current CMOS roadmap. These emerging devices have unique characteristics that set them apart from traditional MOS transistors. In order to attain performance that CMOS cannot achieve, new computing models that are uniquely suited to the characteristics of these emerging devices are required to be explored.

Recently, it was experimentally demonstrated that the spin polarized currents can switch nano-scale magnets due to spin-transfer torque (STT) [4][5]. Compared with CMOS transistors, STT devices have the characteristics of non-volatility, zero current leakage and high integration density, which make them promising candidates for designing compact, low power memory and Boolean logic [13]-[24]. It is well accepted that STT devices are suitable in on-chip memory design, while the suitability of spin-transfer torque devices for logic applications is debatable [24]. In this dissertation, we focus on a wider perspective on the application of STT devices involving exploring combination of spin and charge devices and searching for computation models enabled by their unique capabilities.

1.2. Spin Devices

Every atom is composed of a nucleus and one or more electrons, where electrons are orbiting around the nucleus. Thus electron has an orbital angular momentum. However,

the experimental evidence suggests that an electron has an intrinsic angular momentum, which comes from the spin of the electron. Electrons with unidirectional electron spin moment results in magnet with non-zero moment, or in other words, electron is a magnet. Some atoms, such as Fe^{2+}, Co^{3+}, Mn^{2+}, have oxidation states with incomplete electronic sub-shells, occurring in the 3d shells of the transition elements. These elements can produce magnetic moments. The electron spin can be manipulated using external magnetic field or spin-transfer torque effect [1][2]. In the following subsections, we will discuss the magnet switching energy and the above two magnet switching mechanisms.

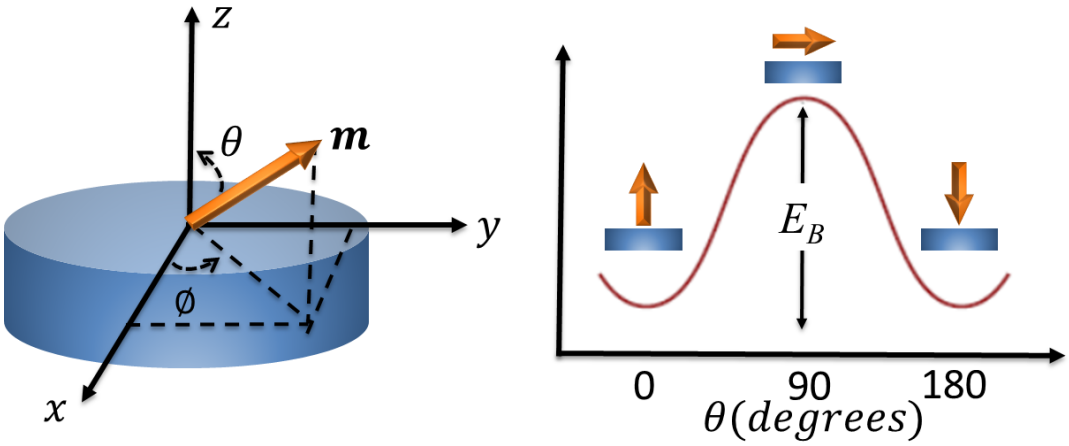


Fig. 1.1 Nano-magnet with uniaxial anisotropy and corresponding energy landscape

1.2.1. Magnet Switching Energy

In a nano-magnet, up-spin (0^o) and down-spin (180^o), as shown in Fig. 1.1, are used to denote two stable states. The anisotropy barrier is introduced to stabilize the magnetic moment along one direction as illustrated in Fig. 1.1. The information can be encoded as nano-magnet magnetization directions. Note that, nano-magnets can be used as non-volatile bi-stable elements due to the fact that the anisotropy barrier exists without the need for an external power supply. The information retention time (T_{rt}) of a nano-magnet is expressed as follow:

$$T_{rt} = T_0 \exp(\frac{K_u V}{K_B T}) \tag{1.1}$$

where, T_0 denotes the characteristic time, K_u is the magnetic anisotropy, V represents the nano-magnet volume, K_B is the Boltzmann's constant and T is the temperature in Kelvin [27]. Typically, around ten years of retention time can be achieved when the energy barrier ($E_B=K_uV$) is around $40K_BT$.

1.2.2. Magnetic Field Switching

One way of manipulating the magnetization direction of a nano-magnet is using an external magnetic field generated by a current-carrying wire. The minimum magnetic field required to switch the magnet is called *critical magnetic field* (H_c), which can be expressed as:

$$H_c = 2\frac{K_u}{M_s} \tag{1.2}$$

where, *Ms* denotes saturation magnetization. For example, if we want to switch the magnet from up-spin (0^o) to down-spin (180^o), in general, there are two scenarios to switch the magnet using external magnetic field. In the first scenario, a critical magnetic field (H_c) is first applied perpendicular (90^o) to the easy-axis, namely along the hard-axis. Then a small bias field (H_{bias}), which can be $\sim 10\% H_c$, is applied along the easy-axis (180^o). When the 90^o H_c magnetic field is removed, the magnet can be switched from up-spin (0^o) to down-spin (180^o).

The layout of two orthogonal wires generating two orthogonal fields are shown in Fig. 1.2a. The relationship between the current and the generated magnetic field can be described by Biot-Savart law:

$$B = I\frac{\mu_0}{4\pi}\int \frac{dl \sin\theta}{r^2} \tag{1.3}$$

where, as shown in Fig. 1.2b, I is the current flowing through the wire, l is the distance from the point-current element to the closest point of the wire to the nano-magnet, r is the distance from the point-current element to the nano-magnet.

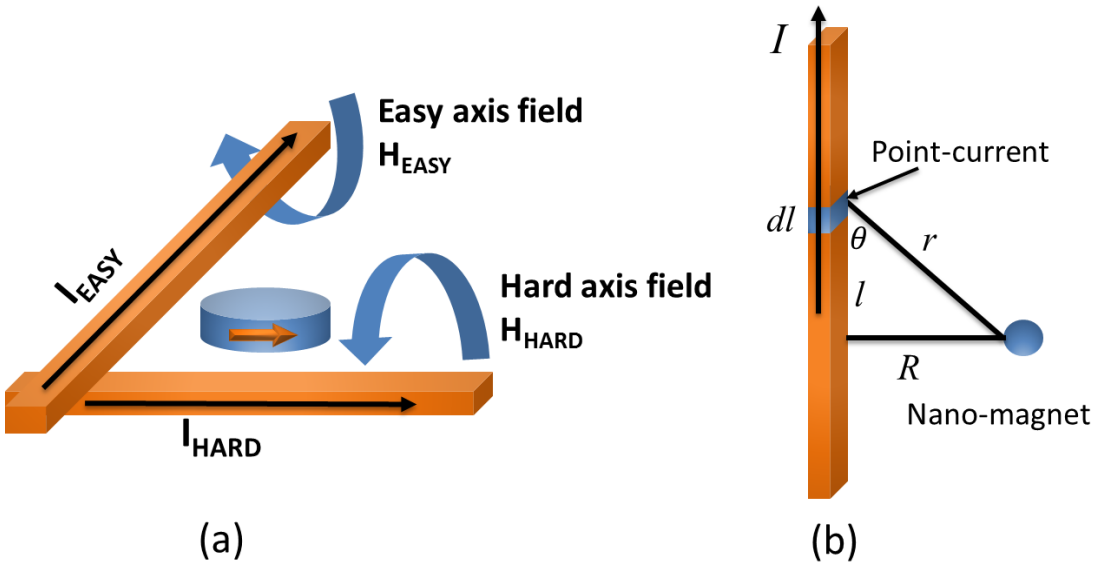


Fig. 1.2 (a) Two orthogonal wires generate two orthogonal fields H_{HARD} and H_{EASY} (b) magnetic field generation using current carrying wire

In the second scenario of magnetic switching, the critical magnetic field is directly applied along the easy-axis (180º). Compared to the first scenario, it only requires one magnetic field. However, the nano-magnet switching time is slower than that of first scenario. In both scenarios, the magnetic field is not localized and is energy inefficient. In addition, the magnetic field switching method is also not scalable for applications that require high density of on-chip nano-magnets.

1.2.3. Spin-Transfer Torque Switching

A more efficient way to switch a nano-magnet involves exploiting the current induced spin-transfer torque effect as we will describe next.

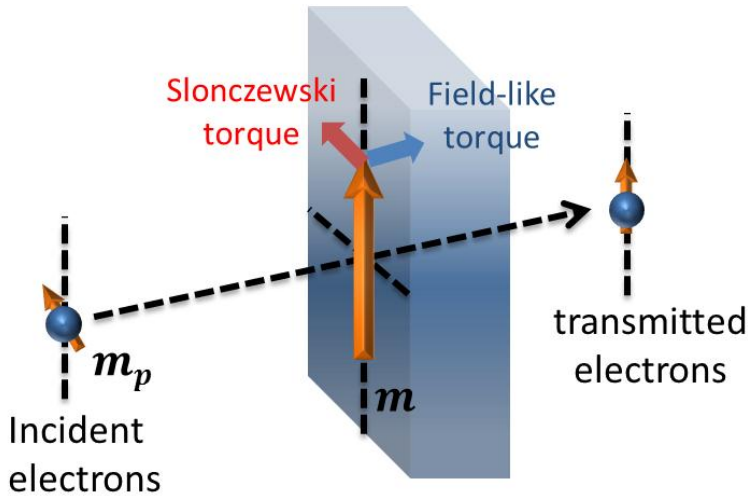


Fig. 1.3 Slonczewski torque and field-like torque on the nano-magnet due to the flowing of spin-polarized current

The behavior of the magnetization of the nano-magnet can be modeled using the Landau-Lifshitz-Gilbert equation with other terms describing the interaction between spin current and nano-magnets [1][2][23]:

$$\frac{d\boldsymbol{m}}{dt} = -|\gamma|\boldsymbol{m}\times\boldsymbol{H}_{eff} + \alpha\boldsymbol{m}\times\frac{d\boldsymbol{m}}{dt} + \boldsymbol{\tau} \tag{1.4}$$

where $\boldsymbol{m}$ is a unit vector pointing to the magnetization direction of nano-magnet, γ is the gyromagnetic ratio, $\boldsymbol{H}_{eff}$ denotes the effective magnetic field, and α is the Gilbert damping factor. $\boldsymbol{\tau}$ represents the current induced torques that we will describe in details in this subsection.

$$\boldsymbol{H}_{eff} = \boldsymbol{H}_{ani} + \boldsymbol{H}_{exch} + \boldsymbol{H}_{ext} + \boldsymbol{H}_{M} + \boldsymbol{H}_{noise} \tag{1.5}$$

The first term in equation-1.4 describes the magnetization *precession* resulting from effective magnetic field, which may include anisotropy field ($\boldsymbol{H}_{ani}$), exchange magnetic field ($\boldsymbol{H}_{exch}$), external magnetic field ($\boldsymbol{H}_{ext}$), magneto-static field ($\boldsymbol{H}_{M}$) and thermal noise term ($\boldsymbol{H}_{noise}$), as shown in equation-1.5 [115]. Anisotropy field comes from the anisotropy effect observed in ferromagnetic bodies resulting from the lattice structure and the particular symmetries in certain crystals. The *easy directions* in this dissertation are certain energetically favorable directions in a given magnetic materials without external

magnetic field. The exchange field arises from the exchange phenomenon (i.e. ferromagnetism and anti-ferromagnetism) observed in a large magnet composed of many smaller ferromagnetic particles. Unlike the exchange fields coming from the nearest neighbor coupling between magnetic particles, the magneto-static field (i.e. demagnetizing field) represents the long range coupling. It comes from the fact that the magnetic particle in a ferromagnetic body can be affected by the magnetic fields generated from the rest of the magnetic particles. Thermal effects are modeled using a stochastic Gaussian magnetic field, H_{noise}= ($H_{noise\text{-}x}$, $H_{noise\text{-}y}$, $H_{noise\text{-}z}$,). The mean of this Gaussian distribution is zero, while the standard deviation is $\sqrt{2\alpha K_B T/\gamma M_s V \Delta t}$ [127], where K_B is Boltzmann's constant, T is temperature in Kelvin, M_s is the saturation magnetization, V is the volume of free layer and Δt is the time step used in solving LLG equation.

The second term, called *Gilbert damping* term, describes the nonlinear spin relaxation phenomenon due to spin-orbit coupling [147]. It represents the damping rate at which $\boldsymbol{m}$ reaches equilibrium.

In general, the last term $\boldsymbol{\tau}$ represents current induced torques that take Slonczewski (i.e. spin-transfer torque) term and field-like term as shown in Fig. 1.3. Spin-transfer torque effect was theoretically predicted by Slonczewski1 [1] and Berger [2]. It comes from the interaction between spin current and a nano-magnet. Since a nano-magnet has unequal up-spin and down-spin density of states, the currents flowing through a nano-magnet is spin-polarized. Thus, a nano-magnet can act as a spin-polarizer whose direction is determined by the magnetization. The non-collinear spin-polarized electrons experience an exchange field trying to align the electron spins in the same direction of the nano-magnet, when they flow through the nano-magnet. This exchange field is the same field that aligns all the spins in the nano-magnet. Correspondingly, due to angular momentum conservation, the nano-magnet also experience a torque of equal magnitude but opposite sign at the same time. This torque is called *spin-transfer torque* (STT), which can be employed to switch the magnetization. It can be expressed as follows:

$$\tau_{STT} = -\frac{\gamma\hbar}{2eM_sV}\boldsymbol{m}\times(\boldsymbol{m}\times\boldsymbol{I}_s) \tag{1.6}$$

where, $\hbar$ is the Plank's constant, e is the electron charge, M_s denotes the saturation magnetization of the magnet, and V represents the volume of the magnet. The spin current ($\boldsymbol{I_s}$) is determined by the device geometry and materials combination, which will be described in the next chapter.

Generally, a field-like torque is also generated in asymmetric layered systems shown as follows:

$$\tau_{fl} = -\frac{\gamma\hbar}{2eM_sV}\beta \boldsymbol{m} \times \boldsymbol{I}_s \tag{1.7}$$

where, β is the ratio of this field-like torque strength to the Slonczewski torque. The magnitude of these two torques is dependent on the material and the device structures. Note that, for giant-magneto resistance (GMR) devices, the field like term is typically negligible as transverse spins dephase rapidly [184]. While for tunneling magneto resistance (TMR) devices, besides the in-plan torque predicted by Slonczewski, this field-like (out of plane) torque is proven significant in modeling the dynamics of magnet [187].

Following the recent discoveries of various physical phenomena involved in the current induced switching of nano-magnets, there have been various devices based on spin-transfer torque for memory and logic applications. In next chapter, we will discuss various spin-transfer torque devices that are employed in our research.

1.3. Organization

This dissertation conducts synergistic exploration in architecture, circuit and device levels for Boolean and brain-inspired computing using spin-transfer torque devices. Compared with state of the art CMOS designs, the spin based Boolean threshold logic design and brain-inspired computing can achieve ultra-low energy consumption. The remaining part of this dissertation is organized as follows.

Chapter 2 reviews several spin-transfer torque devices, including vertical spin value, lateral spin valve, magnetic domain wall strip and spin-orbit torque devices. The associated underlying physical phenomena in these STT devices are also described in this chapter.

Chapter 3 explores the spin based Boolean computation in threshold logic design. Memristive cross-bar array is employed to perform current mode summation of binary

inputs in the proposed spin-memristor threshold logic gate design. The low power, current mode spintronic threshold device is used to carry out the energy efficient threshold operation. Compared with state of the art CMOS threshold logic design, the proposed spin-memristor threshold logic achieves around two orders of magnitude lower energy consumption.

In chapter 4, we propose an energy efficient hardware mapping of a novel brain-inspired computing scheme - Hierarchical temporal memory (HTM) that tries to mimic the computing in cerebral neocortex. In HTM design, ultra-low power, magneto metallic hard-limiting spin-neurons combined with memristive cross-bar array (MCA) are explored in the dot product based pattern matching, which is the core computing block in HTM hardware. Such a direct mapping of the core-computing primitive of the cortical computing system can be very attractive for large-scale and energy efficient design.

In chapter 5, we present a spin-transfer torque (STT) device based on Domain Wall Motion (DWM) magnetic strip that can efficiently implement a Soft-limiting Non-linear Neuron (SNN) operating at ultra-low supply voltage and current. In contrast to previous spin-based neurons that can only realize hard-limiting (i.e. step function) transfer functions, the proposed STT-SNN displays a continuous resistance change with varying input current, and can therefore be employed to implement a soft-limiting neuron transfer function. We also present an artificial neural network (ANN) hardware design employing the proposed STT-SNNs and MCA as synapses. The ultra-low voltage operation of the magneto metallic STT-SNN enables the programmable MCA-synapses, computing analog domain weighted summation of input voltages, to also operate at ultra-low voltage. We modeled the STT-SNN using micro-magnetic simulation and evaluated them using a feed-forward ANN for character recognition. Comparisons with analog and digital CMOS neurons show that STT-SNNs can achieve more than two orders of magnitude lower energy consumption.

Chapter 6 shows that the dynamics of injection locked Spin Hall Effect Spin-Torque Oscillator (SHE-STO) cluster can be exploited as a robust primitive computational operator for associative computing. A cluster of SHE-STOs can be locked to a common frequency and phase with an injected AC current signal. DC inputs to each STO from

external stimuli can conditionally unlock some of them. Based on the input DC signal, the degree of synchronization of the SHE-STO cluster is detected by CMOS interface circuitry. The degree of synchronization can be used for associative computing/matching. We present a numerical simulation model of SHE-STO devices based on Landau-Lifshitz-Gilbert (LLG) equation with spin-transfer torque (STT) term and Spin Hall Effect (SHE). The model is then used to analyze the frequency and phase locking properties of injection locked SHE-STO cluster. Results show that associative computing based on the injection locked SHE-STO cluster can be energy efficient and relatively immune to device parameter variations and thermal noise.

Finally, the concluding remarks are available in chapter 7. Spin-transfer torque devices are unlikely to be drop-in replacements for CMOS. They may be integrated with CMOS and other charge based devices to model energy efficient computing systems. The proposed new computing models in Boolean and brain-inspired computing are inherently suited to the characteristics of STT devices, thereby attaining performance that CMOS cannot achieve.

2. SPIN-TRANSFER TORQUE DEVICES

In this chapter, we review several spin-transfer torque devices, including vertical spin value, lateral spin valve, magnetic domain wall strip and spin-orbit torque devices. The associated underlying physical phenomena in these STT devices are also described in this chapter. In the latter chapters of this dissertation, the fundamental STT devices described in this chapter will be employed as the building blocks in Boolean and brain-inspired computing.

2.1. Vertical Spin Valve

The device structure of vertical spin valve is shown in Fig. 2.1. It consists of a fixed ferromagnetic layer (reference layer), a free ferromagnetic layer (free layer) and a spacer in between. Historically, this device structure is used as a sensor by exploiting the resistance dependence on the magnetic orientation in the vertical spin valve. In 1975, Julliere [6] discovered *Tunneling Magneto-Resistance (TMR)* effect when the spacer between two ferromagnetic layers is insulator. In such TMR vertical spin valve device, the resistance is higher when the magnetization of two ferromagnetic layers are anti-parallel compared to the resistance of parallel magnetization configuration. The *magneto-resistance* (MR) ratio defined as $\Delta G/G_{AP}$ in percentage is used to characterize vertical spin value, where ΔG is the difference of the conductance between parallel (P) configuration and anti-parallel (AP) configuration and G_{AP} is conductance of AP configuration. In Julliere's work [6], MR is ~14% in a Fe/GeO/Co vertical spin valve at $T = \sim 4.2K$ (T is the temperature). In 1988, Fert and Grunberg [7][29] discovered the similar resistance dependence on the magnetic orientation in a vertical spin valve with a metallic spacer, which is called *Giant Magneto-Resistance* (GMR). In Fert's work [7], a vertical spin valve with Fe/Cr super lattices structure can achieve MR ratio ~80% at T

$= \sim 4.2K$. After these pioneering works, more works on GMR and TMR effects has been developed [10]-[12], [30]-[34].

When the spin-polarized electrons travel through the vertical spin valve with metallic spacer, the spin scattering effect causes the GMR effect. Specifically, electrons experience little scattering and can pass through the vertical spin valve easily when the device is in parallel magnetization orientations. While for anti-parallel configuration, electrons experience more spin scattering when passing through the vertical spin valve. It makes electrons difficult to pass through the device. Thus, the conductance of parallel configuration is higher than that of anti-parallel configuration.

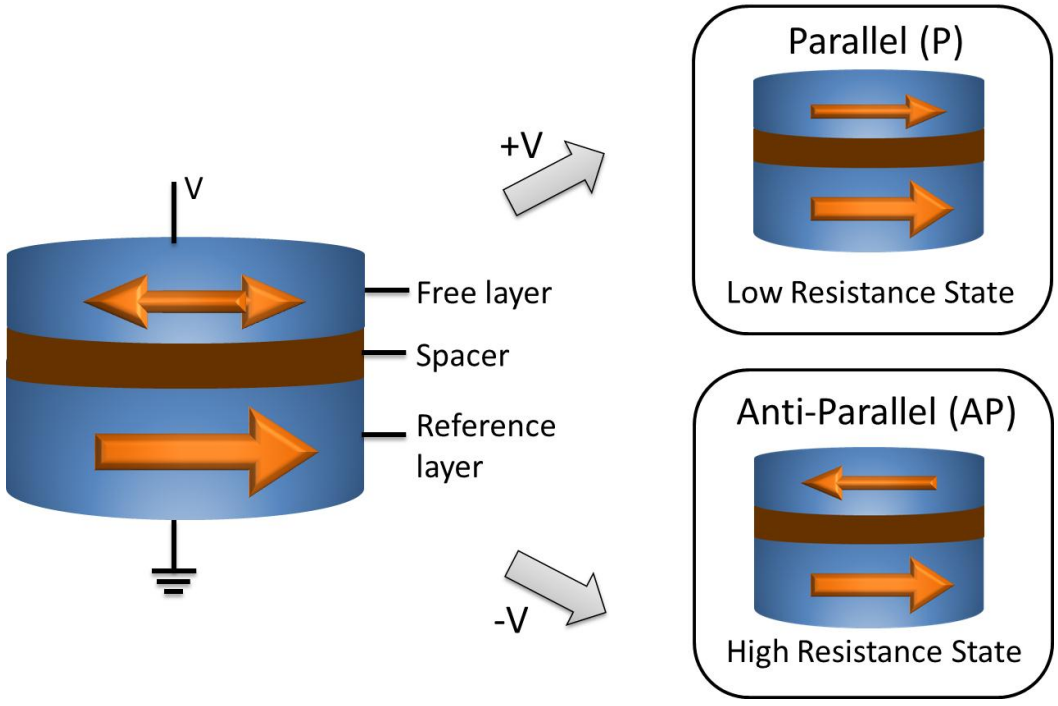


Fig. 2.1 Physical structure of a vertical spin value and its two states, corresponding to low and high resistance state

A Magnetic Tunnel Junction (MTJ), as shown in Fig. 2.2, consists of two nano-magnets separated by an insulator. The band structures of parallel MTJ and anti-parallel MTJ are shown in Fig. 2.2a and Fig. 2.2b, respectively. In MTJ, the ferromagnetic layer

acts as polarizer of electron spin. The TMR effect in an MTJ can be explained by the *spin filtering effect*, where the tunneling probability of electrons across the tunnel barrier depends on the relative magnetization of the reference and free ferromagnetic layers [6], [10]-[12]. As shown in Fig. 2.2, the electrons can only tunnel into the sub-band of the same spin orientation in the absence of spin-flip processes. For example, in the MTJ parallel configuration shown in Fig. 2.2a, the sub-bands of two ferromagnetic layers (FM) are well matched, namely the number of filled and empty electronic states for each spin are well matched. On the other hand, the sub-bands of anti-parallel MTJ is not matched. Thus, larger number of electrons can tunnel through the parallel MTJ than anti-parallel MTJ, leading to a larger tunneling conductance of parallel MTJ than anti-parallel MTJ.

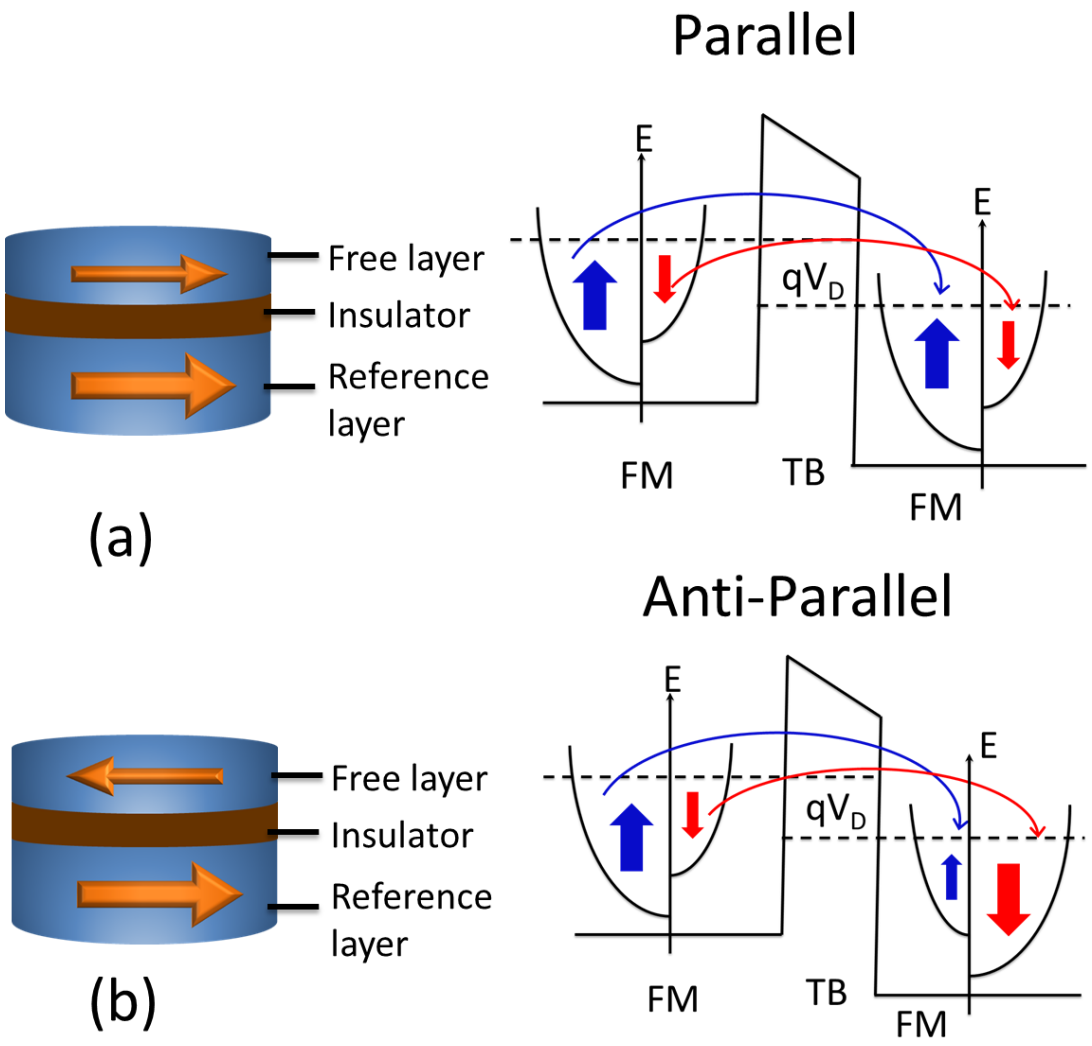


Fig. 2.2 Physical structure of MTJ and its band structure of parallel and anti-parallel configurations

The MTJ tunneling conductance can be expressed as:

$$G(\theta) = \frac{1}{2}(G_P + G_{AP}) + \frac{1}{2}(G_P - G_{AP})\cos\theta \tag{2.1}$$

where θ is the relative angle of two ferromagnetic layers, G_{AP} and G_P are the anti-parallel (θ=180°) and parallel (θ=0°) MTJ conductance, respectively. Then the TMR ratio is defined as:

$$TMR\ Ratio\ =\ \frac{G_P - G_{AP}}{G_{AP}} \tag{2.2}$$

Typically, the spin filtering effect enhances the MR ratios of MTJ, which makes it much higher than those of GMR based vertical spin valves. Furthermore, the MTJ resistance difference between P and AP configurations are much higher than that of GMR based vertical spin valve due to the employment of insulator as spacer.

So far, we have discussed that the vertical spin valve can be easily used as a sensor to detect the magnetic state of a nano-magnet by exploiting the GMR or TMR effects. However, for memory and logic applications, manipulation of the magnetic state is also necessary. Next, we will discuss how to manipulate the magnetization of the free layer in the vertical spin valve using current induced spin-transfer torque as we described in the previous subsection.

Since the electron spins get polarized flowing through the FM layer, they exert spin-transfer torque on the FM layer magnetization. For the reference layer, the magnetization is strongly pinned so that STT is negligible. While for the free layer, the STT can switch the magnetization direction or drive the magnetization into a sustained oscillation based on the orientation of the magnetization and the spin current polarization. Thus, the spin current generated by the vertical spin valve can be expressed as:

$$\boldsymbol{I}_s = \eta \mathrm{I}\, \boldsymbol{m}_p \tag{2.3}$$

Where $\boldsymbol{I}_s$ is the spin-polarized current, I is the charge current, $\boldsymbol{m}_p$ is the FM layer magnetization direction and η indicates the ratio of charge current magnitude to spin-polarized current magnitude. The magnitude of η may depend on the voltage across vertical spin valve, $\boldsymbol{m}$ and $\boldsymbol{m}_p$ [185]. Note that, MTJ is more efficient at generating spin-polarized current than GMR based vertical spin valve due to the spin filtering effect.

2.2. Lateral Spin Valve

Fig. 2.3 shows the physical structure of lateral spin valve (LSV) with local and non-local measurements [35]-[39]. LSV consists of two ferromagnetic contacts (FM) deposited on a non-magnetic (NM) channel. As shown in Fig. 2.3, there are two ways to measure the magneto-resistance effect in LSV, namely local and non-local measurements. For local measurement, it is similar to a vertical spin valve with a structure of

FM/NM/FM. Thus the magneto-resistance effect is also observed [35]. For non-local measurement of lateral spin valve, authors in [36][37] discovered that voltage on the FM detector contact can be modulated by current injection through the FM injector contact. It depends on the current injection and the distance between the injector and the detector.

In the non-local measurement, though a current is injected through the FM injector contact, this current does not flow through the FM detector contact or the NM channel underlying the detector contact. The electron transport spin drift diffusion model can be used to explain LSV non-local effect [41]. Firstly, the FM injector spin-polarizes the injected electrons. As a result, the number of spins with same magnetization direction as FM injector is larger than that of opposite spins in the underlying non-magnetic channel. This imbalance of spins leads to non-equilibrium spin accumulation, thus a spin voltage in the NM. Note that, the spin voltage is the defined as the *electrochemical potential* (ECP) difference between the *up-spin potential* (μ_{up}) and *down-spin potential* (μ_{dn}). Due to the spin voltage across the non-magnetic channel, one type of spins flow in one direction (I_{up}), while the other type of spins flow in the opposite direction (I_{dn}). As shown in Fig. 2.3, the charge current is defined as $I_Q=I_{up}+I_{dn}$ and the spin current is defined as $I_s=I_{up}-I_{dn}$. Since I_{up} and I_{dn} have the same magnitude, but opposite directions, the charge current in the non-magnetic channel is zero and the spin current is non-zero.

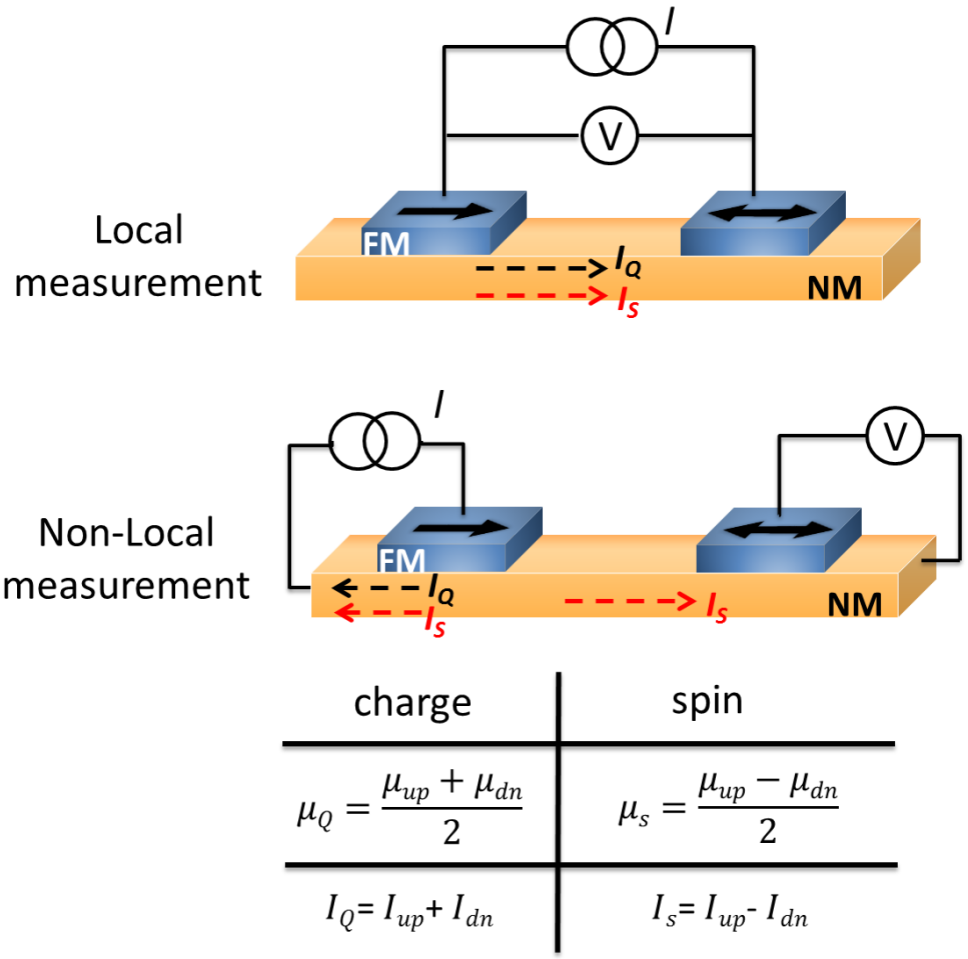


Fig. 2.3 Physical structure of a lateral spin valve of *local* and *non – local* measurements that has been made to experimentally observe magneto-resistance effect and pure spin-current generation.

Typically, the spin current generation efficiency in lateral spin valve is mainly limited by two factors: spin injection efficiency at the FM injector contact and the spin flip length in the NM channel. The spin injection efficiency at FM/NM interface can be improved by inserting a tunnel barrier between FM and NM, as shown experimentally in [37]. In the non-magnetic channel, the spin current decays exponentially because of the spin flip processes, leading to reduced magnitude of spin current to the FM detector contact. Several research works [35]-[38], [42][43] have investigated different NM channel materials with different spin flip lengths (λ_{sf}) for implementing energy efficient

lateral spin valves. The pure spin current generation efficiency can be improved by exploring LSV material combinations. In this non-local LSV experiment [44], the non-local spin current was proven large enough to switch the magnetization of the FM detector. Therefore, by measuring the magnetization of FM detector contact, we can detect the non-local spin current in the LSV. Based on non-local LSV, "all-spin" based computation was proposed in [21].

2.3. Magnetic Domain Wall Strip

So far, the spin current we have discussed are generated by injecting current through a spin valve and the exerted STT is used to manipulate another nano-magnet. Besides spin valve structure, it has been experimentally shown that the spin current can also be generated in a magnetic domain wall strip. Fig. 2.4 shows a ferromagnetic wire, called magnetic domain wall stripe (DWS), with a nanowire-like geometry and opposite magnetization at its two ends. The magnetization transition region along the DWS from one direction to the opposite direction is called *domain wall* (DW), whose structure and size are dependent on the DWS geometry and material properties.

Fig. 2.4 shows several typical DW structures in a DWS [45][46], [189]-[193]. When the shape anisotropy dominates in materials such as Permalloy, NiFe or Py, the magnetic domains lie along the wire axis (in-plan magnetic anisotropy, IMA). The domain wall in such materials can be either transverse or vortex type. In a thin and narrow magnetic nano-strip, a transverse DW is typically formed. While, the vortex domain wall occurs when the magnetic nano-strip is relatively wider and thicker [189][190]. As shown in the right column of Fig. 2.4, the DWS has a strong perpendicular magnetic anisotropy (PMA, such as Co/Ni magnetic multilayers), where the magnetic domains are magnetized in the out-of-plane directions. A Neel type DW usually occurs in a narrower PMA DWS, while a Bloch type DW typically forms in a wider PMA DWS. Typically, the probabilities of left-handed and right-handed rotations of the DW are equal. However, an additional Dzyaloshinskii-Moriya interaction (DMI) [191][192] can favor and stabilize a particular DW configuration [45][46][193] in the presence of broken inversion symmetry.

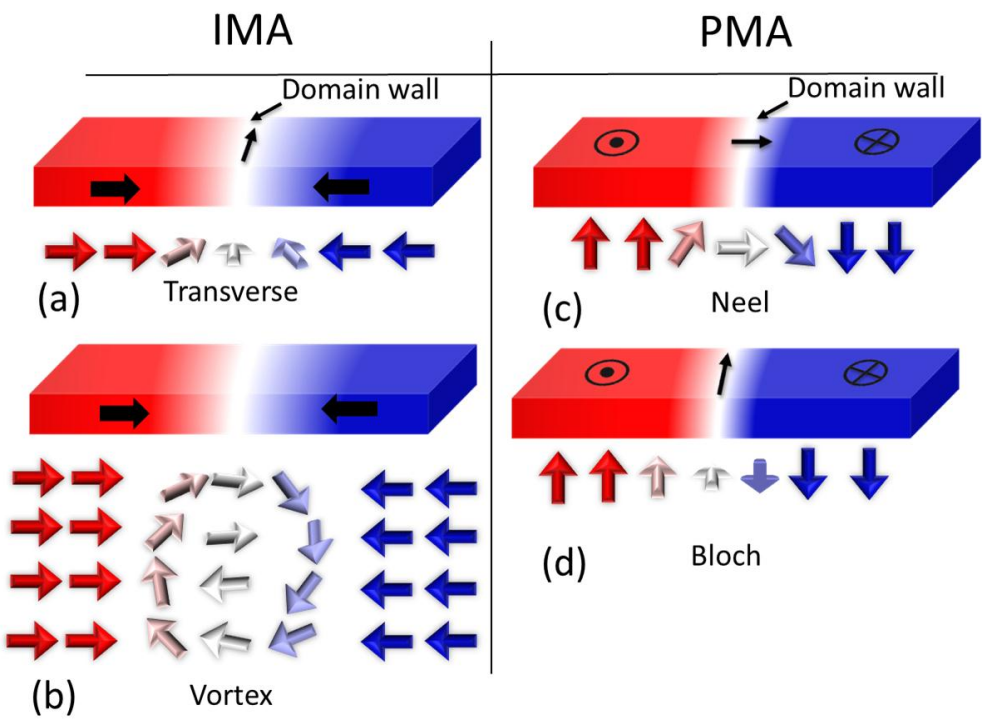


Fig. 2.4 Typical domain wall structure (a) in-plane magnetic anisotropy transverse head-to-head transverse DW (b) IMA vortex DW, (c) perpendicular magnetic anisotropy nanowire with Neel DW, and (d) PMA Bloch DW.

An external magnetic field can be used to move domain walls in magnetic nano-wire. However, similar to the switching of magnetization in spin valves due to current induced STT, a more energy efficient way to induce domain wall motion is applying an electrical current along the domain wall strip [40][46], [77]-[80]. When electrons flow through a fixed domain in the DWS, they become spin-polarized. The spin-polarized electrons exert spin-transfer torque on the magnetic moments in and around the domain wall region. If the applied current density is above the *critical current density*, the exerted STT can overcome the pinning force, leading to steady domain wall motion (DWM). The critical current density is defined as the minimum current density applied along DWS to induce a steady DWM. Its magnitude is proportional to hard-axis anisotropy and the domain wall length. Earlier current induced domain wall motion experiments are based on IMA ferromagnetic nanowires with the critical current density in the order of ~10^8A/cm^2.

Several issues, such as excessive Joule heating and reliability concerns, may accompany this relative high current density. In order to reduce the critical current density of DWM, a scaled PMA DWS is used in [126]. The hard-axis anisotropy of a PMA device reduces with lower device thickness and becomes much smaller than that of an IMA device. Moreover, the DW length in a PMA DWS is in general smaller than that in an IMA DWS. Therefore, a scaled PMA magnetic nano-strip can achieve much lower critical current density to induce steady DWM, leading to smaller power consumption.

2.4. Spin-Orbit Torque

In spin valve or DWS, the spin current is generated by passing charge current through a FM or spin polarizer. In these cases, the efficiency of generating spin current from charge current is limited by the polarization efficiency of the FM. Recent experiments show that spin current can be generated more efficiently through spin-orbit interaction (SOI) [50]. Later on, current induced SOI was experimentally demonstrated in I/FM/HM structure (I: Insulator, FM: Ferro-magnet, and HM: non-magnetic heavy metal) and applied in efficient magnetization switching [48]-[55], domain wall motion [45]-[47], [56], and spin-torque oscillations [49][57][58].

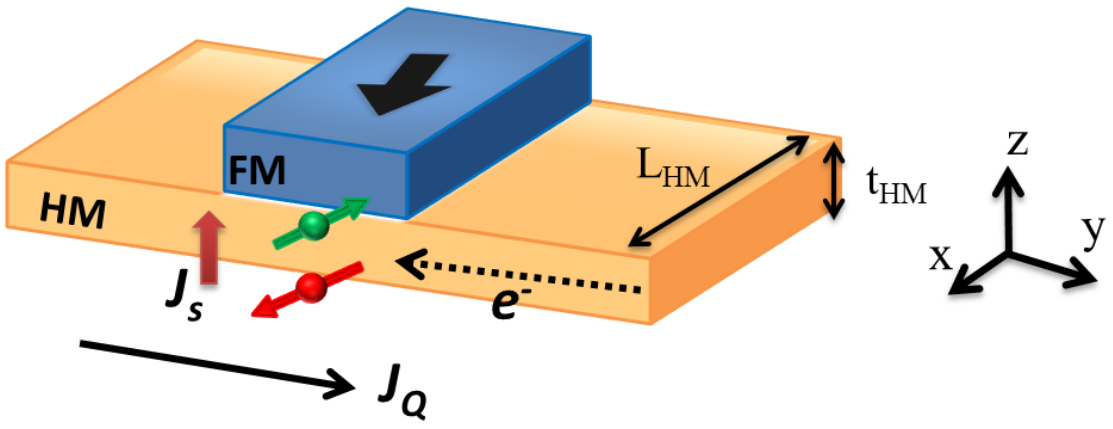


Fig. 2.5 Charge current applied in non-magnetic heavy metal with strong spin-orbit coupling is converted to pure spin current due to spin Hall effect

The observed phenomenon can be explained by either Rashba effect [47][60] or Spin Hall Effect [61]. Rashba effect arises from the broken structural inversion symmetry in a material system [47][60]. This structural inversion asymmetry first produces an electrical potential along the asymmetry direction. When electrons travel through this electrical potential, they experience an effective magnetic field. This magnetic field induces spin polarization of electrons based on the magnetic momentum. Therefore, a pure spin current can be generated. The other explanation of the observed phenomenon is based on Spin Hall Effect (SHE) [61]. Referring to Fig. 2.5, when electrons flow through a non-magnetic heavy metal (HM) (in ±y direction) with strong spin-orbit coupling, opposite electron spins accumulate on the opposite surfaces of HM. Thus, a pure spin current ($\boldsymbol{I}_s$) in the ±z direction is generated, which exerts a spin-transfer torque on the adjacent FM. The STT will switch the magnetization or drive the FM into steady oscillation. The relationship between the generated spin current ($\boldsymbol{I}_s$) and the applied charge current ($\boldsymbol{I}_Q$) can be expressed as follows:

$$\boldsymbol{I}_s = \frac{A_{FM}}{A_{SH}} \theta_{SH} (\boldsymbol{\sigma} \times \boldsymbol{I}_Q) \tag{2.4}$$

Where A_{FM} is the area of the adjacent FM area and A_{SH} is the cross-sectional area of HM. θ_{SH} is the spin Hall angle, which is defined as the ratio of generated spin current density to the applied charge current density. Recently, large spin Hall angle was experimentally demonstrated in different heavy metal materials, such as Pt [62][63], β-Ta [49][64], β-W [65], and CuBi alloys [66]. $\boldsymbol{\sigma}$ is the electron spin polarization, which is transverse to both the spin current and charge current directions.

By observing the above equation, it can be easily seen that the generated spin current can be larger than the applied charge current if $\theta_{SH} A_{FM} / A_{SH}$ is larger than 1. The reason comes from the scattering of electrons at the HM and FM interface, which generates multiple units of angular momentum. The spin current generation efficiency (η shown in equation-2.3) in spin valve is usually less than 1. Therefore, it is more efficient to generate spin current utilizing spin Hall effect.

2.5. Summary

In this chapter, we briefly reviewed several fundamental spin-transfer torque devices and the associated underlying physical phenomena. Firstly, we discussed the GMR and TMR effects in vertical spin valves. Then, local and non-local measurements in lateral spin valve were introduced. We also presented current induced domain wall motion in magnetic domain wall strip and efficient spin current generation due to Spin Hall effect. In all of the STT devices discussed in this chapter, the magnetization of the nano-magnets can be manipulated to perform various Boolean and brain-inspired computing that will be presented in the latter chapters of this dissertation.

3. BOOLEAN LOGIC DESIGN: SPIN-MEMRISTOR THRESHOLD LOGIC

In this chapter, we present a Spin-Memeristor Threshold Logic (SMTL) gate design, where memristive cross-bar array (MCA) is used to perform current mode summation of binary inputs and low power, current mode spintronic threshold device (STD) is employed to carry out the energy efficient threshold operation [133].

3.1. Introduction

Recently, a CMOS compatible and programmable resistive device, called *memristor*, has earned a lot of interest [67]-[69]. Such devices can be integrated into metallic cross-bars to obtain high density memristive cross-bar arrays (MCA). The continuous resistance range can be obtained in memristors, leading to a possible design of multi-level, non-volatile memory [69][70]. Application of the specific device characteristics of memristors in unconventional computing schemes like neural networks [71][72] and threshold logic (TL) [73]-[75], has been explored in recent years.

A threshold logic gate (TLG) operation essentially constitutes of summation of weighted inputs, followed by a threshold operation [76]. While a memristor array can be employed to perform analog summation of binary voltage input signals, the thresholding operation requires the application of a current comparator circuit. Such a comparison operation can be obtained using conventional analog circuits based on current mirrors [73] or voltage comparators [74][75]. However such analog CMOS circuits often consume significant power and area, thereby eschewing the energy and density benefits of nano-devices. Rather than depending upon analog CMOS circuits for implementing current comparison, it would be desirable to explore nano-devices that can directly provide such a current mode thresholding characteristic.

Recently, high speed magnetization switching of a nano-magnet due to current induced spin-transfer torque (STT) have been demonstrated in experiments [77]-[80]. Such a phenomenon can be used to design compact and low power current mode spintronic switches and simultaneously provide energy efficient current-to-voltage conversion. Application of such spin-transfer torque switches in memory [94][182], digital [81][183], analog [82], and neuromorphic computing applications [83], have been explored earlier. Such nano-scale, spintronic devices inherently act as compact, ultra-low voltage and fast current comparators and hence, can be highly suitable for memristor based TLG design.

In this chapter, we present a spin-memristor threshold logic (SMTL) design using such spin-transfer torque switches based on magnetic domain wall (DW) motion [79]. The magneto-metallic domain wall switch allows ultra-low voltage operation of memristive TLGs, leading to low energy dissipation at the gate level. We name our proposed domain wall switch structure as spintronic threshold device (STD). It can facilitate ultra-low voltage current mode interconnect for the design of fully programmable, large TL blocks. This helps to achieve highly reduced energy dissipation in programmable interconnects. Notably, in CMOS look up table (LUT) based conventional FPGAs, more than 90% of energy can be ascribed to programmable switches and interconnects [85]. Further, the STD being non-volatile magnetic switches inherently act as a latch and hence can facilitate fully pipelined connection of multiple TLG stages without the insertion of additional memory elements like flip-flops. This can provide high performance and integration density for complex data processing blocks. The aforementioned factors combined together lead to ultra-low energy consumption of the proposed design.

In this chapter, we also present a comprehensive methodology for SMTL design, synthesis and optimization and compare its performance with conventional CMOS FPGAs.

3.2. Design of TLG First Stage using Memristive Cross-bar Array

In this subsection we review the recent progress in memristive cross-bar array design, programming and its application as the first stage of threshold logic computation.

A threshold logic operation shown in Fig. 3.1a, can be expressed as follow:

$$Y = sign(\sum_{i=1}^{n} X_i W_i + b_i) \tag{3.1}$$

where, X_i's are multiple binary inputs to a threshold gate, W_i's are scalar weights with which the corresponding inputs are multiplied (or scaled) and b_i is the bias for the i^{th} gate. Note that, W_i can be either positive or negative. Hence, depending upon the input combination (assuming unipolar values of inputs, i.e., 1 and 0) the summation can yield either a positive or a negative value, result of which is determined by the sign function (involving a comparison operation). The first stage of the threshold logic computation is the scaling and summation of the inputs, which can be implemented using an MCA, as shown in Fig. 3.1b. The detailed design and programming of MCA will be introduced in this subsection. The second stage of threshold logic computing is a 'sign' (in equation-3.1, or threshold) function, which will be implemented using the proposed spintronic threshold device described in the next subsection.

3.2.1. Multi-level MCA

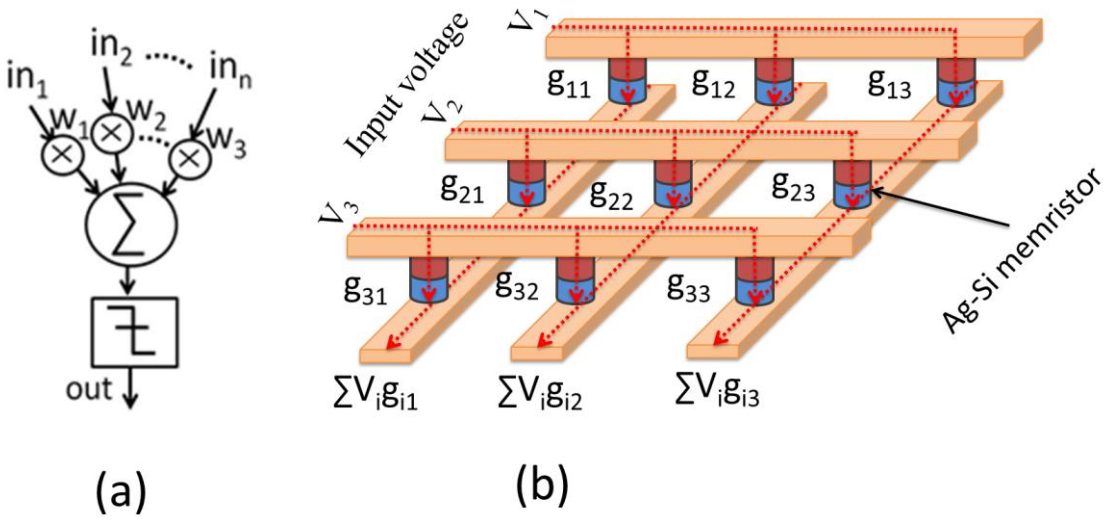


Fig. 3.1 (a) A Schematic representation of a threshold logic gate (TLG), (b) memristive cross-bar array

Fig. 3.1b depicts a MCA with two sets of metal bars (horizontal bars and in-plan bars). In such a MCA, memristor with conductance-g_{ij} interconnects i^{th} horizontal metal bar and j^{th} in-plane metal bar. More than 8-bit write accuracy for isolated memristors have been proposed and demonstrated in literatures [69][70]. However, for threshold logic design the bit-precision requirement can be significantly less (less than 4-bit, explained later in this chapter). The programming voltage applied across two cross-connected memristor, in a large-scale cross-bar array, results in sneak current paths through neighboring devices, which disturbs the state of unselected memristors. The application of access transistors and diodes can facilitate selective and disturb-free write operations to overcome the sneak path problem [90]. If the programming speed is not a major concern, the technique that can only program a single device at one time is also proposed in [91] without access transistors or diode.

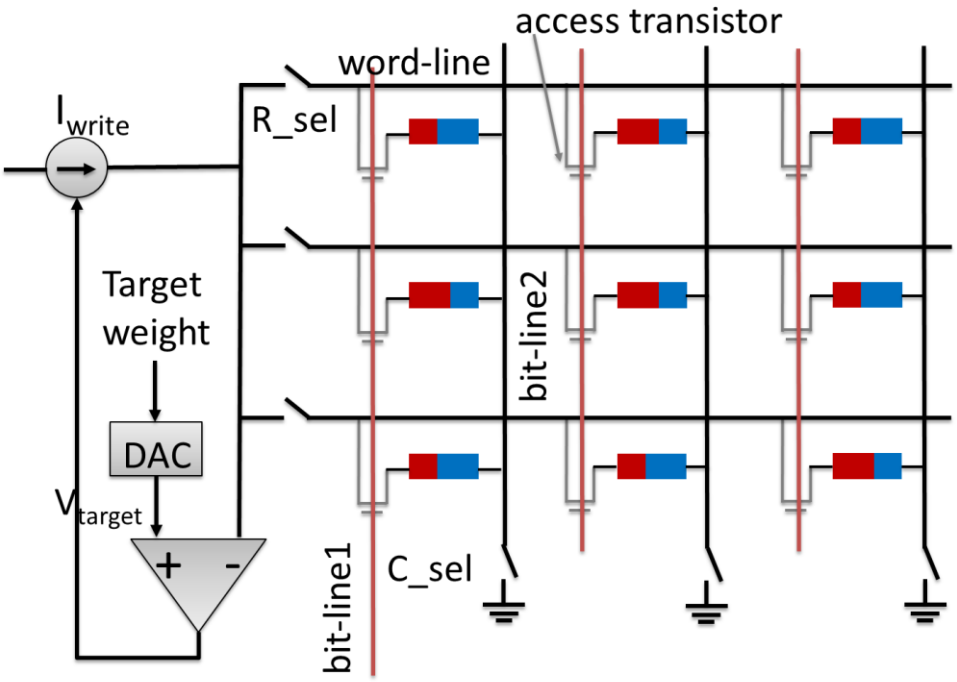


Fig. 3.2 A resistive memory array with multi-level programming periphery

A multi-bit memristor programming array-level scheme employing adjustable pulse width is shown in Fig. 3.2 [70][130]. In this scheme, when programming one specific memristor cell in the array, the corresponding set of the word line, the source-line and the bit line will be selected. In Fig. 3.2, only a single write unit is shared among all of the rows for infrequent write operations, while a dedicated programming cell can be assigned to each row for maximum write speed. This would allow writing of one column at a time, by selecting a particular world line. During the writing operation, a constant current will be injected into the selected cell and the voltage developed on the source line is compared with a comparator threshold. A digital to analog converter (DAC) is used to set the threshold proportional to the target resistance. As soon as the accessed memristor is programmed as the target value, the current source is disconnected. More precise tuning of memristor value can be achieved by applying a lower value of write current resulting in slower ramp in the resistance value. The write precision in the method described above is mainly limited by the random offset of the comparator, inaccuracy in the current source and DAC. Larger accuracy would entail higher design complexity for these blocks and lower write speed. The memristive devices (including Ag-Si) do exhibit a finite write threshold for an applied current/voltage, below which there is negligible change resistance [92]. Since the application of spin-based current comparator facilitates ultra-low voltage (and hence low current) operation of the memristors for computing as will be described in the following sections, the state of memristor in the MCA will not be disturbed for reading.

3.2.2. Threshold Logic Computing using MCA

For a TLG, the scaling and summation operations can be implemented using a MCA, as shown in Fig. 3.1b. If we assume that the outward terminals of the in-plane bars are connected to ground, for a given set of binary voltage inputs, the total current flowing out of an in-plane bar is the dot product of input voltages and the memristors' conductance values [69][76].

The above principle can be exploited in realizing current mode analog scaling (multiplication) and summation that corresponds to the first stage operation of a TLG. Several authors have proposed the design of hybrid TLG hardware based on memristive

cross-bar arrays and analog CMOS circuits, where analog circuits are employed to perform the second stage operation of the TLG, namely, thresholding [73]-[75]. For instance, application of analog current mirrors has been proposed for implementing memristor based hybrid TLG's in [73]. However such a design requires additional interconnect networks to realize fully programmable logic modules. Notably, energy consumption of interconnects dominate the total power budget of an FPGA. Authors in [74][75] applied CMOS voltage comparators for realizing the thresholding operation for memristor based TLGs. Application of analog amplifiers and comparators may lead to significant energy consumption. Authors in [76] recently demonstrated the use of a simple CMOS latch for thresholding operation. Such a scheme would need large voltage inputs (resulting in large current) to the memristors, so that a digital latch can directly sense the voltage mode output of a TLG. This would result in power hungry TLG blocks that may not be suitable for large-scale integration.

Thus, although memristors can provide an efficient mapping of the first stage operation of a TLG (namely current-mode scaling/multiplication and summation), the second operation, namely, the current mode thresholding, does not have a likewise 'matching' device. The above mentioned inefficiencies could be eliminated if an alternate device structure could be found that can perform the current mode thresholding operation in an energy efficient way. Next, we will present a spintronic threshold device design that can be ideally suitable for this purpose.

3.3. TLG Second Stage Design using Spintronic Threshold Device

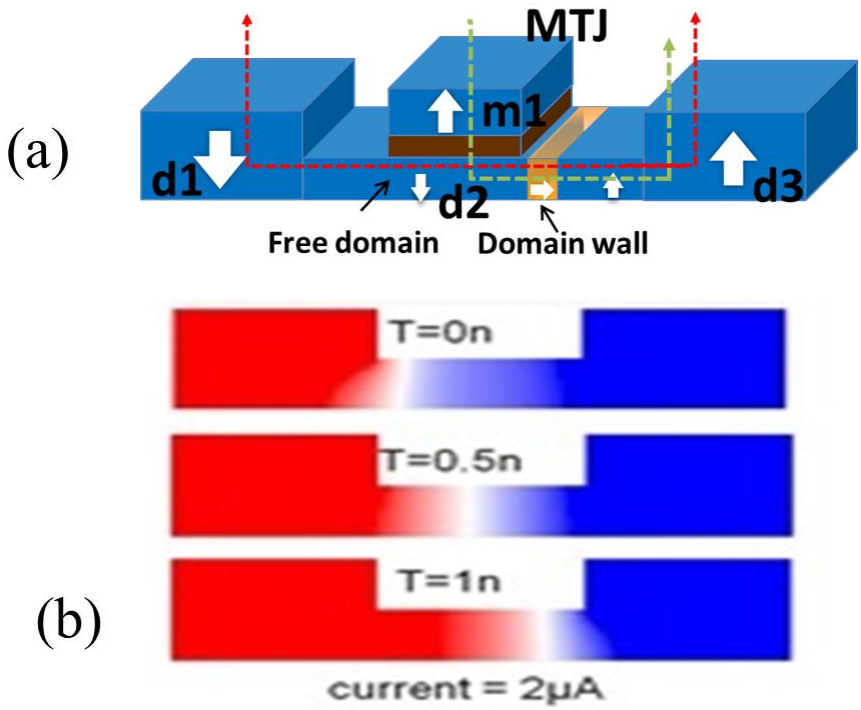


Fig. 3.3 (a) Device structure for Spintronic Threshold Device (b) Transient micro-magnetic simulation plots. Read color represents the 'down spin' corresponding to d1. Blue color represents the 'up spin' in d3. White color is the magnetic domain wall.

In this sub-section, we present the spintronic threshold devices (STD), based on magnetic domain wall, suitable for the design of energy efficient Spin-Memristor Threshold Logic (SMTL). This STD design will serve as the second stage of threshold logic computing, which is a thresholding ('sign') function.

The proposed spintronic threshold device structure is shown in Fig. 3.3a. It constitutes of two fixed magnetic domains (d1 and d3) and a free domain (d2, 20×40×3 nm^3). The magnetization of these two fixed domains are anti-parallel. Domain-1 takes the current input, while the domain-3 is grounded. The magnetization of domain-2 can be written parallel (or anti-parallel) to d1 if the total input current is injected from d3 to d1 (or from d1 to d3) [81]-[83]. The magnetization of free domain-2 is sensed using a

magnetic tunnel junction (MTJ), formed between a fixed magnet (m1) and d2. When d2 and m1 have the same magnetization, the effective resistance of the read MTJ is smaller and vice-versa. Thus, the proposed STD acts as a low power and compact current comparator that can be employed in energy efficient current mode threshold logic design.

Table. 3.1 STD device parameters

t	**L**	**W**	**M_s**	**E_b**
3nm	40nm	20nm	400emu/cm^3	$20K_BT$
α	**A**	**β**	**t_{ox}**	**Area(m1)**
0.01	10pJ/m	0.1	1.8nm	$20\times20nm^2$

The resolution of the device, i.e. the minimum current magnitude required to switch the free layer, is determined by the critical current density for DW motion. Several recent experiments have achieved sub-nanosecond domain wall motion, with a low current density [77][126]. Magnetic domain with perpendicular magnetic anisotropy can provide scaled device dimensions (thickness ~3nm and width <50nm) as well as relatively lower critical current density [78]-[80], [126]. More recently, application of spin-orbital coupling has been explored for reducing the required current for a given speed of domain wall motion by an order of magnitude [80]. These device optimizations can be used to engineer current thresholds of the order of ~2μA for 1ns switching. Fig. 3.3b shows the transient micro-magnetic simulation plots for the proposed STD design using Object Oriented Micro-Magnetic Framework (OOMMF, [95]) when supplied with a 2 μA current. The device parameters of STD are shown in Table 3.1. It can be seen the magnetic domain wall moves from the left free domain boundary to the right boundary within 1 ns. We will analyze the effect of STD resolution on the energy efficiency of SMTL later in this chapter.

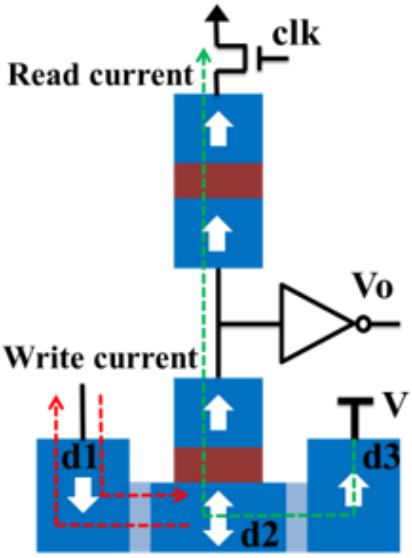


Fig. 3.4 STD state sense circuit

The effective resistance of the MTJ formed between m1 and d2 is smaller when they have the same magnetization and vice versa. The ratio of the two resistances is defined in terms of tunnel magneto resistance ratio (TMR). STD forms a voltage divider with a fixed reference MTJ, as shown in Fig. 3.4. A TMR of ~400% can provide a voltage swing close to Vdd/2 that can be detected using a simple CMOS inverter. Static current in the voltage divider can be minimized for a given operation speed by increasing the MTJ oxide thickness. For 500MHz clock frequency, the oxide thickness was determined to be ~1.8nm that resulted in a total power dissipation of ~0.15µW for the sensing unit (including the clocking power), for a supply voltage of 0.6V.

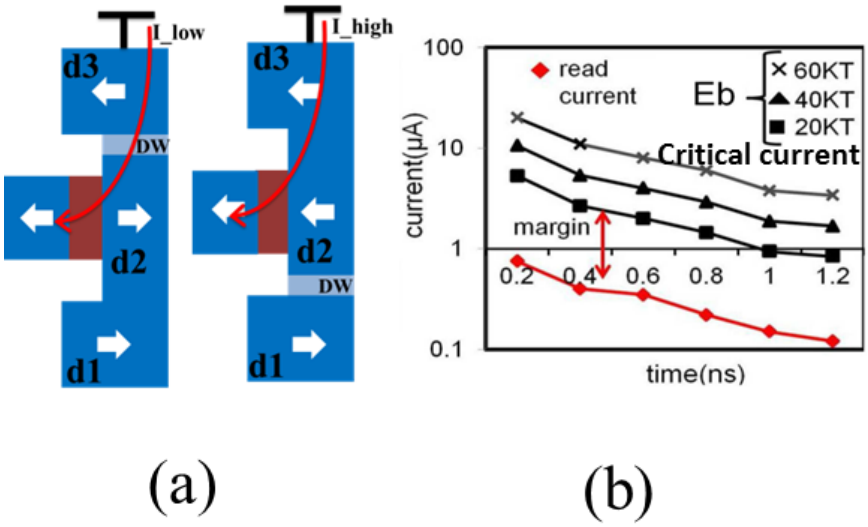
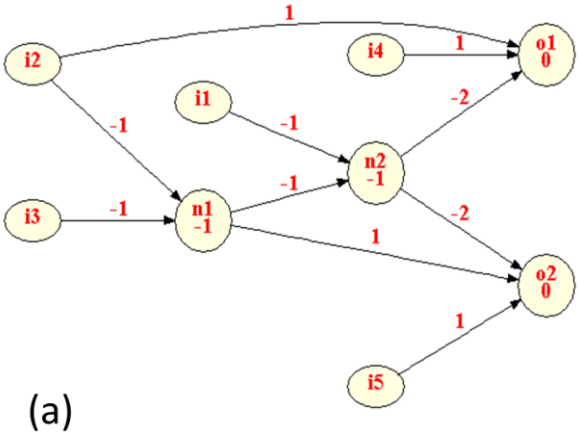


Fig. 3.5 (a) read current for different d2 state (b) read current margin to critical current

Note that in the detection circuit, the terminal d3 of the STD is connected to Vdd. Hence, the transient evaluation current flows from d3 to d2 as shown in Fig. 3.5a. The current required for the DW motion increases proportional to the switching speed. The magnetization of d2 is not disturbed by the read current with a short duration and low magnitude. The read margin can be seen in Fig. 3.5b. Apart from device scaling, the STD critical current can also be lowered by manipulating other device parameters, like the anisotropy energy (E_b) of the magnet.

In general, the circuit in Fig. 3.4 forms the 'sign' function required in equation-3.1. The STD works as a current comparator and its input is the output current of the first stage MCA. If the input current to STD is larger than the critical current, the output of the inverter is high, and vice versa. Next, we describe circuit design for combining the MCA and STD to implement threshold logic array design.

3.4. Design of SMTL Array



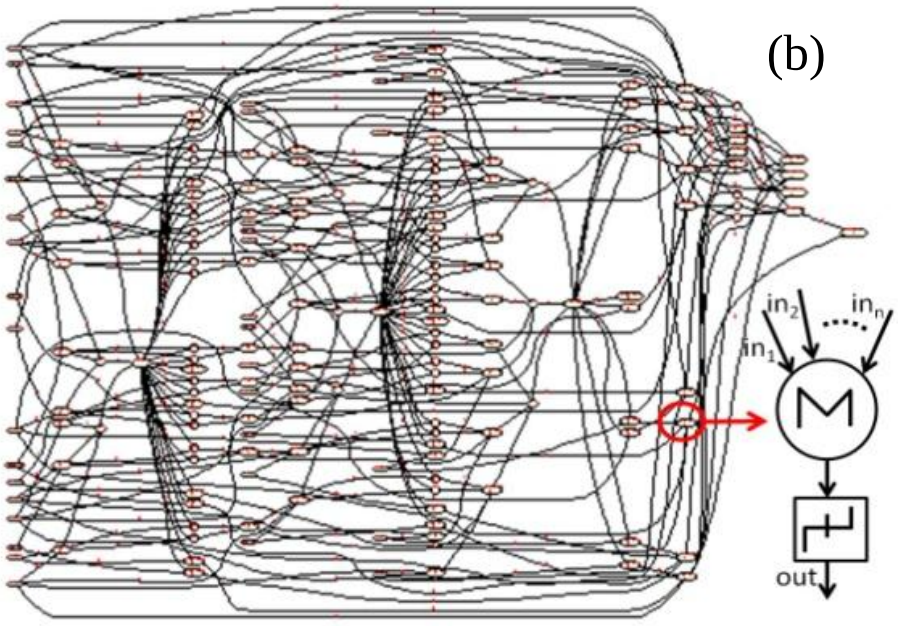


Fig. 3.6 (a) synthesized ISCAS85 benchmark C17 threshold logic network. (b) synthesized ISCAS85 benchmark-C432 (27-channel interrupt controller) threshold logic network.

Fig. 3.6a and 3.6b show two threshold logic networks (TLN) for an ISCAS85 benchmark, C-17 and C-432 [73], obtained using the threshold logic synthesis (TELS) technique presented in [84]. In Fig. 3.6a, each circle represents one threshold logic gate. The connections between each TLG are the fan-ins and fan-outs. The node without fan-ins is the input node. The node without fan-outs is the output node. The weights are labelled along the connections. Node-*i1 to* node-*i5* are the input nodes. Node-*n1* and node-*n2* are internal TLGs. Node-*o1* and node-*o2* are the output nodes. The bias values are labelled inside of the TLGs. The synthesized threshold logic network in Fig. 3.6b consists of 15 stages, while each stage is comprised of N_i threshold logic gates. The maximum number of fun-ins for each TLG is 4. Comparing these two benchmarks, it can be easily seen that 'C-17' is a simple TLN, while 'C-432' is a much larger scale TLN. In order to show our design is compatible to large scale TLN mapping, we will use C432 as a design example in this work.

TLN constitutes of a network of TLGs which can be divided into multiple stages. Each circle in the plot represents one TLG and the TLGs in the same column will be mapped to the same MCA stage. The connections between the TLGs are implemented by the MCA described in previous subsection, whereas the conductance of memristor corresponds to the synthesized weights. In such a multi-stage logic scheme, each MCA stage would comprise a number of TLGs receiving inputs from its previous stage and communicating their outputs to the next stage. Let us consider the design of such a stage using MCA and the STD device.

TLN constitutes of a network of TLGs which can be divided into multiple stages. Each circle in the plot represents one TLG and the TLG in the same columns will be mapped to the same MCA stage. The connections between the TLGs are implemented by the MCA described earlier, whereas the conductance of memristor corresponds to the synthesized weights. In such a multi-stage logic scheme, each MCA stage would comprise a number of TLGs receiving inputs from its previous stage and communicating their outputs to the next stage. Let us consider the design of such a stage using MCA and the STD device.

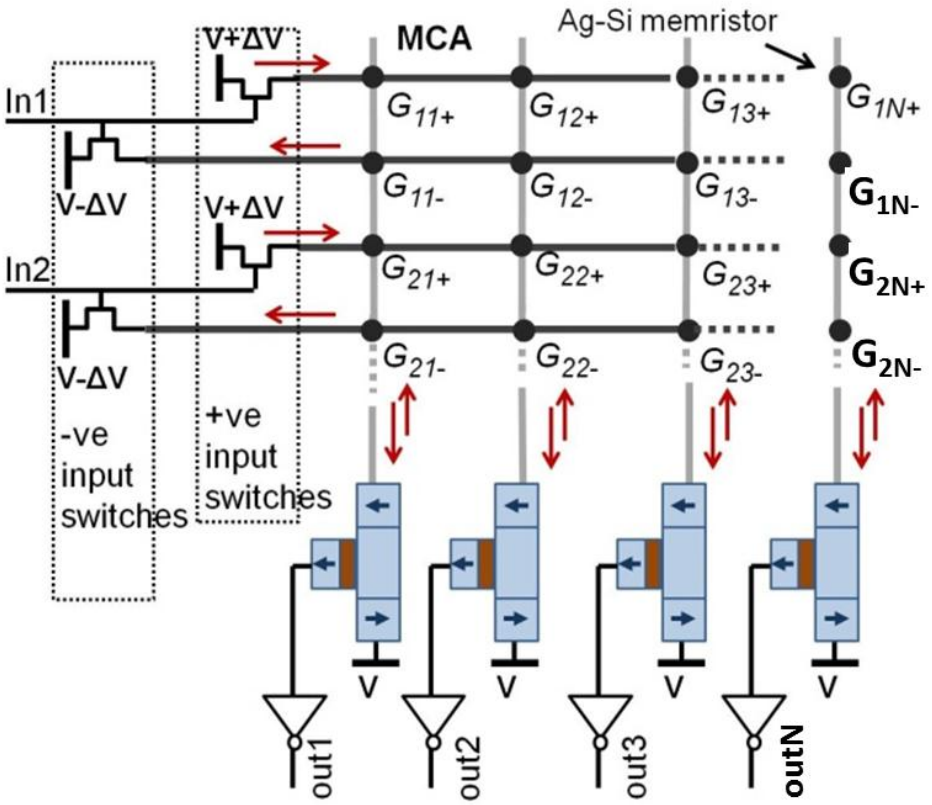


Fig. 3.7 Circuit of one single threshold logic stage using MCA and STD

Fig. 3.7 shows the circuit realization of a single MCA stage that contains N number of TLGs based on STD. Each stage has a maximum of M inputs (which can be set as a parameter during the implemented MCA mapping tool), and N STDs, forming the N TLGs. The i^{th} input to the MCA may connect to the j^{th} STD (i.e. j^{th} TLG) with either a positive, negative or zero weight. This is achieved by programming either of G_{ij+} or G_{ij-} to the corresponding weight value (The bias of each TLG can be viewed as the weight of an extra input whose value is always high). For zero weight (i.e. no connectivity), both G_{ij+} and G_{ij-} are driven to high resistance *off* state. The input signal to MCA is received through PMOS transistors with source terminals connected to a potential $V+\Delta V$ (for positive weights) and $V-\Delta V$ (for negative weights), where ΔV can be less than ~50mV. These input transistors act as deep triode region current sources (DTCS) [82][83]. The STD is connected to a DC supply V. This effectively clamps the potential of all the

vertical metal bars in Fig. 3.7 to the same potential (due to small resistance of the magneto metallic STD). Thus, small static power consumption is achieved due to the fact that the static computing current flows across a small terminal voltage of ΔV. Moreover, the dynamic power dissipation on the metallic interconnects forming the programmable cross-bar is also largely reduced due to ultra-small voltage swing. The direction of current flow at the input of a STD, and hence the output of a TLG, would depend upon the input data and the corresponding weights (determined by the programmed memristor conductance). Note that, the resistance values for the memristors can be chosen large enough to avoid inaccuracy due to resistive voltage division between the DTCS transistors and the memristors in a given row. The output of the MTJ based detection circuit associated with each TLG, in turn, drives a corresponding DTCS transistor that communicates the outputs of the TLGs to the next stage.

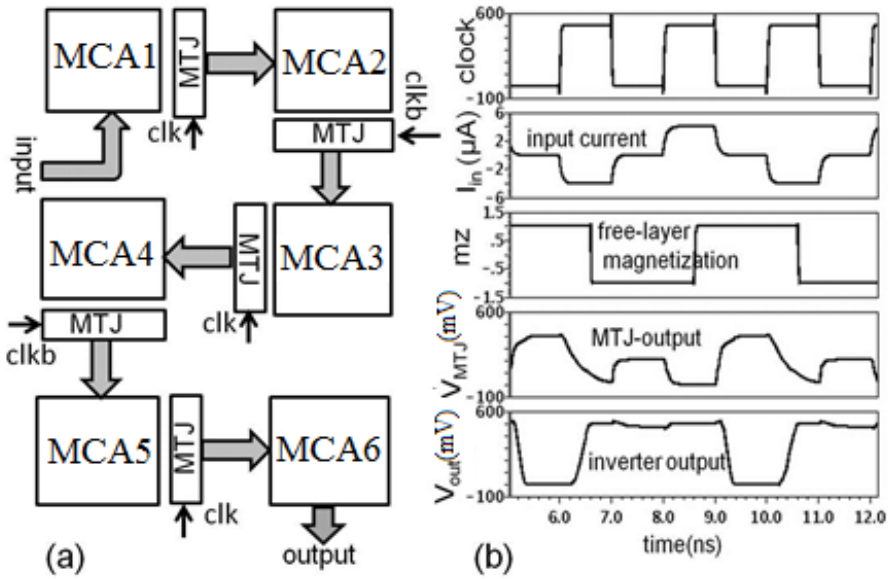


Fig. 3.8 (a) 2-phase pipelined MCA blocks for large-scale logic design, (b) transient simulation plots for a single TLG.

Due to the non-volatility of the STD, the MCA design described above can be extended to realize a 2-phase pipelined architecture composed of large number of such

hybrid arrays without inserting the CMOS latches, as shown in Fig. 3.8a. In such a design, consecutive MCAs operate with complementary clock phases. For instance, in Fig. 3.8, when the clock is high, MCA1 is driving MCA2, and MCA3 is driving MCA4. When the clock goes low, the driver and driven MCAs exchange roles. The exemplary simulation plots for a single TLG is shown in Fig. 3.8b.

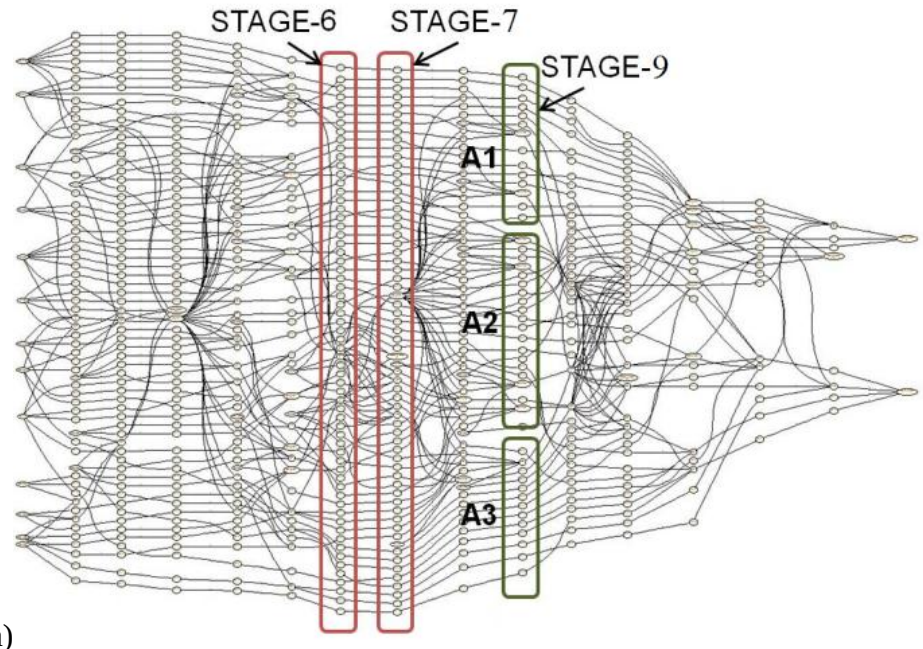
Next we discuss optimal pipelining and partitioning scheme for the mapping of large logic blocks on to the SMTL array.

3.5. Optimal Pipelining and Partitioning of SMTL Arrays for Logic Mapping

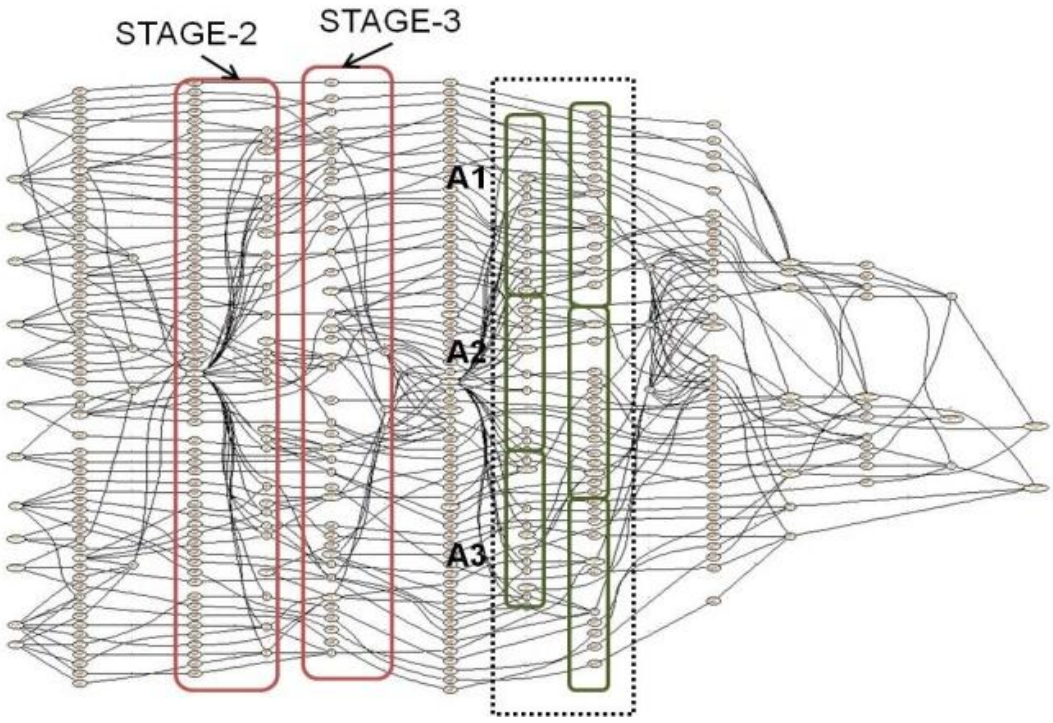
3.5.1. Pipelining Optimization

As mentioned earlier, each STD acts as a non-volatile latch and hence, a multi-stage MCA can be pipelined without insertion of additional CMOS latches. However, logic paths in the threshold logic network (TLN) of a generic logic block (like for C432 shown in Fig. 3.6b) may be unequal. Hence '*buffer-nodes*' need to be inserted to make them equal and to facilitate fine grained pipelining. The number of buffers needed depends upon the granularity of pipelining. In case, each MCA stage is pipelined, the number of buffers is the maximum. Fully pipelined TLN for C432 is shown in Fig. 3.9a, where each circle represents one TLG and the TLGs in the same column are in the same stage. In such a TLN, each stage is mapped into a separate MCA stage. For a given switching speed of the STD, this configuration yields maximum throughput. However, the total energy consumption also depends upon the total number of TLG nodes.

Combining two MCA stages to form a single pipelined stage (Fig. 3.9b) reduces throughput by half, however the total number of nodes for most benchmarks was found to reduce by a larger factor, which leads to reduced energy consumption. Note that, despite using multiple MCA layers per pipeline stage, the same throughput can be maintained by increasing the current injection, i.e., the switching speed of the STD.



(a)



(b)

Fig. 3.9 synthesized C432 pipelined threshold logic network. (a) Fully pipelined architecture (b) two TLG stages combined with one pipeline stage.

Fig. 3.10 shows the power consumption of C432 for different number of MCA levels (note, single MCA level for a pipelined stage implies maximum pipeline granularity) in a single pipelined stage. The power component due to the detection unit ('*Power_det*' due to MTJ voltage divider, clock and inverter) reduces with reducing pipeline granularity, because of reduction in total number of TLG nodes in the resulting TLN (Fig. 3.10b). However, to maintain the same throughput, larger currents need to be supplied by the DTCS transistors, which lead to increase in static power consumption in the MCA ('*Power_MCA*' in Fig. 3.10a). For most ISCAS85 benchmarks a pipelined stage with 2-MCA levels yielded optimal results.

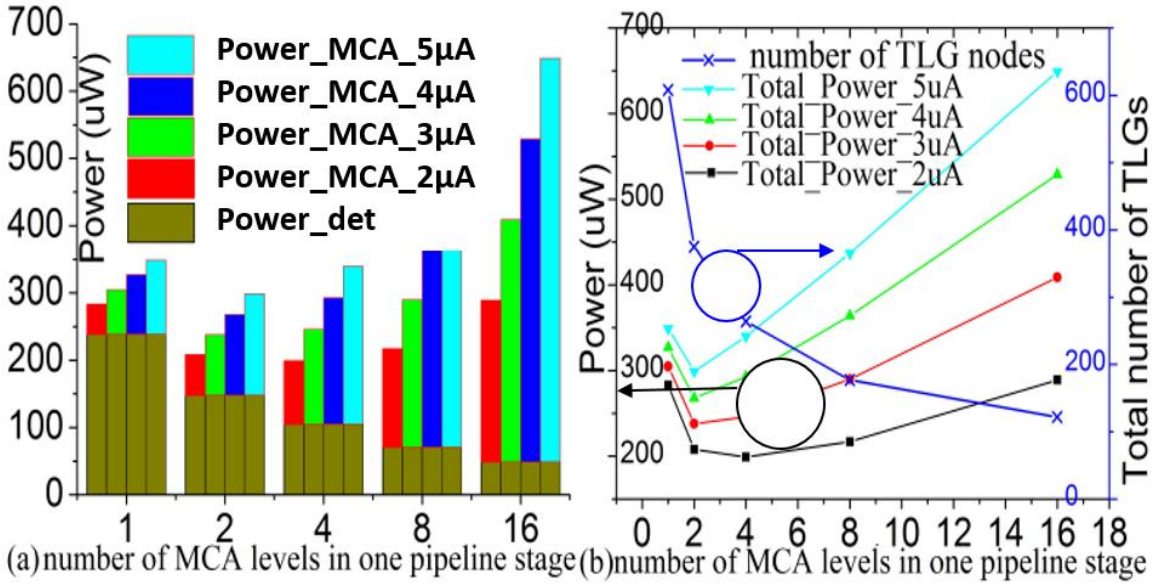
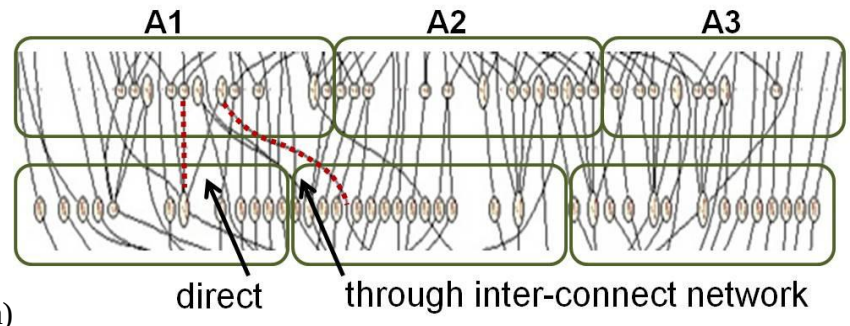


Fig. 3.10 : (a) Power consumption of different pipeline configurations (b) tradeoff between power and area. 'Power_MCA_5uA' represents the power of memristor cross-bar array when the DTCS current is 5uA. 'Power_det' is the power of detection module including MTJ-voltage divider, clock and inverter

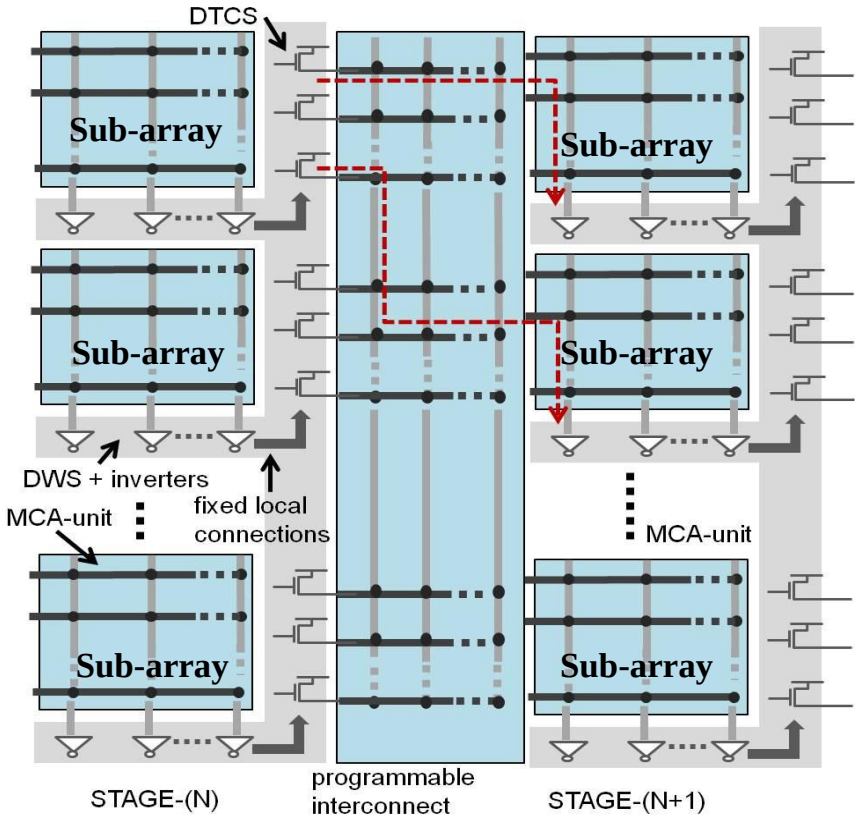
3.5.2. Partition and Interconnects

So far we assumed that each stage of the pipelined TLN is assigned to a single large dimension MCA. In such a design no additional interconnect network is required, as, the outputs of the n^{th} MCA stage can directly connect to the inputs of the $(n+1)^{th}$ MCA stage

using the scheme shown in Fig. 3.9. Due to the absence of additional interconnect power dissipation, this leads to the minimum energy solution (Fig. 3.12a). However, in this case, the MCAs have sparse connectivity (due to having large number of inputs but each input connecting to only few outputs, determined by the fan-in limitation) due to which the overall area efficiency is significantly sacrificed, as shown in Fig. 3.12b. To reduce the overall area, each pipeline stage can be divided into multiple smaller dimension sub-arrays (*Ai's* shown in Fig. 3.9b and an enlarged version in Fig. 3.11a). In this case, some of the inter-layer connections can still be directly routed to the next stage (Fig. 3.11a). However, some others (between nodes that are not located on directly opposite MCAs) need to be routed through an additional routing network. Such a design scheme is shown in Fig. 3.11b. For reducing MCA dimensions (implying the use of large number of smaller MCA modules in a single stage), the usage of the interconnect network increases. This also necessitates larger and longer interconnect array, leading to larger parasitic resistance drops along the current signal paths, mandating the use of larger voltage. As a result, energy component due to interconnect increases. Fig. 3.12 shows the tradeoff between area and power of SMTL with respect to the size of the sub-MCA array size. A design choice can be made based on priority.



(a)



(b)

Fig. 3.11 (a) Enlarged green square part of Fig. 3.9b (b) SMTL network partition architecture

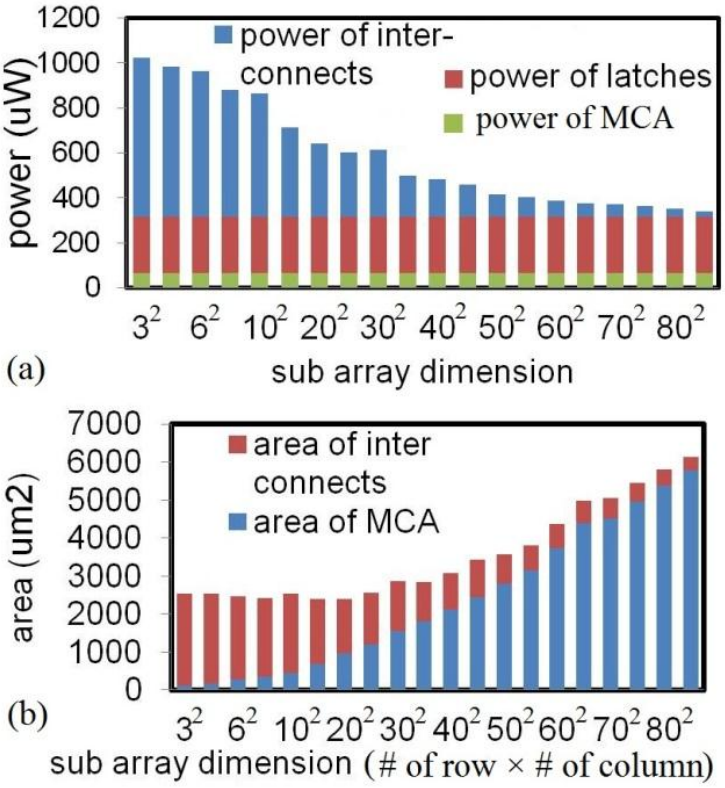


Fig. 3.12 Relationship between (a) power, (b) area and sub-array dimension, (larger dimension implies lower number of sub-arrays needed)

3.6. Simulation and Synthesis Algorithm

In this subsection, we discuss the synthesis scheme used in this chapter.

Fig. 3.13 shows the high level overview of the SMTL synthesis and hardware mapping methodology employed in this work. We employed threshold logic synthesis (TELS) algorithm proposed in [84] to do the initial synthesis, which reads a logic description and generates the functionally equivalent threshold network. Some important parameters like the fan-in restriction of TLGs and defect tolerance in the weights can be preset as parameters [84].

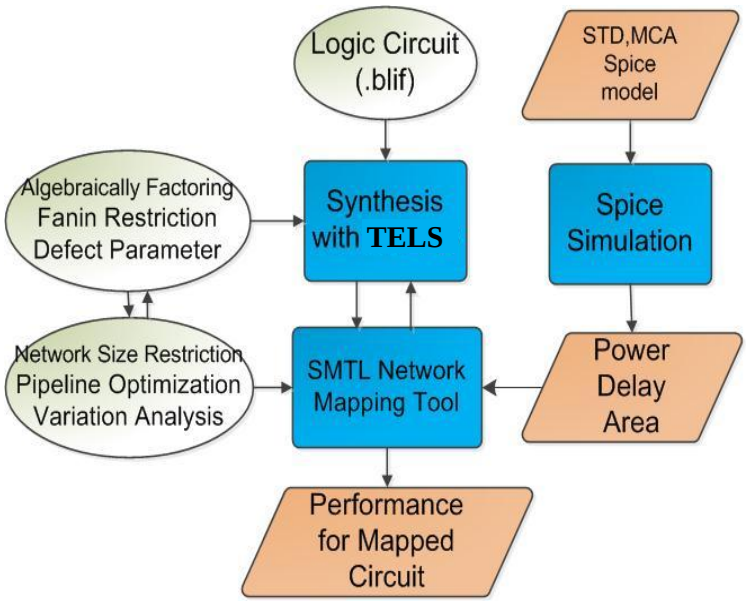


Fig. 3.13 Proposed design methodology

The SMTL mapping algorithm proposed and implemented in this paper, shown in Fig. 3.14, reads the synthesized TLG network and maps it to SMTL hardware. The tool first reorders the positions of TLGs in each stage so as to minimize the use of the interconnect network. This is achieved by placing the TLGs in the sub-arrays such that the use of direct links between face-to-face MCAs (as depicted in Fig. 3.11a) is maximized. Next, if the number of nodes in the current stage exceeds the restriction (number of MCA in a given stage times MCA size), one or more nodes are moved to next stage. This is done in a way that minimizes the number of intermediate buffers. The nodes without fan-out to next one stage are selected with highest priority, following which, the nodes with minimum fan-in's are shifted.

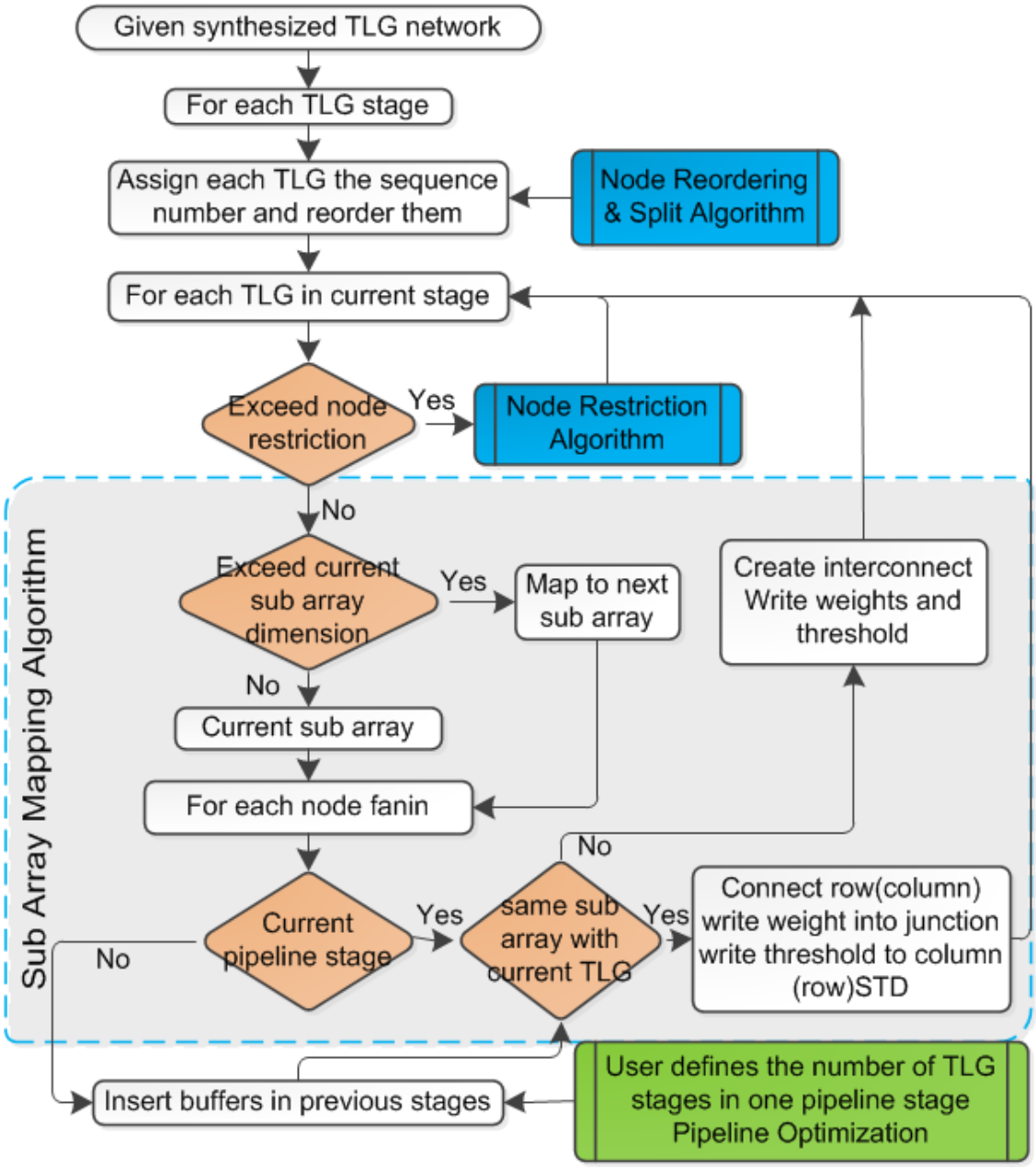


Fig. 3.14 SMTL network mapping algorithm

Some of the layers in the SMTL netlist may have very small number of nodes, for which, the use of a separate MCA unit may be wasteful. In TELS such nodes are incorporated in the MCA units corresponding to the previous stage, through the provision

of a small numbers of programmable backward connections (from output of an MCA back to its input).

The fan-out number of some nodes can be very large. Such TLGs communicate evenly to all the MCAs in the next level, making heavy use of the interconnect network. Such high loading can lead to significant voltage division between the DTCS source and the receiving memristors, leading to significant lowering of the input voltage and the current for the loads. A simple way to address this issue is to split the large fan-out nodes into multiple smaller nodes.

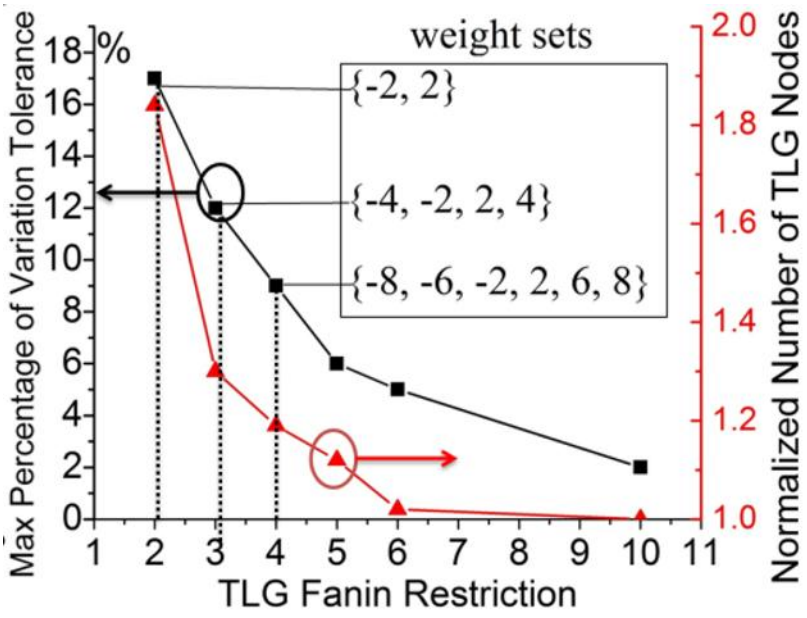


Fig. 3.15 the relationship between variation tolerance, TLG fan-in restriction and number of TLGs

Larger TLG fan-in generates denser SMTL network with smaller number of TLG nodes. This can provide larger area and energy efficiency. However, simulations show that larger fan-in restriction leads to reduced variation tolerance for memristor values, as seen in Fig. 3.15. In this plot, variation tolerance is defined as the standard deviation (σ) value for which total 10^5 test vector simulation gave zero errors. The variation tolerance

increases for lower fan-in restriction, but the use of lower fan-in TLGs results in larger number of nodes, leading to increase in overall area (Fig. 3.15). In this work we choose the fan-in restriction to be 4 (leading to a variation tolerance of ~9%). There are only 6 different levels of memristor conductance needed for mapping the TLG weighs, therefore the programming bit resolution for memristor is 3 bit. Note that, in this work we have assumed that the memristor programming thresholds are large enough, such that passing small computing currents (few μA) does not significantly disturb their state [68].

Next we discuss the performance of SMTL and compare it with conventional CMOS programmable logic based on CMOS LUTs.

3.7. Performance and Prospects

In the conventional FPGA based TLG design, the total power consumption is dominated by the interconnect power. Note that more than 90% of energy can be ascribed to programmable switches and interconnects [85]. The reason is the fact that the FPGA interconnect circuit has an extremely low utilization rate (~12%) for purpose of programmability. The energy and delay of 4-input LUT based FPGA for ISCAS85 benchmark using 45 nm technology is shown in Fig. 3.16. While in our proposed SMTL design, the energy efficiency mainly comes from four aspects. **1)**: The interconnect energy dissipation in the metallic cross-bars as well as the interconnect network is drastically lowered due to ultra-low voltage (~50mV), current mode signaling between the MCA layers, which comes from low voltage, low current operation of spin-transfer torque based threshold logic gates. The STD device can sense and compare the ultra-low current (few μA) enabling ultra-low voltage biasing of the MCA and hence, low voltage operation of the threshold gates. As a result the static power consumption, due to direct current paths, is largely reduced. Note that in the SMTL design, memristors play the dual role of computing elements as well as programmable interconnects. This can be contrasted with earlier approaches where memristors were employed only as programmable interconnects [87] or only as computing elements [73]. **2)**: In our proposed threshold logic network design, the output inverters of a particular MCA layer drives only the DTCS transistors that in turn supply current to the next MCA stage. Since a small terminal voltage ΔV is applied across the MCA, the dynamic power consumption

(CV^2f) in large number of programmable interconnects is largely reduced. Such low voltage operation of the MCA can also significantly reduce the disturb rate of the programmed memristors and can enhance the retention time of the hardware. **3)**: The STD achieves energy efficient current to voltage conversion with the help of MTJ based voltage divider. This eliminates the need of analog trans-impedance circuits based on current mirrors and amplifier, leading to high energy and area efficiency. **4):** Due to the non-volatility of STD, the proposed SMTL design can be extended to realize a pipelined architecture without inserting the CMOS latches. The throughput of the design is determined by a single stage delay. This delay in turn, is limited by the switching speed of the STD device. As discussed earlier, larger current per input can be used to increase the STD switching speed. Domain wall velocities of more than 60m/s has been demonstrated in literature [126], hence, for a 40nm long free domain more than 1GHz processing speed may be achievable. In this work a clock frequency of 500MHz has been used, corresponding to STD switching time of 1ns. Recently application of Spin Hall effect has been explored for bringing large reduction in domain wall current thresholds [79][114]. Such phenomena can be exploited in improving the resolution of scaled STD devices.

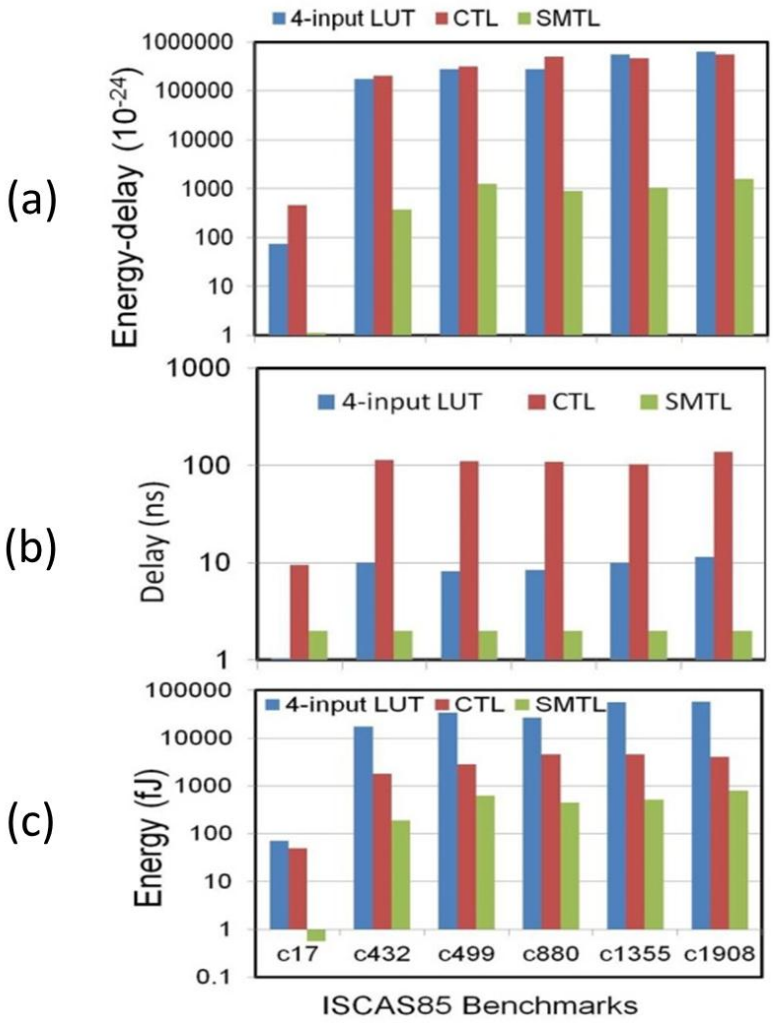


Fig. 3.16 (a) Energy-delay product, (b) delay and (c) computation energy of SMTL compared with 4-input LUT based FPGA [73] and CTL [73] for ISCAS85 benchmarks.

Fig. 3.16c compares the computation energy of the proposed SMTL design with that of 4-input lookup table (LUT) based FPGA and with capacitive threshold logic (CTL, a CMOS based implementation style for TLG [73]). The computing energy of proposed design is reduced by two orders of magnitude compared to the LUT based FPGA TLG. SMTL also shows much smaller delay compared with LUT and CTL, as shown in Fig. 3.16b. Results in Fig. 3.16a show around three orders of magnitude lower energy-delay product as compared to both the CMOS based schemes.

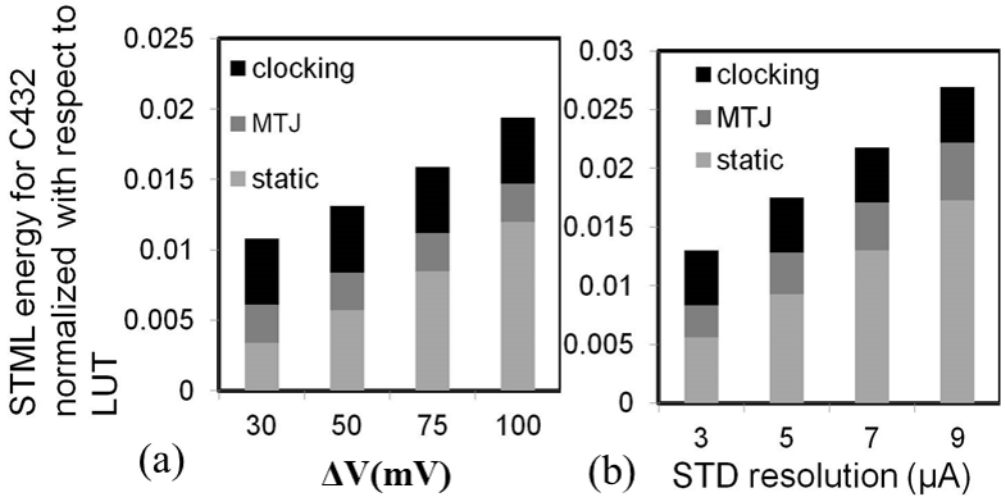


Fig. 3.17 SMTL energy for C432 normalized with respect to 4-input LUT for the case of (a) increasing ΔV, (b) increasing STD threshold for a fixed ΔV of 50mV ;LUT delay is ~10ns

The energy efficiency of the proposed design is dependent on two critical design parameters. First is the minimum achievable ΔV (voltage swing across MCA) in such a hybrid circuit. Fig. 3.17a shows that increasing ΔV increases the static power consumption due to current mode computing in MCAs (strength of DTCS transistors is reduced to keep the current drive constant). The second important parameter is the resolution of the STD device. As mentioned earlier, a poor resolution would require larger current per-input for a TLG. Corresponding results are shown in Fig. 3.17b, showing almost linear increase in computation energy with reducing resolution.

Integration of Ag-Si memristors with CMOS has been demonstrated in recent years [68][69]. The same is true with magnetic domain wall based memory cells [79][89][93]. However, integrating two novel technologies with CMOS to realize the proposed SMTL scheme can be significantly more challenging, especially when scaled dimensions of STD devices, such as used in this work, is targeted. However, the possibility of large energy benefits of the proposed design can be a motivating factor.

Some critical design parameters used in this work are given in table 3.2. The device characteristics for STD were obtained using the simulation framework for magnetic

domain wall strip presented in [89]. The system functionality is simulated in SPICE based on statistical behavioral STD model.

Table. 3.2 SMTL Design Parameters.

Free-domain size	$3\times20\times40$ nm^3	MTJ-t_{ox}	1.8nm
Ms	400 emu/cm^3	R_{MTJ} (parallel)	300KΩ
Ku_2V	$20K_BT$	MTJ-TMR	400%
β	0.1	MTJ area	20×20 nm^2
α	0.01	Memristor	50K~1M Ω
$I_{threshold}$	2μA	ΔV	50 mV
V	0.6V	CMOS tech.	45nm

3.8. Summary

Spintronic threshold device can be combined with CMOS compatible Ag-Si memristors for designing ultra-low energy Spin-Memristor Threshold Logic (SMTL). Such hardware can achieve more than 100× improvement in energy and 1000× improvement in energy-delay product, as compared to state of the art CMOS FPGA based TLG, due to low voltage, low current computing facilitated by a spin-transfer torque device.

4. BRAIN-INSPIRED COMPUTING: HIERARCHICAL TEMPORAL MEMORY BASED ON SPIN-NEURON AND RESISTIVE MEMORY

In this chapter, we present hierarchical temporal memory design based on spin-neuron and resistive memory for energy efficient brain-inspired computing [150]. Hierarchical temporal memory (HTM) tries to mimic the computing in cerebral neocortex. It identifies spatial and temporal patterns in the input for making inferences. This may require large number of computationally expensive tasks like, dot-product evaluations. Nano-devices that can provide direct mapping for such primitives are of great interest. In this chapter, we show that the computing blocks for HTM can be mapped using low power spin based neuron combined with emerging memristive cross-bar array (MCA), and involves comprehensive design at algorithm, architecture, circuit and device levels. Simulation results show possibility of more than 200× lower energy as compared to 45nm CMOS ASIC design.

4.1. Introduction

The human brains are highly efficient in performing cognitive tasks which are thought to involve processing of patterns hidden in different sensory input stimuli, followed by response generation [96][97]. The biological vision system for instance, may incorporate processing of spatial/ temporal patterns, the results of which may be combined with that of the auditory system by the brain, to produce an appropriate physiological response. Several computing models have been explored in literatures [97]-[99] that aim to borrow from the cerebral information processing system, in a quest to realize “intelligent” machines. The earliest efforts involved different mathematical models for artificial neural networks, with varying neuron transfer functions and connection topologies [97]. Deep learning networks (DLN), capable of identifying

patterns under large degree of spatial variations, evolved as a tool for machine learning applications of practical complexity [98][99]. DLNs employ a number of computing levels, with each level processing spatially overlapping region of the inputs, thereby, leading to appreciable tolerance towards spatial modifications of a set of "learned" patterns [98].

Recently, temporal processing was introduced to DLNs as an important new feature. The resulting brain-inspired computing model, called hierarchical temporal memory (HTM), offers the potential of spatial as well as temporal pattern processing, akin to the cerebral neocortex. HTM constitutes of multiple levels of processor arrays. Each processor node "pools" spatial patterns received from the nodes in the lower level of its "perceptive field" and simultaneously identifies the key temporal sequences among those spatial patterns. The pattern identification process may involve computation of conventional distance metrics like, Hamming Distance (HD), Gaussian distance (GD), or dot product (DP) between the stored and the input patterns at each node. A practical HTM hardware may need to store and compute with hundreds of spatial/ temporal patterns at every node. Implementation of such hardware, using the conventional Von-Neumann digital architecture may incur prohibitively high energy and real estate cost [102].

Recent years have seen growing interest in emerging nano-devices that can provide direct and energy efficient mapping of computing primitives required for pattern matching tasks, as in HTM. The pattern matching computations, being inherently variation tolerant, can exploit the "inexact" terminal characteristics of such nano-devices to perform non-Boolean, analog mode operations upon inputs. More importantly, devices that can facilitate direct "in-memory" processing, may be highly attractive for such memory intensive computing. As we described in previous chapter, the memristive cross-bar array can be employed to compute the dot product of multi-dimensional input vector and the stored data. Thus, it can provide a direct mapping of correlation evaluation required in non-Boolean pattern matching applications [72][83][107]. In MCA based pattern matching computation, the direct usage of nano-scale memory for computing leads to high parallelism and elimination of memory read. However, in pattern matching

applications, after the correlation evaluation between the test vectors and stored data, the best match is required to be detected. In previous works, this best match detection requires analog or digital CMOS circuits to process the outputs from MCA [106][107], failing to fully leverage the energy efficiency of MCA based pattern matching application.

In this chapter, we present a STT device structure that can implement ultra-low power current summation and thresholding operation, just like an artificial hard-limiting neuron. Thus we call it ‘spin-neuron’ [83][113][114]. We also present a hybrid Spin-CMOS processing element to detect the best match required in MCA based pattern matching application. Then we propose energy efficient HTM computing blocks based on MCA and the spin-neurons.

4.2. HTM Algorithm and Architecture

In this subsection, the basic computing algorithm and architecture for HTM are described. We focus on the hardware mapping of the inference computing algorithm. The training process is done offline (by software).

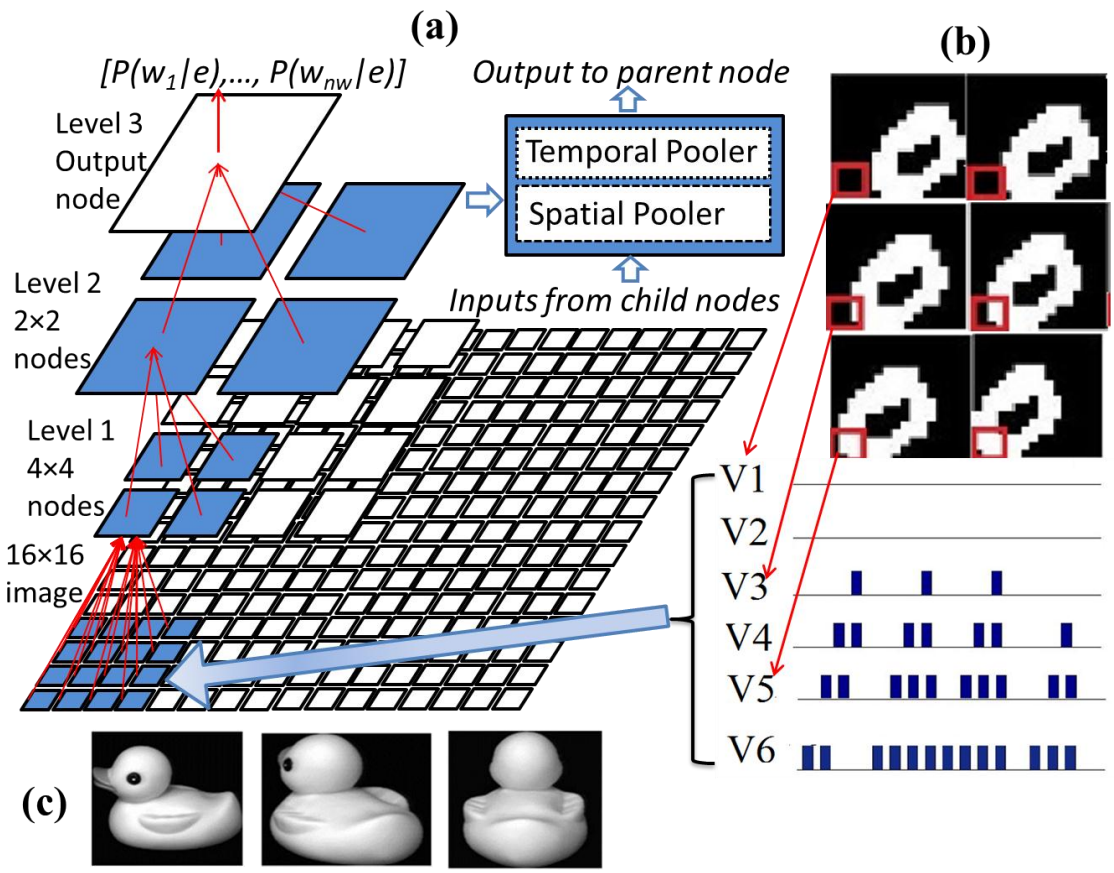


Fig. 4.1 (a) A three-level HTM architecture designed to work with 16×16 pixel images (b) HTM Training Sequence generated by zigzag scan and part of the training sequence of the highlighted lower left node in level 1 (c) snap-shots of a moving duck.

4.2.1. HTM Architecture and Training

HTM computing architecture constitutes of a tree-like network of large number of processing nodes, arranged across multiple levels, having pyramidal connectivity. Each node receives inputs from *N* "child nodes" in its "receptive field" in the immediate lower level. The first level nodes receive inputs from an input stimulus (like, an image). Both forward as well as backward connections between the nodes of non-adjacent levels may also be used, depending upon the training algorithm and the applications [99][100]. In this work, the specific application considered requires only feed-forward flow.

HTM network can work in two phases: training and inference. In this work, we propose the hardware design only for HTM inference phase. Training phase is for the

HTM to learn and memorize the patterns, which mainly involves the extraction of spatial and temporal patterns from the time varying input data, which we assume is done offline. The training proceeds from bottom to top. The parent nodes are trained only after all the child nodes in the lower levels are trained. When the output node finishes training, it is called *fully trained*. All but the top-most (output) level are trained in unsupervised mode [100][103]. The following subsection describes the training process.

4.2.1.1. Spatial Pooling

During the training process, HTM network is exposed to time varying inputs, such as that produced by an object moving smoothly across the network's visual field [101][103]. Fig. 4.1b shows a simple training sequence generated by the moving image of a numeric character, which may be shifting, rotating and scaling (by moving towards or away from the scanner) across the visual field. Training with such time-varying snap-shots of an object can help recognize it with different perspectives using a fully trained network. A more realistic example can be given as that of a moving object, like a duck (taken from COIL-20 data-set [118]), as shown in Fig. 4.1c.

The level-1 nodes (L1-nodes) receive $M \times M$ pixels (M=4 in this work) of the input image, which can be viewed as a 1-D spatial pattern (of length $M \times M$). The L1-nodes detect and store the frequently recurring patterns in their receptive fields. During the training process, each spatial pattern or "*coincidence*"-ci is compared with the present set of patterns for similarity. It is added to the "*spatial pool*" as a new pattern, if it is found to be sufficiently distinct from the existing set. The distinctiveness of a new pattern, with respect to the present set can be determined by placing a threshold on a distance metric, like dot product (DP). This threshold can have a significant impact on the number of spatial/ temporal patterns and the overall training accuracy (will be described later). The probability of occurrence $P(c_i)$ of each spatial pattern is also stored in the form of its count of appearance during the training process.

4.2.1.2. Temporal Pooling

Computation of the temporal patterns for a particular node involves identifying the group spatial patterns c_i*'s* that are likely to occur close in time. A '*temporal group*', g_i, is

a subset of coincidences that possibly originate from simple variations of the same '*class*' of input that is smoothly moving throughout receptive field of the network [101]. Different algorithms can be used to partition the spatial patterns into a set of disjoint temporal groups $G= \{g_1, g_2, .., g_n\}$ [101][103]. In this work, we employ an ad-hoc greedy algorithm for the sake of simplicity [101]. It employs a temporal activation matrix (T_{AC}), where $T_{AC}(i, j)$ denotes the number of times the coincidence c_i was followed by c_j during the training. To start, we pick the element $T_{AC}(i, j)$ in the matrix with the highest value of $P(c_i) \times T_{AC}(i, j)$. This implies selecting c_i as the first element of the first temporal group. The largest non-zero value of $T_{AC}(i, j)$ implies that the coincidence c_j has highest temporal connection with c_i. Hence, c_j is added as the next element to current temporal group g_i. The next element to be added is c_k, where $T_{AC}(j, k)$ has the highest value among the elements in the row $T_{AC}(j, :)$ (j^{th} row). The elements already included in a temporal group g_i are marked as 'assigned' and are not assigned to any other group. This recursive process terminates when the length of one temporal group exceeds the predetermined maximum group size. Thereafter, a new coincidence is selected as the beginner of a new temporal group.

4.2.1.3. Computation of the matrix PCG:

The final step for training a node is the creation of *PCG* matrix, which essentially relates the spatial coincidence c_i*'s* of a node to its temporal groups-g_i*'s*. The element $PCG(i, j)=P(c_i \mid g_j)$ represents the conditional probability of c_i given g_j. The elements of the *PCG* matrix are defined as in equation-4.1 [101].

$$PCG(i, j) = \begin{cases} P(c_i) \text{ if } c_i \in g_j \\ 0 \quad \text{otherwise} \end{cases}, \text{ for each } i = 1...nc,\ j = 1...ng \tag{4.1}$$

where, *nc* and *ng* are the maximum number of spatial patterns (coincidences) and temporal groups respectively. During the inference mode, the *PCG* matrix of a node is used to evaluate the probability distribution over the stored temporal groups, g_i*'s*, in that node, based on its current spatial inputs. Hence, it can be termed as the '*inference matrix*' of a node. The index of the temporal group with the highest probability value constitutes the output information of the node. During the training of a parent node (nodes not connected directly to the input image), all its child nodes (which are already trained),

operate in the inference mode. Their outputs, (which are the indices of the winning temporal groups of the respective nodes, obtained based on current input image) form an effective spatial pattern for the parent node.

4.2.1.4. Training of the output node

As mentioned earlier, the training steps of the output node (the node at the top of the HTM tree) is supervised. The computation of spatial pool (with elements c_i*'s*) is identical to the other levels. The inference matrix, however, is constructed through supervised learning, under a set of specified "desirable" output classes w_i*'s*. The inference matrix of the output node is called *PCW* matrix. The elements of the *PCW* matrix are updated based on the a priori knowledge of the current image class. For example, if the current input image belongs to class w_j, and current coincidence to the output node is identified to be c_i (using DP with all c_i*'s* in the output node), the value of *PCW(i,j)* is incremented by 1.

4.2.2. HTM Inference

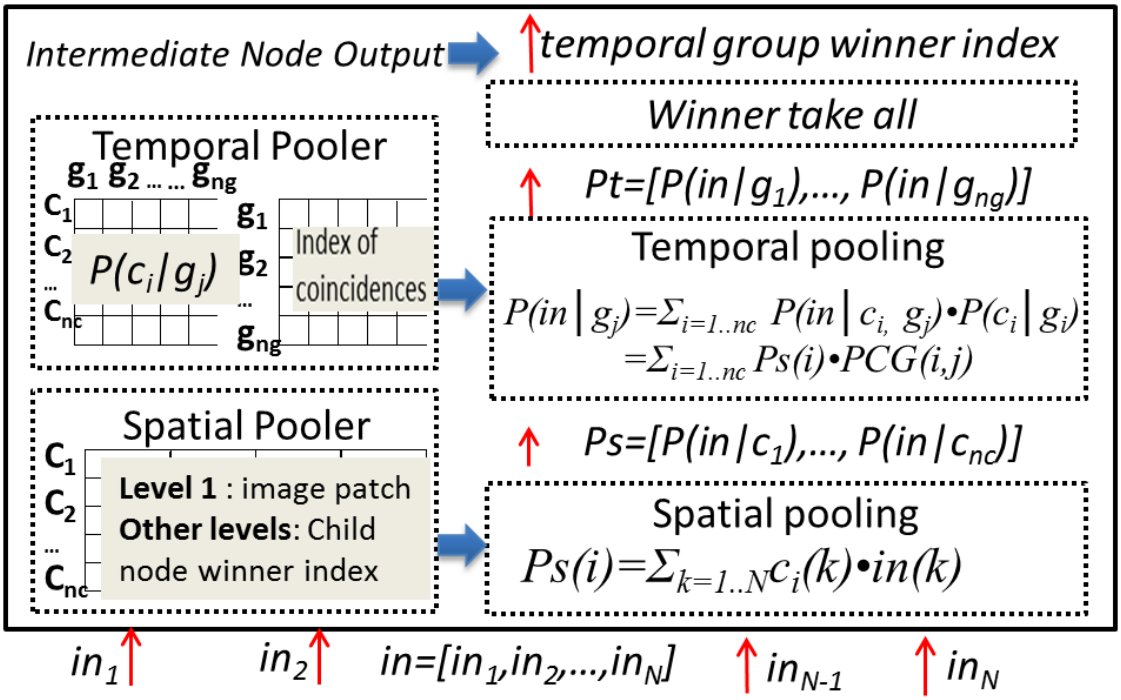


Fig. 4.2 HTM-node structure and the associated inference-steps

Fig. 4.2 shows the node structure and mathematical formulations of the inference steps used in this work [101][103][150]. Inference steps for a node can be divided into the following steps:

4.2.2.1. Composition of spatial input

The spatial input to a node $in=[in_1,in_2,\ldots,in_N]$ is the juxtaposition of the output messages from its N child nodes. As described earlier, for the L1 nodes, the spatial inputs are received directly from the input image (being tested). For the higher level nodes however, the spatial inputs are constituted by the winner indices of the temporal groups of their child nodes.

4.2.2.2. Probability densities over spatial coincidences (spatial pooling)

The vector P_s shown in Fig. 4.2 constitutes of the conditional probability distribution of the input spatial pattern (expressed as '*in*' in the equations) over the stored coincidences: $Ps(i)=P(in|c_i)$, $i=1\cdots nc$. It encodes the spatial similarity between the input pattern (*in*) and the stored spatial coincidences (c_i'*s*). It can be computed as the dot product (DP) between the input and the stored patterns as follows:

$$P_s(i) = \sum_{k=1\ldots N} c_i(k) \bullet in(k) \tag{4.2}$$

Note that, for the output (L3) node, a winner take all (WTA) circuit is needed for this step to detect the "winner" and set the winner output to be 1, while the others to be zero.

4.2.2.3. Probability densities over temporal groups (temporal pooling)

Note that, $P_s(i)$ computed in step-2, denotes the probability distribution of the current input vector over the pooled set of spatial coincidences (c_i'*s*). The vector $PCG(:, j)$ (j^{th} column of *PCG* matrix) on the other hand, denotes the probability of c_i'*s*, "in context" of the particular temporal group, g_j. Hence, the conditional probability of the input given j^{th} temporal group can be computed as follows:

$$\begin{aligned} P(in \mid g_j) &= \sum_{k=1\ldots nc} P(in \mid c_i, g_j) \bullet P(c_i \mid g_j) \\ &= \sum_{k=1\ldots nc} P_s(i) \bullet PCG(i, j) \end{aligned} \tag{4.3}$$

We assume that $P(in|c_i,g_j)=P(in|c_i)$, since g_j and c_i are irrelevant.

4.2.2.4. Computation of output message

The output message of a node is the index of the "winner" temporal group, which is the group with the highest value of $P(in|g_j)$, computed in step-3.

The inference computation of the output node is similar to the other nodes, except for the use of *PCW* matrix, in place of *PCG* matrix.

From the above discussion, we note that the core computing function for the inference mode operation of HTM is the dot product computation. At each node, this function is evaluated twice. At the first step, the operands are the analog vectors corresponding to the input spatial patterns (*in*) and the spatial coincidences stored in the

node. The result, $P_s(i)$, depicts the input dependent probability distribution over the pooled spatial patterns. For the second stage of computation, the input to the dot product function are $P_s(i)$, and, the columns of the *PCG* matrix, corresponding to each of the temporal groups associated with the node. The last step involves determining the index of the "winner" temporal group, which is '*j*' if the second dot product computing (temporal pooling) yields the highest value for $DP(Ps(i), PCG(:,j))$.

Before we move to hardware mapping of the aforementioned HTM computing scheme, we briefly discuss the choice of design specifications for HTM hardware in the following subsection.

4.2.3. HTM Design Specification

In the previous subsections, we introduced the algorithm for training and inferring patterns using HTM, where the main computing process involves DP-evaluation. The algorithm was applied to MNIST [117] data-set for handwritten digits recognition (Fig. 4.3a) and COIL-20 data-set for object recognition [118]. For training, each image was scaled to 16×16 pixels and scanned to generate a sequence of training images, incorporating a sequence of shifts, rotation and scaling of the original image. The character images were taken as binary, whereas, 4-bit resolution was chosen for the grey level COIL-20 images. In this paper, we focus on the HTM inference hardware implementation, whereas the training of HTM is done offline, or in other words, the training is done by software. During software training process, an important parameter is the "*matching threshold*" that determines the addition of a new spatial pattern to a node's memory. The relationship between the numbers of spatial patterns, the numbers of temporal groups in each node and matching threshold are shown in figure 4.3b-c. These plots show that larger threshold and hence, larger number of spatial and temporal patterns ensures higher accuracy. However, this requires increased number of DP-evaluations and hence higher computation cost. In this work, the matching threshold was chosen close to the value for which the computation accuracy saturated to the maximum value of ~95% (corresponding to 0.7). The bit resolution required for the input and the spatial/ temporal memory elements was determined by the maximum variation tolerance for which matching accuracy close to the ideal case (with non-truncated grey scale values for

memory and input) was retained (Fig. 4.3e). During the training phase, appropriate noise models were added to the memory data and the computing function in order to account for the approximate nature of the devices-circuits characteristics used in this work.

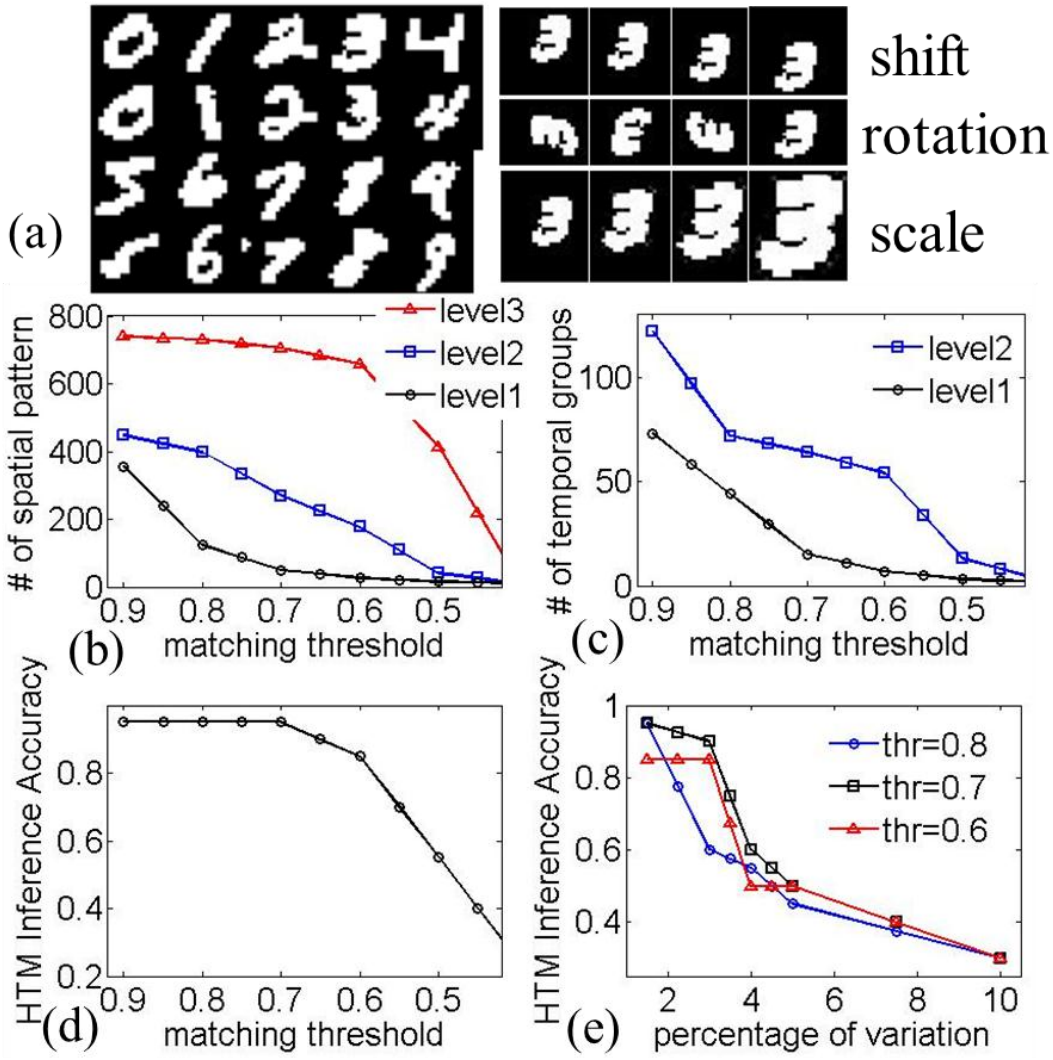


Fig. 4.3 (a) 20 image samples in MNIST benchmark and the shift, rotation and scale variations. (b) Numbers of spatial patterns in each node vs. matching threshold. (c) Numbers of temporal groups in each node vs. matching threshold. (d) HTM inference accuracy vs. matching threshold. (e) HTM inference accuracy vs. percentage-variation in the elements of spatial-temporal memory.

4.3. Computing with memristive cross-bar array

Fig. 4.4 depicts a MCA with two sets of metal bars (horizontal bars and in-plan bars). In such a MCA, memristor with conductance-g_{ij} interconnects i^{th} horizontal metal bar and j^{th} in-plane metal bar. More than 8-bit write accuracy for isolated memristors have been proposed and demonstrated in literatures [69][70]. In a cross-bar array, consisting of large number of memristors, write voltage applied across two cross connected bars for programming the interconnecting memristor also results in sneak current paths through neighboring devices. This disturbs the state of unselected memristors. To overcome the sneak path problem, application of access transistors and diodes have been proposed in literature [90] that facilitate selective and disturb free write operations. Methods for programming memristors without access transistors have also been suggested, but using such techniques, only a single device in an array can be programmed at a time [89][91]. Such schemes can be applicable only if programming speed is not a major concern.

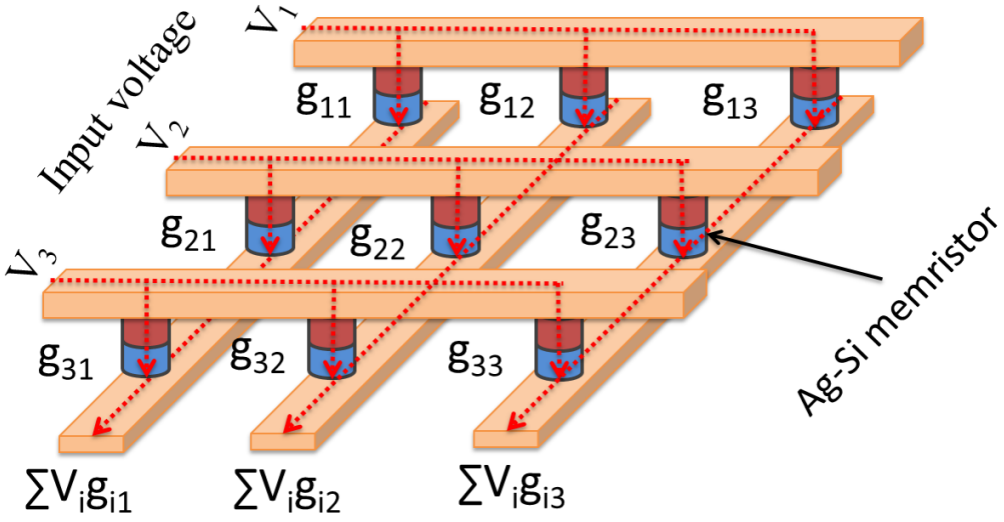


Fig. 4.4 Correlation evaluation between input vector and stored vectors using a memristive cross-bar array

In MCA based pattern matching applications, the input vector can be represented as input voltages applied across the horizontal metal bars as shown in Fig. 4.4, assuming the in-plane metal bars are grounded. j^{th} stored template is mapped to the conductance of memristors connected to j^{th} in-plan bar. During computing, the current flowing through the memristor with conductance- g_{ij} is $V_i \bullet g_{ij}$ and the total current flowing out of j^{th} in-plan metal bar is $\Sigma_i V_i \bullet g_{ij}$. Therefore, this MCA structure can be used to compute the degree of match between one analog vector and the stored templates. The best match of test vector to the templates would be the one corresponding to the highest MCA output current. In order to detect the best match, a winner take all (WTA) circuit is required. In general, the WTA circuits can be categorized as binary tree WTA [111][112] and current conveyor WTA [112]. However, both of these CMOS based WTA design consumes large static power and may be several times larger than the MCA power consumption in the pattern matching applications. Thus, they may fail to fully leverage the energy efficiency of nano-scale resistive memory based computing.

In next subsection, we will describe an ultra-low power STT device structure that can be employed in a spin based WTA circuit design, resulting in energy efficient MCA based computing hardware design.

4.4. Spin-Neuron with Heavy Metal Layer

In this subsection, the spin-neuron with heavy metal layer (will be called spin-neuron for simplicity in this chapter) device structure and operation is described [83][114][150]. We also present the interface circuit design of the spin-neuron to implement an ultra-low power current comparator.

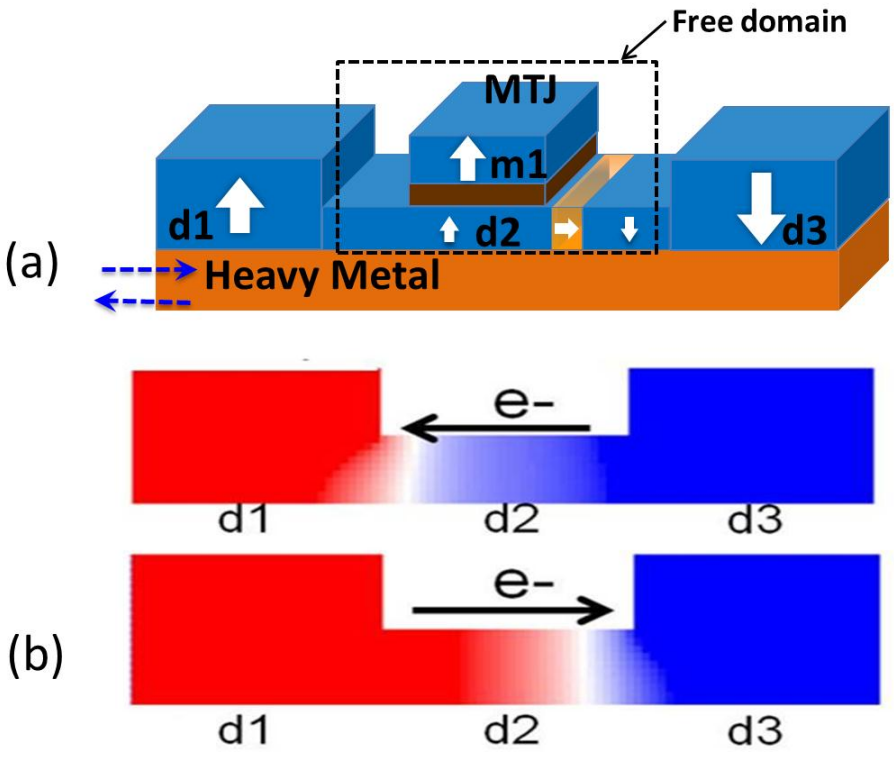


Fig. 4.5 (a) Spin-neuron with heavy metal layer, (b) micro-magnetic simulation of domain wall motion with applied current along spin hall metal layer [114]

Fig. 4.5a shows a three terminal spin-neuron based on magnetic domain wall strip [86]. It has a free magnetic domain d2 which forms an MTJ with a fixed magnet m1 at its top. The magnetization of d2 can be written parallel or anti-parallel to the two fixed spin-domain d1 and d3, depending upon the direction of current flow between d1 and d3. Thus, this device can detect the direction or polarity (positive if going in and negative if going out of its input domain d1) of current flow across its free domain. Hence this device can be used for current-mode thresholding operation [82]. The minimum magnitude of current flow required to flip the state of the free domain d2 depends upon the critical current density for domain wall motion across the free magnetic domain d2. Since the critical current density of domain wall motion is non-zero, a hysteresis in the spin-neuron switching characteristics can be observed as shown in Fig 4.7a. This

hysteresis effect can be reduced by lowering the domain wall motion critical current density to make the switching function closer to a step function.

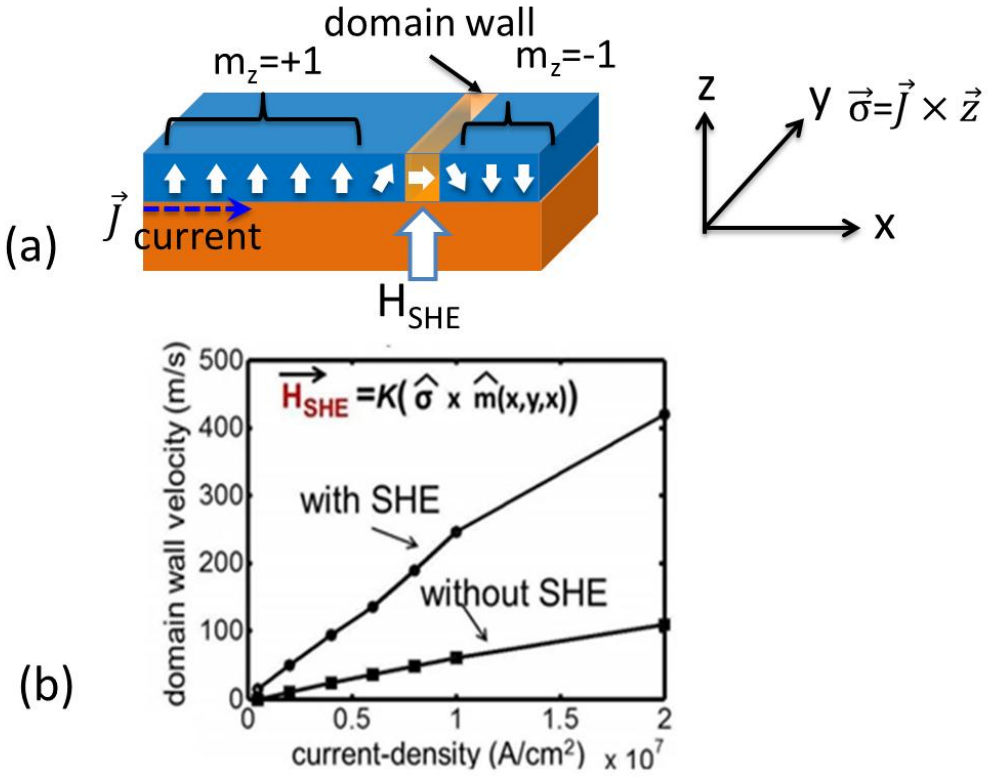


Fig. 4.6 (a) spin orbit torque induces higher domain wall velocity, (b) domain wall velocity vs. applied current density with and without SHE

In order to sense the magnetization of free domain-d2, a fixed magnet-m1 and an MgO layer are placed on top of d2 to form a MTJ. The MTJ resistance is larger when the magnetizations of m1 and d2 are anti-parallel. On the contrary, if the magnetizations of m1 and d2 are parallel, the MTJ resistance is smaller. The dynamic CMOS latch shown in Fig. 4.7b is used to sense the MTJ resistance state.

In the detection latch, the terminal d3 of the spin-neuron is connected to V_{dd}. The current required for the DW motion increases proportional to the switching speed. Since the transient read current flows only for a short duration, it does not disturb the state of

d2. Note that, the transistor mismatch may introduce wrong output of the dynamic latch. The possible solutions can be: 1) Increasing the transistor size. This is a tradeoff between power and device matching. 2) Adding inverter buffer at the latch output terminal. This technique can both isolate load capacitance and minimize the offset errors.

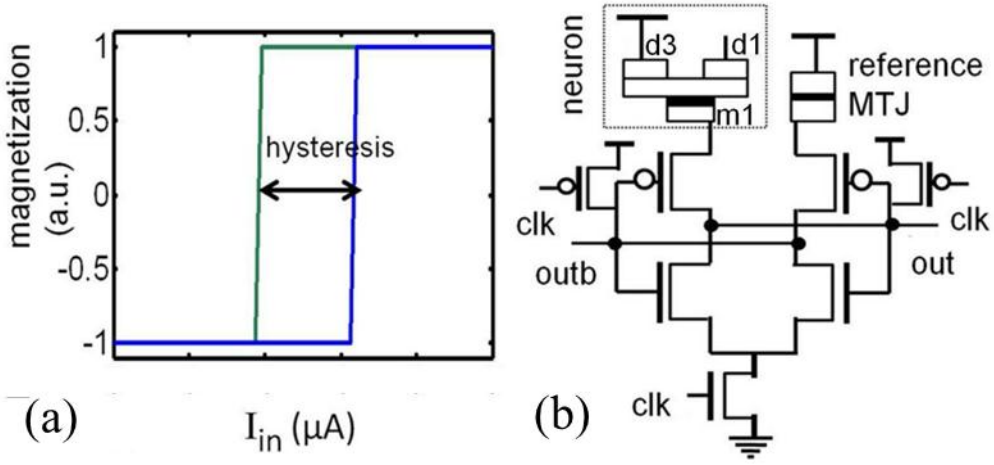


Fig. 4.7 (a) transfer characteristics of the spin-neuron with spin hall metal layer (E_b=20KT), (b) dynamic CMOS latch to sense spin-neuron state

Robustness to read disturb can be further enhanced by the appropriate design choice of m1. Notably, the branch with effective lower resistance draws comparatively higher read current. By setting the polarity of m1 parallel to d1, it can be ensured that for the parallel configuration of the spin-neuron MTJ (and hence, lower resistance) the free layer (d2) is already parallel to d1 and hence a larger transient current does not disturb d2. This technique facilitates lowering of spin-neuron threshold to physical limits of scalability without the concern of read disturb. Apart from device scaling, the spin-neuron threshold can also be lowered by manipulating other device parameters, like the anisotropy energy (E_b) of the magnet [82].

Recently, application of spin-orbital (SO) coupling in the form of Spin Hall Effect (SHE) has been proposed for low-current, high-speed domain-wall motion [47][80][114][116]. For Neel-type DW, SHE induced from an adjacent metal layer

results in an effective magnetic field ($\boldsymbol{H_{SHE}}$) [80], that can be expressed as, $\boldsymbol{H_{SHE}}=K(\boldsymbol{\sigma}\times\boldsymbol{m})$. Here, $\boldsymbol{m}$ denotes the unit magnetization of magnetic domains, $\boldsymbol{\sigma}$ is a current dependent vector defined as $\boldsymbol{\sigma} = \boldsymbol{j}\times\boldsymbol{z}$, where, $\boldsymbol{j}$ is the current vector (which can be positive or negative depending upon direction of current flow) and $\boldsymbol{z}$ is the direction perpendicular to the magnetization plane (along easy axis). As shown in Fig. 4.6a, $\boldsymbol{\sigma}$ can be in-plan or out of plane of the figure, depending upon the direction of the current flow [114]. K is a quantity dependent upon material parameters of the magnet and is proportional to the effective Spin Hall angle, θ_H [80]. Notably, θ_H determines the effectiveness of the Spin Hall interaction, larger θ_H implies larger effective torque due to Spin Hall effect.

For a Neel-type domain wall shown in Fig. 4.6a, the magnetization in the region of the domain wall lies along the length of the magnetic nano-strip [80]. For this configuration, the effective H_{SHE} acting on the domain wall region can be visualized to be perpendicular to the plane of the magnet. The H_{SHE} assists the non-adiabatic spin-transfer torque (which results from the current flow) acting on the domain wall region. For a θ_H of 0.2, micro-magnetic simulations showed an increase of ~5× in the domain wall velocity for a given current density, due to the H_{SHE} term (Fig. 4.6b) [114]. This effect can be used to achieve higher switching speed for a given current, or, to reduce the required switching current for a given switching time for the free domain in the spin-neuron.

In this work switching current threshold of ~2μA for 1 ns switching speed has been chosen for a neuron with SHE-assisted free domain size of $20\times2\times60nm^3$, which corresponds to the current density of $4MA/cm^2$. The state of the free domain can be sensed by injecting a small current across the high resistance MTJ formed between fixed magnet-m1 and free domain-d2.

4.5. Design of HTM Computing Block using Spin-Neuron and MCA

In this subsection, we will present the HTM computing block design composed of spatial pooler, temporal pooler and winner take all (WTA) circuits. Based on equation-4.2 and 4.3, the fundamental operation of spatial pooling and temporal pooling is the dot product between inputs and stored matrixes (spatial/ temporal patterns), where the energy

efficient dot product operation is implemented using the combination of MCA and spin-neuron. The spin-CMOS hybrid processing element based on spin-neuron that achieves analog to digital converter (ADC) and WTA functionality at ultra-low energy will also be introduced.

4.5.1. Spatial and Temporal Pooler Design

Each HTM block consists of two *'pattern matching'* networks using dot product, corresponding to the spatial pooling (density over coincidences) and temporal pooling (density over temporal groups). The node data structure and mathematical equations can be seen in Fig. 4.2. The dot product functionality can be implemented by MCA described in previous section and the spin-CMOS hybrid process element (spin-neuron based SAR-ADC) is used to detect the output.

4.5.1.1. Dot product operation circuit

As described in previous subsection, the dimension of each MCA based dot product computing block is ($n\_child \times nc$, $nc \times ng$), where *n_child* is the number of child nodes, *nc* is the number of spatial patterns stored in current node and *ng* is the number of temporal groups. The input vectors to first MCA (spatial pooling) are respectively the real image pixels for level-1 nodes and the child node temporal group winner indices for the other level nodes. The input vectors to the second MCA (temporal pooling) are the outputs of the first MCA. As shown in Fig. 4.3e, ~4% parameter variation can be tolerated based on our choice of matching threshold during training. Thus, the bit-length of the *PCG* matrix (and of spatial pooler) was chosen to be 5.

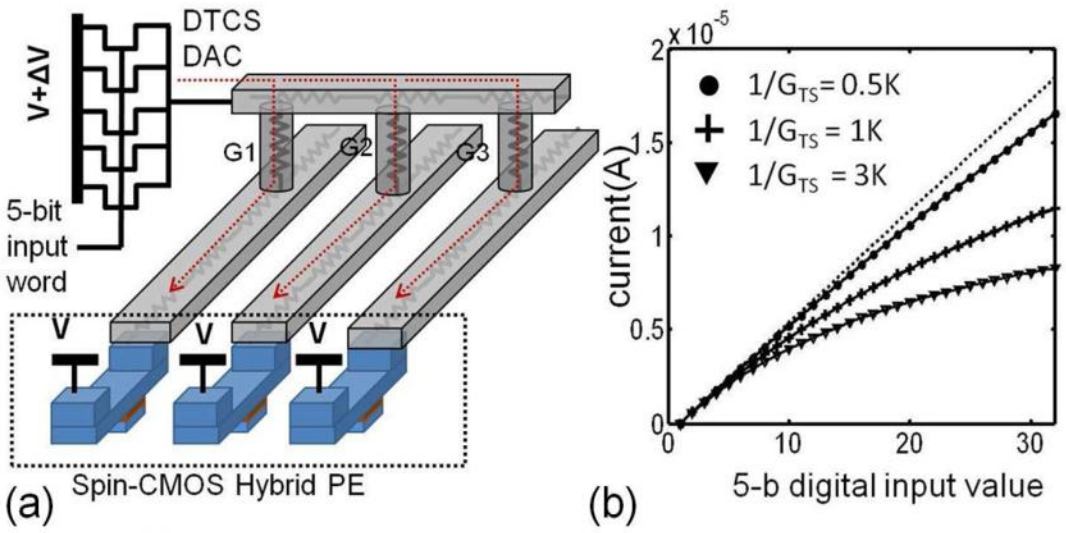


Fig. 4.8 (a) DTCS DAC provides inputs to MCA, while spin-CMOS hybrid PE takes the MCA outputs (heavy metal layer is not shown for simplicity) (b) DTCS DAC non-linearity with different G_{TS} [83]

Fig. 4.8a shows the architecture of dot product computing block required in HTM spatial and temporal pooling operations. It consists of deep triode current source (DTCS) based digital to analog converter (DAC), MCA and the spin-CMOS hybrid PE. Since the test vectors from the image or the HTM child node are digital values, a DAC is required to convert digital test vector to analog voltage or current. In this design, we employ a DTCS based DAC design as shown in Fig. 4.8a. The binary weighted PMOS transistors are working in deep triode region by applying a voltage of V+ΔV to the source terminal and a voltage of V to the spin-neuron. ΔV can be ~50mV to ensure the transistors working in deep triode region. As shown in Fig. 4.9a, the DTCS transistor shows near linear drain current to gate voltage.

An alternative low power DAC for the input digital data can be a compact switched MOS capacitor DAC (Fig. 4.9b). This analog voltage can be used to drive the DTCS transistors that supply current to the MCA for computation. Analog mode driving can achieve lower data bus width, thereby reducing the power consumption due to dynamic switching of the data bus. As described in previous subsection, the output current of each column is the dot product of the input voltages (currents) and the programmed

conductance of the memristors. The analog output currents will be converted into digital values using the proposed spin neuron based SAR-ADC (will be described later).

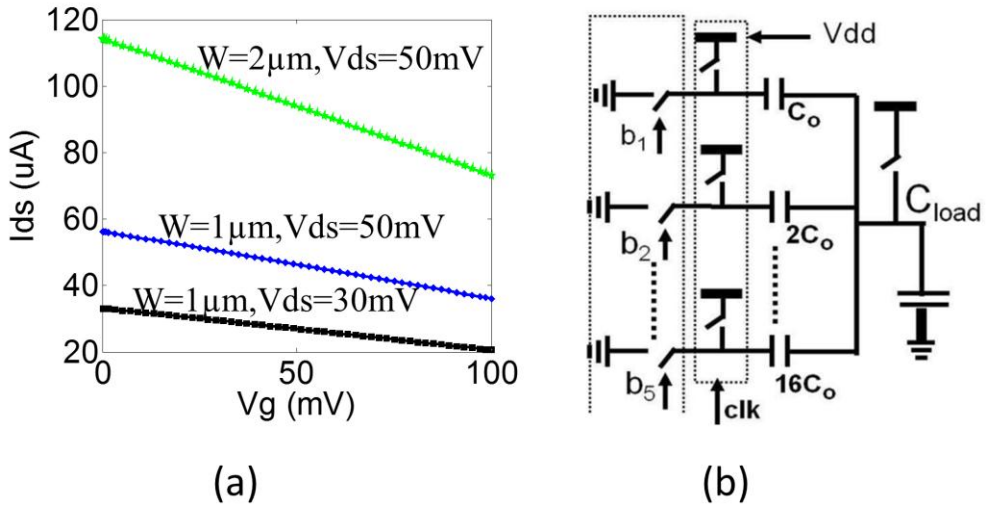


Fig. 4.9 (a) Near-linear drain-current (I_{ds}) vs. gate voltage (Vg) with different Vdd and ΔV (b) compact switched capacitor DAC scheme [83]

Assuming the parasitic resistance of the metal bars can be ignored, the i^{th} DTCS-DAC current output can be expressed as follows:

$$I_{DAC}(i) = \Delta V \bullet G_{DAC}(i) \bullet G_{TS} / \left(G_{DAC}(i) + G_{TS}\right) \tag{4.4}$$

where $G_{DAC}(i)$ is the conductance of i^{th} DTCS-DAC depending on the digital input, G_{TS} is the total conductance of memristors connecting to the same horizontal bar. Note that, dummy memristors are added such that G_{TS} is equal for all horizontal bars. Then the current flowing through memristor connecting i^{th} horizontal bar and j^{th} in-plan bar can be written as:

$$\begin{aligned} I(i,j) &= \Delta V \bullet [G_{DAC}(i) \bullet G_{TS} / \left(G_{DAC}(i) + G_{TS}\right)] \bullet \left(g_{ij} / G_{TS}\right) \\ &= \Delta V \bullet G_{DAC}(i) \bullet g_{ij} / \left(G_{DAC}(i) + G_{TS}\right) \end{aligned} \tag{4.5}$$

where g_{ij} is the memristor conductance connecting i^{th} horizontal bar and j^{th} in-plan bar. If $G_{TS} >> G_{DAC}(i)$, the above equation can be approximately written as:

$$I(i,j) = \Delta V \bullet G_{DAC}(i) \bullet g_{ij} / G_{TS} \tag{4.6}$$

It can be seen that the current flowing through each memristor is proportional to the product of $G_{DAC}(i)$ and g_{ij}. Therefore, the current flowing out of j^{th} MCA in-plan bar is the dot product of input vector and the stored template, which can be expressed as:

$$I_{MCA}(j) = \frac{\Delta V}{G_{TS}} \sum_{i=1}^{N} G_{DAC}(i) \bullet g_{ij} \tag{4.7}$$

where, N is the dimension of test vector and stored templates. Note that, lower value of G_{TS} reduces the linearity of the DTCS-DAC characteristics as shown in Fig. 4.8b, so does the HTM accuracy. We add normal distributed device variation to the simulation of HTM node (including process variations on DTCS-DAC based on the model in [110] and memristors model [70]). DAC Integral Non-Linearity (INL) and Differential Non-Linearity (DNL) will degrade with the consideration of process variation, thereby reducing the detection margin (difference between the best match and second best match, shown in Fig 4.10) of MCA outputs (i.e. HTM accuracy). The HTM accuracy vs the percentage variations on DTCS-DAC and memristor can be seen in Fig. 4.3e. Moreover, in case access transistors are employed for improved writablity, the minimum conductance is determined by the 'ON' resistance of the transistors (which is ~1K Ω for a minimum sized 45nm device).

4.5.1.2. Spin-Neuron based SAR ADC Design

The second step of spatial and temporal pooler is the detection of MCA outputs (dot product) and converting them to digital values. Fig.4.10 shows the normalized MCA outputs (ADC inputs) of one HTM level-2 node, for 20 different image samples. It shows the worst case difference between the best and second best matches to be ~4%, at the moment of comparison, which indicates at least a 5 bit resolution ADC circuit is needed to detect the best match.

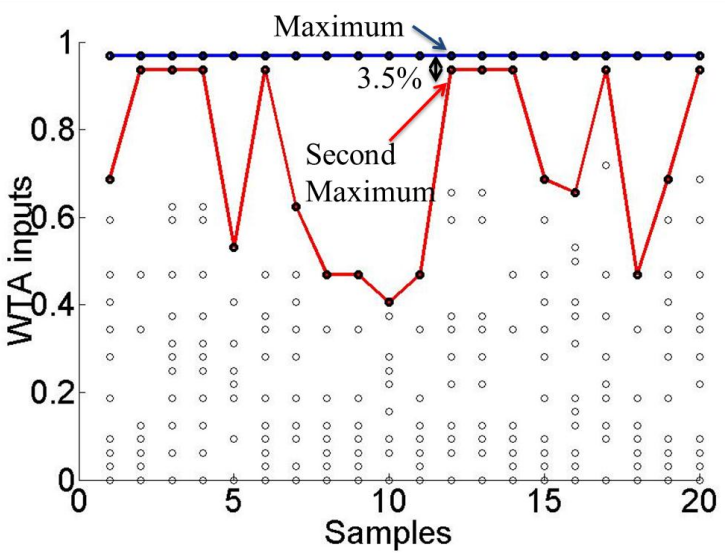


Fig. 4.10 the normalized MCA column outputs (WTA inputs) for different image samples, showing isolation between the best and second best match.

The standard algorithm of successive approximation register (SAR) ADC can be described as follows. Initially, the digital value stored in the approximate register is set in mid-scale. For example, in a 5-bit SAR-ADC, the initial state of approximate register is '10000'. Then a DAC is used to convert the digital value stored in the approximate register to analog value, comparing to the analog input. If the analog input is higher, the MSB of approximate register remains high. If the analog input is lower, the MSB of approximate register changes to low and the next lower bit should be changed to high. The same process is repeated until all of the bits are compared. In the end, the digital value stored in the approximate register is the digitized value of the analog input.

The spin-neuron we described in the previous subsection is used as the ultra-low power current comparator in the SAR-ADC design as shown in Fig. 4.11 [83][130]. At each conversion cycle, the DTCS-DAC converts the digital value stored in the approximation register to an analog current, comparing to the MCA output current by the spin-neuron. The output state of the spin-neuron determines the SAR logic as we described in the SAR-ADC algorithm. Note that, the drain terminal voltage of the DTCS-DAC is V-ΔV and the other node of spin-neuron is powered at voltage-V.

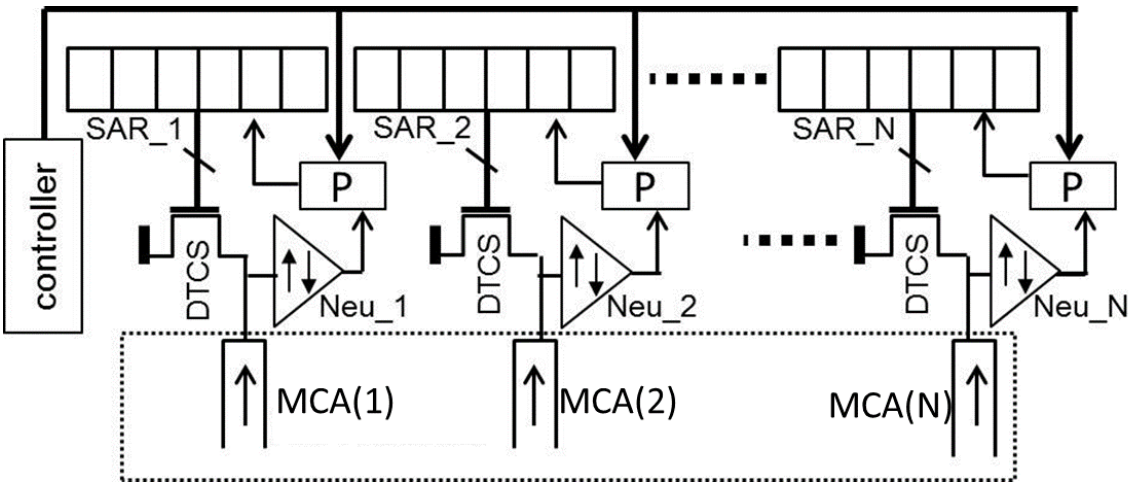


Fig. 4.11 Spin-neuron based SAR ADC circuit diagram [83]

The digital outputs of SAR-ADC are the outputs of the spatial pooler and temporal pooler, which indicate the input densities over spatial patterns and temporal groups (shown in Fig. 4.2). Spatial pooler's outputs are sent to temporal pooler, and temporal pooler's outputs are sent to winner take all circuits described in next subsection.

4.5.1.3. Winner take all circuit design

As described in previous subsection, the output of each node in HTM is the winner index of the temporal group for the non-output node or the winner index of the class for the output node. A WTA circuit is required in each HTM node as shown in Fig. 4.2. The spin-WTA circuits employed in this chapter are shown in Fig. 4.12 [83].

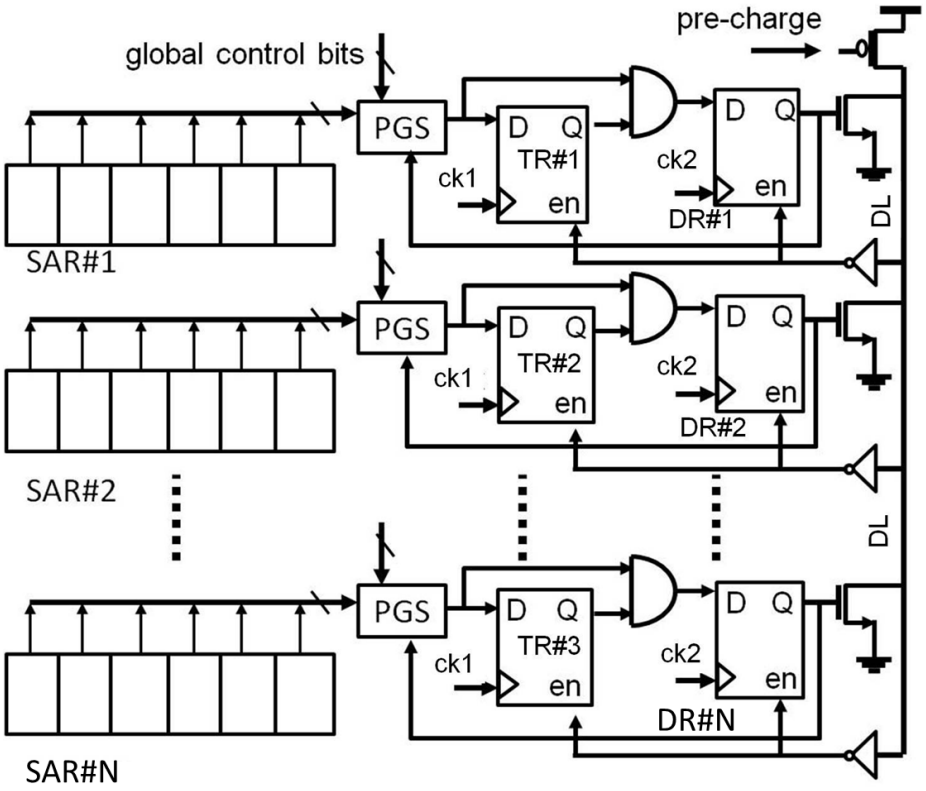


Fig. 4.12 WTA circuit diagram [83]

The WTA algorithm operates in parallel with the ADC operation. It can be explained with the help of the corresponding circuit diagram shown in Fig. 4.12. Results of the first ADC conversion step obtained from the SAR are directly transferred to the tracking registers (TR) shown in the figure through the pass-gate multiplexing switch (PGS). Thus, at this stage, all the TR's with a high output correspond to the ADC results with MSB = '1'. Let us now, consider the second cycle operation. The detection line (DL) is first pre-charged to *Vdd* and the set of discharge registers (DR) driving it are cleared to low output. Next, if for at least one of the SAR's with high MSB, the second MSB also evaluates to '1', the corresponding DR is driven high by the associated AND gate. Thus, DL is discharged to ground and the write of all the TR's is enabled. All the TR's for which both, first and second MSB's evaluated to '1', stay high, but the rest are set to low. In simple terms, if at least one of the SAR's (5-bit) evaluated to '11000' in the second

conversion cycle, the DL is discharged and all the TR's with SAR value '11000' stay high, while those with SAR value '10000' are set to low. In case all SAR's evaluated to '10000' in the second cycle, no change is made to the TR values. Thus, at the end of conversion cycle, if only one of the TR's remains high, it is identified as the winner and the corresponding SAR value is effectively the density over temporal group.

The winner tracking circuitry described above is fully digital. Moreover, owing to the global digital control, it is easily scalable with number of input as well as required bit precision.

4.5.2. HTM Hardware Mapping Using Spin-MCA Based Pattern Matching Network Architecture

We introduced the design of MCA based dot product computing network, spin-neuron based SAR-ADC and WTA in the previous subsections. The architecture of proposed HTM system is shown in Fig. 4.13. The level-1 nodes take the corresponding image patch as the inputs, the first MCA computes the density over spatial patterns (spatial pooling), the spin-ADC converts the current outputs into digital values and sends to the second MCA that computes the density over the temporal groups (temporal pooling). The WTA circuit detects the winner and sends the winner index to its parent node.

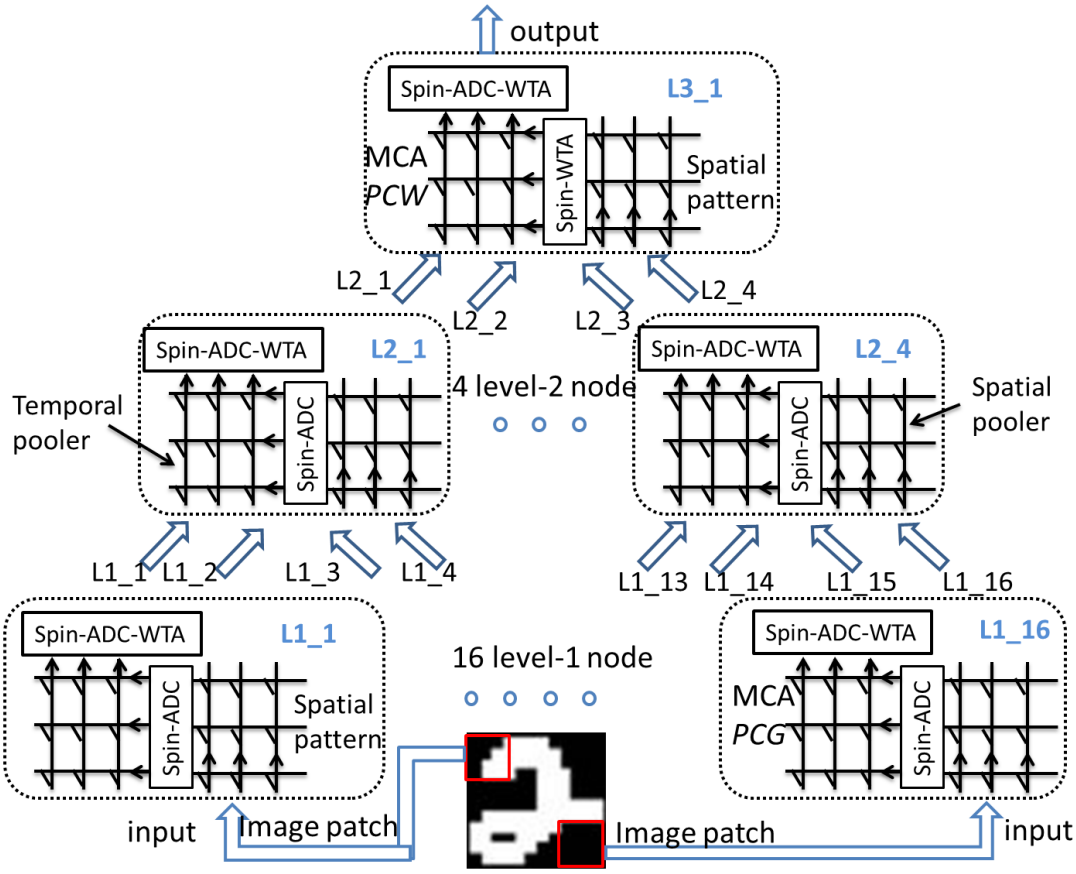


Fig. 4.13 HTM hardware mapping using spin-MCA based pattern matching network architecture

4.6. Performance of Proposed HTM Hardware

As we described earlier, the HTM computing blocks are mainly based on dot product of test vectors and stored templates, which can be implemented using digital CMOS adders and multipliers. As a comparison to our proposed HTM hardware design, we simulated the CMOS digital adders and multipliers based HTM node in IBM 45nm technology. The energy consumption of CMOS and spin based HTM level-2 node are shown in Fig. 4.15. It can be seen that the spin based HTM node design results in a much lower energy consumption (~200× lower) compared with CMOS based design. Such huge energy saving mainly comes from two reasons: 1) In our spin based HTM node design, the voltage across the MCA is drastically reduced to ΔV (~50mV) due to the low voltage, low current requirement of our spin-neuron based processing element (i.e. ADC

and WTA). 2) The fully digital WTA used in this work is a compact and ultra-low power design, compared with relative high power consumption mixed-signal CMOS WTA.

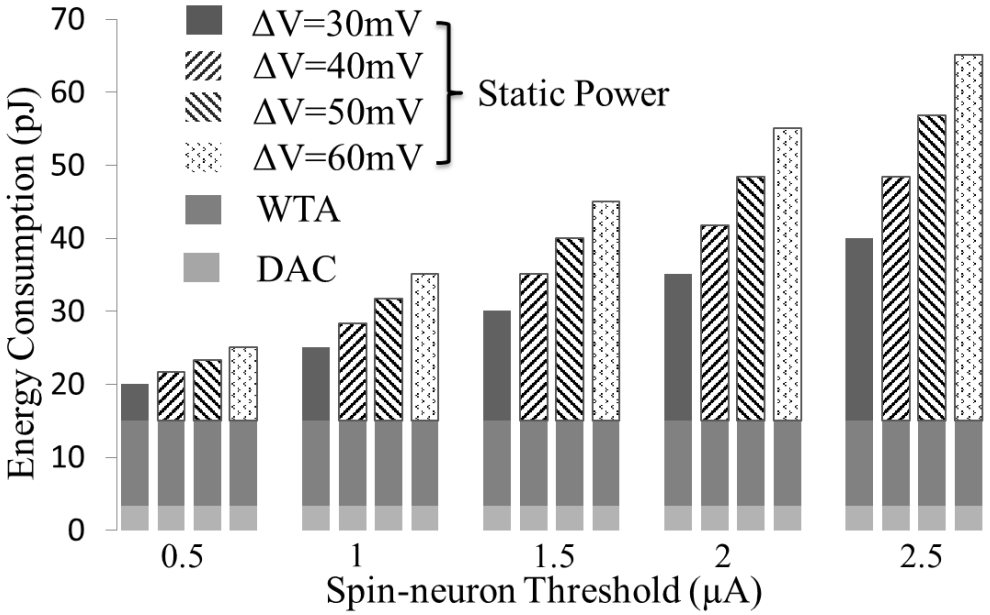


Fig. 4.14 Energy consumption of a single HTM node (level 2) for different values of spin-neuron threshold and ΔV

As shown in Fig 4.3, for appreciable matching accuracy, the average number of spatial (c_i's) as well as temporal (g_i's) groups in the HTM nodes can be more than hundred (for the given application and tree structure). As an example, for most second level nodes, the size of the *PCG* matrix was found to be ~270×64. This would imply DP evaluation between 64 pairs of analog vectors, each of length 270. Here, 270 denotes the length of $P_s(i)$ and that of the *PCG* columns (*PCG(:, j)*, each corresponding to a particular temporal group g_j). The bit-length of the *PCG* matrix (and of spatial pooler) was chosen to be 5 (based on the analysis presented in Fig. 4.3). This calls for more than ~10kB of memory read per cycle of a node's computation. (If a fully parallel design is chosen for the node, it would require, storing of the same amount of data in dedicated registers). CACTI simulations [119] predict more than ~1nJ of energy dissipation, even if zero

leakage digital spin-memory is used. The digital data corresponding to the *PCG* elements needs to be converted into analog voltage (current) levels, before it is subtracted from the analog mode results for $P_s(i)$. This energy was estimated to be ~70pJ for approximate switch capacitor based DACs [83].

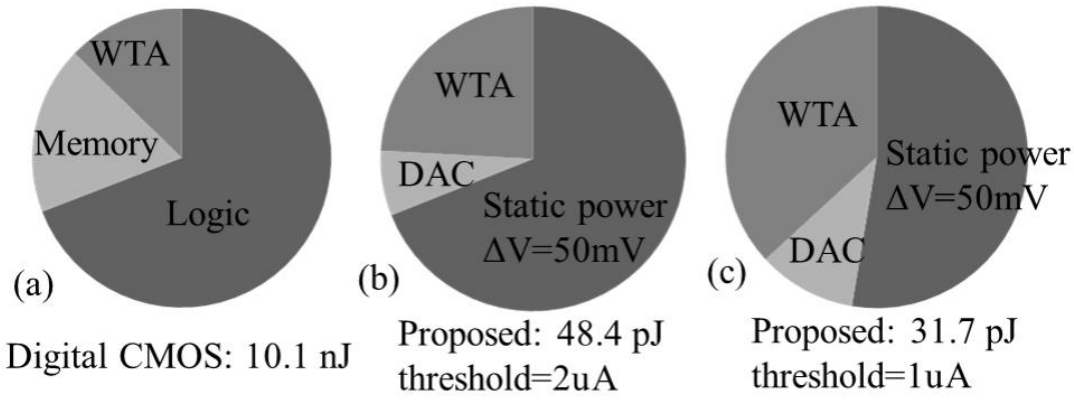


Fig. 4.15 Distribution of energy dissipation for a single HTM node design (level 2 node) (a) fully digital CMOS design, (b) Spin-MCA based design with 2 µA spin-neuron threshold, (c) Spin-MCA based design with 1 µA spin-neuron threshold ('WTA' in the pie chart includes both the ADC and WTA circuit)

Let us now consider the energy dissipation of the proposed computing core of HTM. Based on our simulation, the energy dissipation for the spin-neuron is the dominant part due to the negligible digital WTA static power. The energy dissipation for the spin-neuron has two components. The first is switching energy due to the static current flow between the input voltages and the neuron. This component equals to the product of the total input current flowing across the MCA output columns, the input voltage levels, and the neuron switching time. For an average of 50µA of current flow across input voltage levels of 50mV for 1 ns switching time, this component evaluates to 2.5fJ. The noise considerations in the state of the art on-chip supply distribution schemes may limit the minimum input voltage levels that can be used. Even for 100mV of input levels, the first energy component is limited to 5 fJ. The second component of energy dissipation in the spin-neuron can be ascribed to the spin-neuron read operation. For a supply voltage of

0.8V, this would evaluate to 0.48 fJ. Thus, the total energy-dissipation in a spin-neuron for 1 ns switching speed can be around 3fJ.

Fig. 4.14 shows the energy consumption of a single HTM level-2 node design. It can be seen that the static power consumption mainly depends on the spin-neuron switching threshold and the ΔV across the MCA. However, the dynamic power (Flip-Flops and DAC) is almost constant for different spin-neuron threshold currents and ΔV. With the reducing of the spin-neuron threshold current, the dynamic power starts to dominate. In this work, the spin-neuron threshold current is 2 μA. Lower value of ΔV would imply more energy savings. We have assumed that regulated precision DC levels with ~1mV accuracy are available [120]. The minimum usable ΔV is limited by the precision regulation of DC supply achievable. For the given application, the required bit-precision for the spatial/temporal memory was found to be 5 bit. Hence, even a 1mV noise would mandate a minimum ΔV of ~30mV. We choose ΔV as 50mV in this work to obtain better variation tolerance. With current spin-neuron threshold and ΔV configurations, Fig. 4.15 shows the energy dissipation of the proposed design is around 48pJ for a single HTM level-2 node design. It implies an energy benefit of more than 200× over a digital CMOS design. As mentioned earlier, IBM 45nm technology was used to evaluate the CMOS design energy consumption.

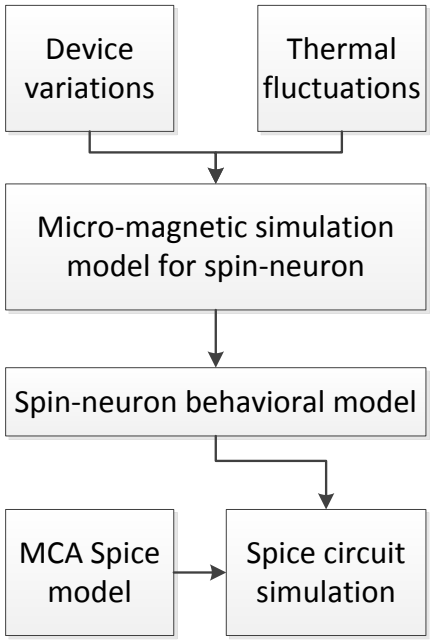


Fig. 4.16 simulation framework used in this work

The simulation framework used in this chapter is shown in Fig. 4.16. The device variation and thermal fluctuations are included in modeling the spin-neuron by a self-consistent simulation framework presented in [89]. The spin-neuron was calibrated with experimental data on domain wall magnets. The addition of device variation and thermal fluctuations in the spin-neuron model creates a variation on spin-neuron threshold current, which will degrade the accuracy of spin-neuron based SAR-ADC. According to our simulation, the effect of spin-neuron variations can be neglectable compared with memristor conductance variations in a 5-bit spin-neuron based SAR-ADC. The HTM node is simulated in SPICE based on a statistical behavioral spin-neuron model. Some important design parameters used are listed in table 4.1.

Table. 4.1 HTM Design Parameters.

WTA resolution	5 bit	Magnet material	NiFe
Input date rate	100 MHz	Free layer size	$20\times2\times60nm^3$
Cross-bar parasitic	1Ω/μm 0.4fF/μm	Ms	800 emu/cm^3
Cross-bar material	Cu	Ku_2V	20KT
Memristor material	Ag-Si	Ic	2μA

4.7. Summary

The low voltage, magneto metallic 'spin-neurons' combined with MCA are explored in the dot product based pattern matching, which is the core computing block in the design of HTM hardware. Such a direct mapping of the core-computing primitive of the cortical computing system can be very attractive for large-scale and energy-efficient design. The simulated spin based HTM computing block results in ~200× lower energy consumption compared to the CMOS based HTM node design.

In this chapter, we focused on the HTM inference hardware implementation, whereas the training of HTM is done offline, or in other words, the training is done by software. In the future, online training of HTM can be explored. We employed dot product based pattern matching as the core computing primitive of HTM. As another extension of this work, other pattern matching scheme, such as Hamming distance or Gaussian distance, can also be implemented using the spin-transfer torque devices.

5. SPIN-TRANSFER TORQUE BASED SOFT-LIMITING NON-LINEAR NEURON

In the previous chapter, we discussed a spin-neuron design which can implement energy efficient current mode thresholding operation. In this chapter, we present a spin-transfer torque (STT) device based on Domain Wall Motion (DWM) magnetic strip that can efficiently implement a Soft-limiting Non-linear Neuron (SNN) operating at ultra-low supply voltage and current [131]. In contrast to previous spin-neurons that can only realize hard-limiting transfer functions (thresholding function), the proposed STT-SNN displays a continuous resistance change with varying input current, and can therefore be employed to implement a soft-limiting neuron transfer function. Soft-limiting neurons are greatly preferred to hard-limiting ones due to their much improved modeling capacity, which leads to higher network accuracy and lower network complexity. We also present an ANN hardware design employing the proposed STT-SNNs and Memristive Cross-bar Arrays (MCA) as synapses. The ultra-low voltage operation of the magneto metallic STT-SNN enables the programmable MCA-synapses, computing analog domain weighted summation of input voltages, to also operate at ultra-low voltage. We modeled the STT-SNN using micro-magnetic simulation and evaluated them using an ANN for character recognition. Comparisons with analog and digital CMOS neurons show that STT-SNNs can achieve more than two orders of magnitude lower energy consumption.

5.1. Introduction

Neural network based computing models have been explored in recent years for realizing hardware that can perform “human-like” cognitive computing [97]-[100], [121]-[123]. The fundamental computing units of such systems are the *neurons* that connect to each other and to external stimuli through programmable connections called *synapses* [97][121]. The basic operation performed by an artificial neuron is computing a weighted

sum of the N inputs and passing the result through a non-linear transfer function, expressed as follows:

$$Y = \varphi(\sum W_i \bullet IN_i - \theta) \tag{5.1}$$

where, Y is the neuron output or *activation level*, IN_i denotes the i^{th} input, W_i is the corresponding synapse weight, θ is the neuron threshold or bias and φ is the neuron *transfer (activation) function*. Fig. 5.1b shows four representative neuron transfer functions. The step function is called *hard-limiting* transfer function because of the binary output states. The saturated linear, logistic sigmoid and hyperbolic tangent functions are *soft-limiting* transfer functions because of the continuous neuron output states [97][121]. Large numbers of neurons can be connected in different network topologies to realize different neural network architectures [98][100][122][123]. For instance, cellular neural networks employ near neighbor connectivity [122], whereas, fully connected feed-forward networks employ all-to-all connections between neurons in consecutive network layers or stages [123]. Several other network paradigms like Convolutional Neural Networks (CNN) [98], and Hierarchical Temporal Memory (HTM) [100][150] provides structured approaches to design large-scale networks. Irrespective of the network topology, neurons connect to each other in effect to communicate their probabilities (neuron activation levels) of being part of the final output [121]. The binary neuron output levels seriously hamper the possibility of neuron-to-neuron communication [121]. Soft-limiting neuron transfer functions are therefore preferred and greatly improve the neural network modeling capability while reducing network complexity. The reason behind this can be intuitively understood as follows. With hard-limiting functions, each neuron is required to decide whether it will be turned completely “on” or completely “off”, which requires a step-like function. On the other hand, with soft-limiting functions, each neuron can be in any of a continuous range of activation levels between ‘0’ and ‘1’, allowing much more information to be communicated across neurons. Various functions that meet these requirements have been explored as artificial neuron transfer functions [97][121][142]. The optimal neuron transfer function is highly dependent on the dataset and network topology. In this work, we do not attempt to implement the optimal neuron transfer function, but rather propose an energy efficient spin-transfer torque based device

that can implement a continuous non-linear function. This function can be used as a soft-limiting artificial neuron transfer function.

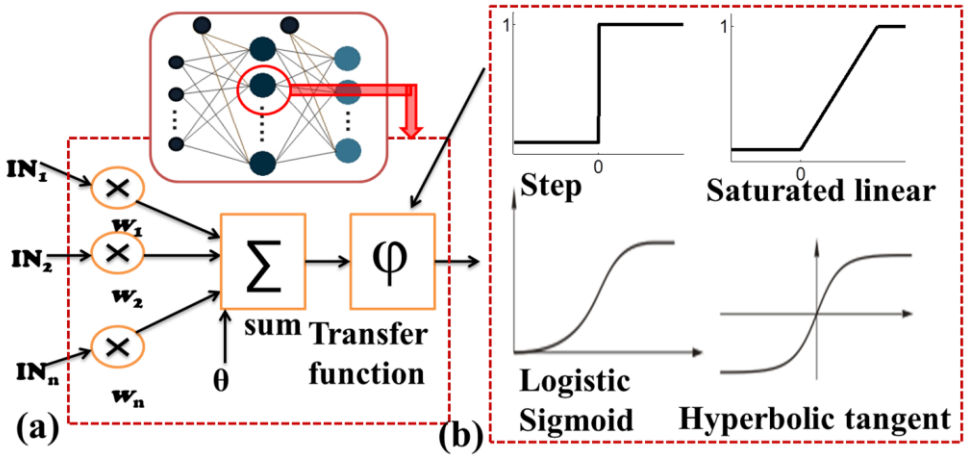


Fig. 5.1 (a) artificial neuron: it takes weighted sum of n inputs and passes the result through an transfer/activation function (b) four representative transfer (activation) functions

The energy efficiency, performance, and integration density of ANN hardware is governed by the design of the fundamental computing units that realize neurons and synapses. In previous works [124][125], the artificial neurons and synapses are implemented using CMOS circuits, which in general requires large numbers of transistors and high power consumption. Therefore, it is of great interest to use post-CMOS devices to realize the ANN algorithmic models into powerful cognitive computing hardware in an energy efficient manner. Recent experiments [77][109][126][128] have shown that nano-magnets can be switched at reasonable speed with small current density using a mechanism called spin-transfer torque (STT). Such STT based magneto-metallic devices can be used to implement current mode summation and non-linear operation, mimicking an artificial neuron in an energy efficient manner. We previously proposed the application of spin-neurons based on domain wall motion (DWM) magnet for designing ultra-low power neural networks [114][129][130]. However, all of the previously proposed spin-neurons implement the hard-limiting step-function, which leads to larger

network size, and simply cannot provide adequate modeling accuracy for complex classification problems.

In this chapter, we present a Spin-Transfer Torque based Soft-limiting Non-linear Neuron (STT-SNN) having an output which is a rational function of the total incoming synapse currents, leading to compact network size and ultra-low power consumption. Instead of binary output states, our proposed STT-SNN can have continuous output voltages. We also present an ANN hardware design employing deep-triode current source (DTCS) transistors as interfacing circuits and memristor cross-bar arrays (MCA) as synapses. The fact that STT-SNNs operate at ultra-low voltages enables the programmable MCA synapses, computing analog domain weighted summation of input voltages, to also operate at ultra-low voltage for low overall energy consumption. Compared with state of the art digital and analog CMOS neurons, the proposed STT-SNN can achieve around two orders of magnitude lower energy.

5.2. Proposed Spin-Transfer Torque based Soft-limiting Non-linear Neuron

In this subsection, we describe the device structure and operation of the proposed soft-limiting neuron. The CMOS circuits employed to interface to the neuron are also discussed.

The proposed Spin-Transfer Torque based Soft-limiting Non-linear Neuron (STT-SNN) is based on a composite device structure consisting of a DWM magnetic strip and a magnetic tunnel junction (MTJ) as shown in Fig. 5.2a. The MTJ consists of two ferromagnetic layers with an MgO barrier sandwiched between them. The “free” ferromagnetic layer (d4) connects laterally to two anti-parallel fixed domains - d1 and d2 [128][139]. The larger thickness at the edges of the free layer is used to stabilize the DW at an intermediate position within the free layer [128]. In general, the application of current induced domain wall motion faces the problem of stable control of domain walls. It comes from many reasons, such as DW structural change, bidirectional displacements, thermal effect of Joule heating, stochastic nature of DWM and the local pinning effect [145]-[149]. The reduction of critical current density to de-pin DW from a pinning site can largely solve those problems. A small DWM critical current density in the range of $10^{11}A/m^2$ was demonstrated experimentally in a scaled magnetic nano-strip with

Perpendicular Magnetic Anisotropy (PMA) [126]. The reason why PMA device has a smaller DWM critical current density compared with In-plane Magnetic Anisotropy (IMA) device can be explained as follows. In the magnetic nano-strip, when the current is injected through a fixed domain, it becomes spin-polarized and exerts a torque on the DW. This torque induces the rotation of magnetization to the hard-axis direction, resulting in the pinning force. If the current density is above a certain threshold, the spin-transfer torque can overcome this pinning force, leading to steady domain wall motion. Thus, the critical current density can be lowered by increasing the STT (narrower domain wall) or decreasing the pinning force (lower hard-axis anisotropy). In summary, the critical current density-$j_{th} \propto K_{h.a.} L_{DW}$, where $K_{h.a.}$ is hard-axis anisotropy and L_{DW} is the domain wall length [145]-[149]. The hard-axis anisotropy of a PMA device reduces with lower device thickness and becomes much smaller than that of IMA device. Moreover, the DW length in a PMA device is in general smaller than that in an IMA device. Therefore, a scaled PMA magnetic nano-strip is used in our work to achieve lower critical current density to induce steady DWM. The free layer dimensions are $2\times20\times100nm^3$ as shown in Fig. 5.2a. A *Neel* type DW is formed because of the small strip width (20nm) [126]. The DW length $L_{DW}=\pi\sqrt{(A_{ex}/Ku)}$= ~17nm based on our device parameters listed in table-5.1.

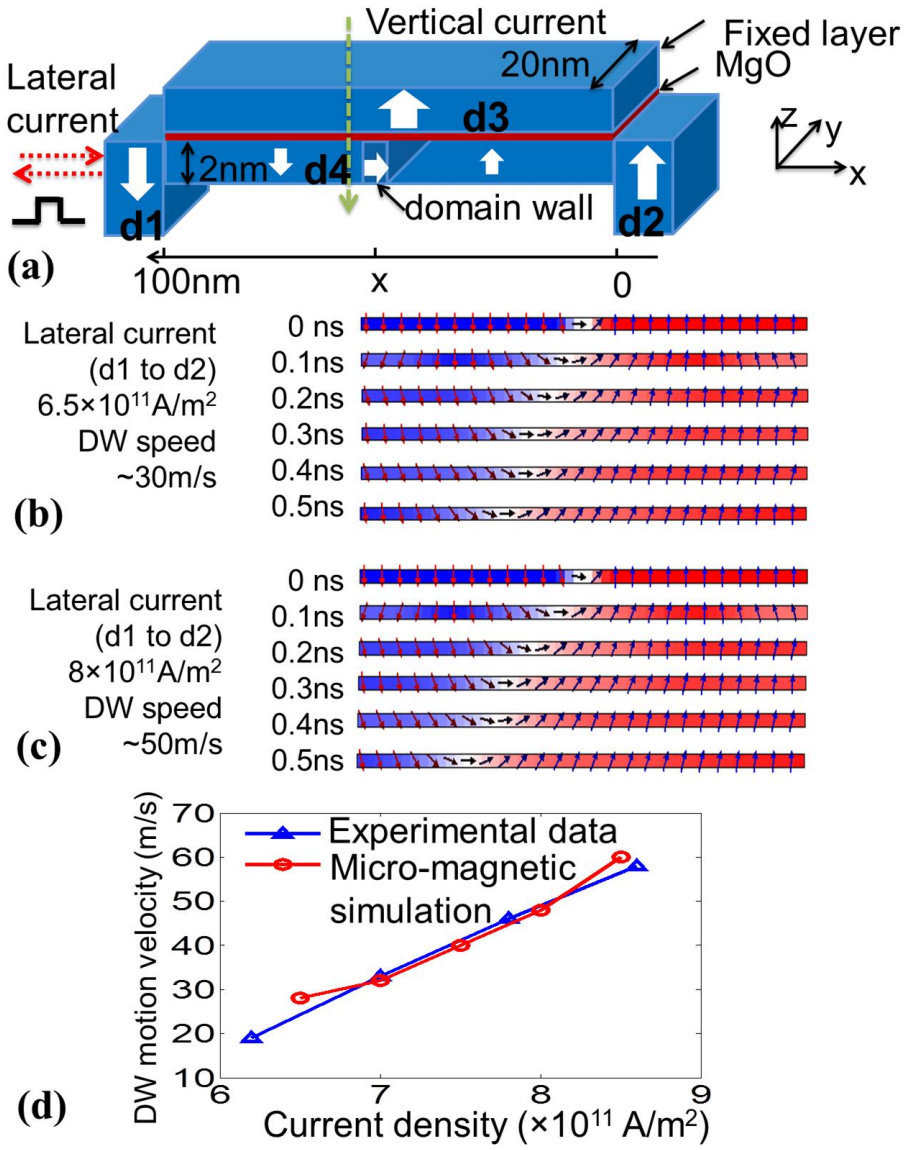


Fig. 5.2 (a) The proposed STT-SNN device structure, (b) the micro-magnetic simulation of free layer DW motion when the injected lateral current density is 6.5×1011 A/m2 and (c) 8×1011 A/m2 , (d) simulated DW motion velocity vs. current density, showing a good match with experimental data reported in [126]

The proposed STT-SNN device can be treated as a four terminal device with lateral and vertical current paths. For the lateral path (d1 to d2, ± x direction), d1 forms the input programming port, assuming d2 is supplied with a constant voltage. The domain wall can be moved along the free layer depending on the lateral current pulse magnitude,

direction and duration [77][109][126], leading to a continuous resistance change of the MTJ in the vertical direction. The transient micro-magnetic simulation plot of the free layer using *mumax*[3] [135] is shown in Fig. 5.2b&c, where a 0.5ns current pulse with magnitude of 6.5×10^{11}A/m^2 and 8×10^{11}A/m^2 are applied from d1 to d2. It can be seen that the domain wall moves to the left (along the direction of electron flow) with a different speed. The device parameters used in the simulation are listed in table-5.1. We benchmarked the micro-magnetic simulation with the experimental data in [126] (the same nano-strip width of 20nm is fabricated in the reference) and it shows a good match as shown in Fig. 5.2d. A relatively high *Ku* (i.e. high energy barrier) is preferred in the memory application for the sake of good thermal stability [126]. In the computing applications, a lower energy barrier can be used to reduce the critical current density to de-pin the DW, which leads to lower energy consumption.

Table. 5.1 STT-SNN Device Parameters used in Simulation

Symbol	Quantity	Values
α	damping coefficient	0.02
Ku	uniaxial anisotropy constant	3.5×10^5 J/m^3
Ms	saturation magnetization	6.8×10^5 A/m
A_{ex}	exchange stiffness	1.1×10^{-11} J/m
P	polarization	0.6

The vertical path (from d3 to d4, $\pm$ z direction) is used for sensing the position of DW in terms of MTJ vertical resistance. MTJ resistance is a function of voltage, tunneling oxide thickness (t_{ox}) and the angle between free layer and pinned layer magnetizations. The atomistic level simulation framework based on Non-Equilibrium Green's Function (NEGF) formalism [137] can be used to evaluate the MTJ resistance, which includes the device variation and thermal fluctuation. The system functionality in this work is simulated in SPICE using a statistical behavioral model. In this model, the

STT-SNN is simulated as three parallel MTJs with variable resistance depending on DW positions (Fig. 5.4a):

$$R_L = RA_{AP} / \left(W \bullet \left(L - x - 0.5L_{DW}\right)\right) \tag{5.2}$$

$$R_R = RA_P / \left(W \bullet \left(x - 0.5L_{DW}\right)\right) \tag{5.3}$$

$$R_{DW} = RA_{DW} / \left(W \bullet L_{DW}\right) \tag{5.4}$$

where, R_L, R_{DW} and R_R are respectively the vertical resistance of left anti-parallel, domain wall and right parallel equivalent MTJ resistances; x is DW position (middle point), L is the length of free layer (100nm), W is the width of free layer, RA_{AP}, RA_{DW} and RA_P are respectively MTJ resistance-area product for anti-parallel, DW and parallel configurations. The resistance of the STT-SNN can then be computed as:

$$R_{neuron} = R_L // R_{DW} // R_R = \frac{A}{Bx + C} \tag{5.5}$$

$$A = RA_{AP} \bullet RA_P \bullet RA_{DW} \tag{5.6}$$

$$B = (RA_{AP} - RA_P) RA_{DW} \bullet W \tag{5.7}$$

$$C = RA_P \bullet RA_{DW} \bullet W \bullet L + (RA_{AP} \bullet RA_P - 0.5RA_P \bullet RA_{DW} - 0.5RA_{AP} \bullet RA_{DW}) W \bullet L_{DW} \tag{5.8}$$

where, R_{neuron} *is* the vertical resistance of STT-SNN. A, B and C are constants depending on the MTJ resistance area product and device dimensions as shown in equation-5.6-equation-5.8. Note, this model is used for SPICE simulation in sensing the neuron state. DW position (x) is a function of total input currents, modeled using micro-magnetic simulation as described earlier.

The interface circuit of STT-SNN is shown in Fig. 5.3a. It works in three phases – programming, sensing and reset phase. In the programming phase, the lateral programming current (total synapse current) programs DW position along the free layer. Then, for the sensing phase, a voltage divider circuit is used to sense the STT-SNN state. The reference MTJ voltage is treated as neuron output voltage which will be transmitted through '*axon*' to its fan-out neurons (axon circuit will be explained in next subsection). For maximum power efficiency and the isolation of two paths, different phases should be separately powered. The clocked power supplies called *pClocks* can be used (as shown in

Fig. 5.3b). When in the programming and the reset phases, *PclkB*+ and *PclkB*- are in *floating state*, while *PclkA* provides a constant voltage *V* to d2, enabling the lateral programming path. When it is in the sensing phase, *PclkA* and the input terminal (d1) are in the floating state. Meanwhile, *PclkB*+ and *PclkB*- supply 50mV and -50mV, respectively (choice of sensing voltage will be explained later). The clocked power supply is implemented using widely used power gating technique [138]. Finally, a reset current pulse (-50μA, 1ns) is applied to the STT-SNN free layer to set the DW location in the rightmost corner, ready for the next computation cycle.

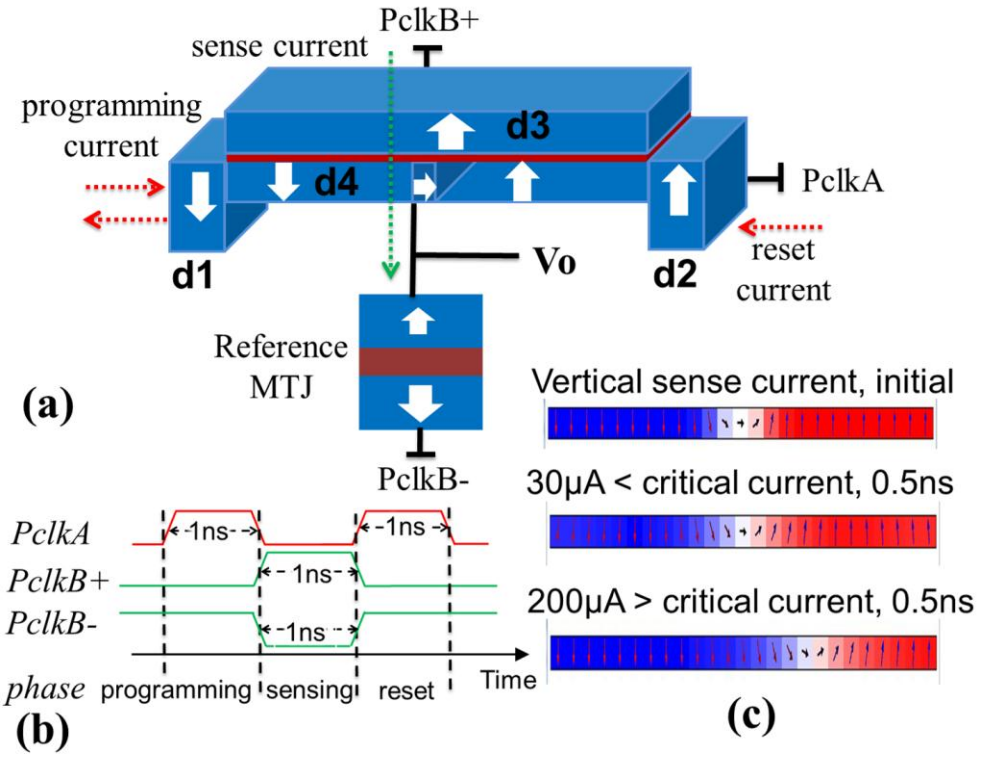


Fig. 5.3 (a) The programming and sensing circuit of the proposed STT-SNN, (b) the clocked power supply waveforms, (c) the micro-magnetic simulation of STT-SNN free layer with different vertical sense currents.

The authors in [128] have experimentally shown that the vertical current may also shift DW when the current density is above a critical value because of the out-of-plane (i.e. field-like) torque. DW position displacement is what we want to avoid in sensing the

STT-SNN resistance. Note, the DW position essentially indicates the state of the neuron. Based on the micro-magnetic simulation for vertical current injection, the vertical critical current density to de-pin the DW was found to be ~5×10^{10}A/m^2 [128], corresponding to a critical current of ~100μA. Thus, based on our simulation, the largest allowed voltage difference between *PclkB*+ and *PclkB*- is ~350mV. In order to keep a good amount of sensing margin, *PclkB*+ and *PclkB*- are set to be 50mV and -50mV, respectively, which corresponds to a maximum of 30μA vertical sensing current. From the micro-magnetic simulation shown in Fig. 5.3c, DW position is stable when the vertical sensing current is 30μA.

Based on the compact STT-SNN model, the output voltage in Fig. 5.3a) can be computed as:

$$V_0 = V_s \frac{R_{ref}}{R_{ref} + R_{neuron}} = V_s\left(1 - \frac{A}{R_{ref}Bx + R_{ref}C + A}\right) \tag{5.9}$$

where, V_s is the voltage difference between *PclkB*+ and *PclkB*- (100mV), R_{ref} is the reference MTJ resistance, x is the domain wall location, A, B, C are the constants expressed as equations-5.6, equations-5.6 and equation-5.8. It can be observed that the output voltage is a *rational function* of DW positions ($0<x<100$nm). Note, rational function is defined as the ratio of two polynomials (two linear functions with the same slope in our case). 'x' is a function of the total lateral programming current as described earlier. Fig. 5.4b shows the STT-SNN resistance vs. DW position. It can be seen that the STT-SNN resistance can be adjusted in a continuous range of values based on the DW position, enabling continuous output voltages as shown in Fig. 5.4c. Based on the micro-magnetic simulation of DW motion velocity dependence on the injected current density shown in Fig. 5.2d, the neuron output voltage vs. programming current (assuming 1ns clock cycle) is plotted in Fig. 5.4d. The positive current direction is defined as from 'd1' to 'd2' as shown in Fig. 5.3a. Note that, the programming current here is the total synapse current (weighted sum of inputs in ANN model). If the programming current is smaller than the DW depinning critical current (*th1)*, DW is stable at the initial position and the output voltage is minimum. When the programming current is larger than '*th2*', DW will

be pushed to the other end and the output voltage saturates to the maximum. '*th2*' can be defined as the minimum current to push the domain wall from one end to the other end using 1ns clock cycle. This two threshold currents (*th1* and *th2*) can be tuned by proper device dimensions and material parameters to adapt different ANN designs.

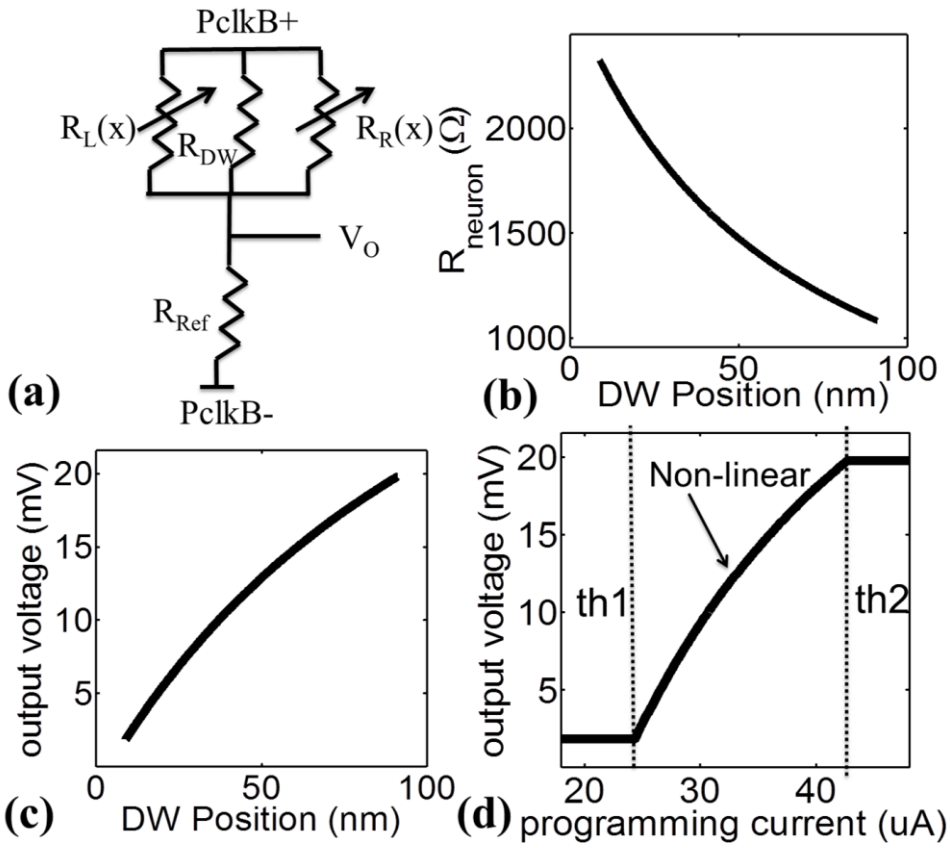


Fig. 5.4 (a) Behavioral STT-SNN SPICE model, (b) STT-SNN resistance vs. DW positions, (c) output voltage vs. DW positions, (d) output voltage vs. programming current. Note, the positive current direction is defined from d1 to d2. Clock cycle is 1ns.

From the above discussions it is clear that the proposed device can be used to implement the low current, soft-limiting non-linear function of an artificial neuron. Next, we will show that the weighted summation of inputs can be efficiently implemented by MCA-synapse.

5.3. Memristive Cross-bar Array Synapses

The two-terminal synapse bears striking resemblance to memristor whose conductance can be precisely modulated by charge or flux through it [140]. In the ANN model shown in Fig. 5.1a, the inputs go through the associated synapses (multiplied by weights) and are summed up as input to the neuron transfer function. This operation can be implemented efficiently using a memristive cross-bar array (MCA) shown in Fig. 5.5 [67][83]. In an MCA, the memristor (e.g. Ag-Si) with conductivity g_{ij} interconnects the i^{th} horizontal metal bar and j^{th} in-plane metal bar. If the outward ends of in-plan bars are grounded and input voltages V_i are applied to horizontal bars, the current going through the interconnected memristor is $V_i \bullet g_{ij}$. Thus, the total current coming out of the j^{th} in-plan metal bar equals to the dot product of the inputs V_i and the associated memristor conductance g_{ij}, namely $\Sigma_i V_i \bullet g_{ij}$. In ANN, the memristors can be employed to store the synapse weights in terms of conductance and the MCA can be used to evaluate the weighted summation of the inputs.

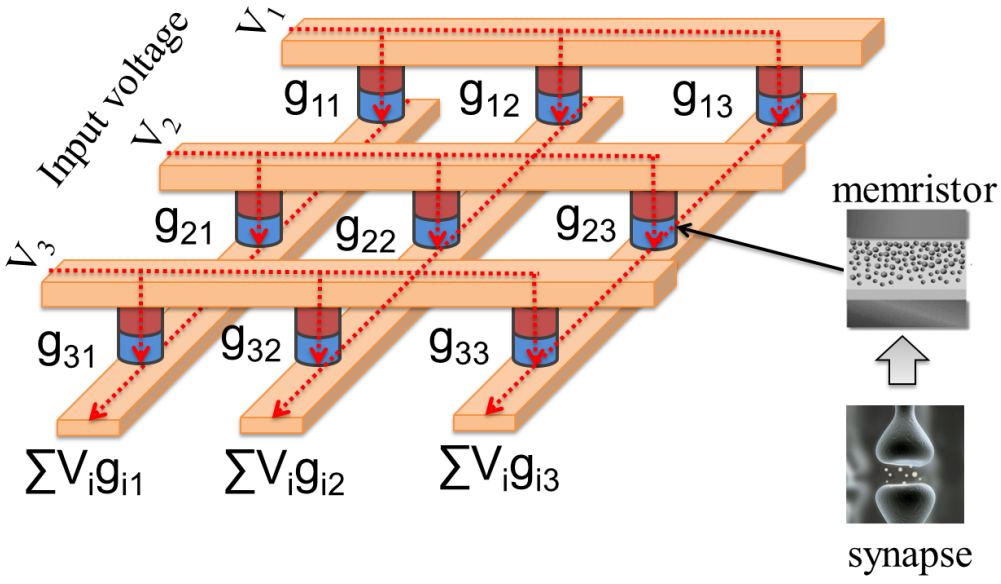


Fig. 5.5 (a) Memristor crossbar array used for evaluating the weighted sum of inputs for ANN

More than 8-bit write accuracy for isolated memristors was proposed and demonstrated in [104]. In our work, 5 bit accuracy was used for demonstrating system functionality. Note that, lower synapse weight resolution can be used by increasing the number of neurons. It is a trade-off between the resolution of the weights and the number of neurons. Even binary weight configuration can be used, however, it would require much more number of neurons. In a cross-bar array consisting of large number of memristors, write voltage applied across two cross-connected bars for programming the interconnecting memristor can result in *sneak current paths* through neighboring devices [90]. This disturbs the state of unselected memristors. To overcome the sneak path problem, application of access transistors and diodes have been proposed in literature [90], which facilitates selective and disturb free write operations. A multi-bit memristor array-level programming scheme employing adjustable pulse width is described in previous chapter and shown in Fig. 3.2 [130]. In this scheme, when programming one specific memristor cell in the array, the corresponding set of the word line, source line and bit line will be selected. During the writing operation, a constant current will be injected into the selected cell and the voltage developed on the source line is compared with a comparator threshold. A digital to analog converter (DAC) is used to set the threshold proportional to the target resistance. As soon as the accessed memristor is programmed to the target value, the current source is disconnected [130]. More precise tuning of memristor value can be achieved by applying a lower value of write current resulting in slower ramp in the resistance value. The memristive devices (including Ag-Si) do exhibit a finite write threshold for an applied current/voltage, below which there is negligible change in resistance [92]. Since the application of spin based neuron facilitates ultra-low voltage (and hence low current) operation of the memristors for computing, the state of memristor in the MCA will not be disturbed during read operations.

5.4. ANN Hardware Using STT-SNN and MCA

In this subsection, we describe our proposed ultra-low power ANN hardware design combing MCA synapses and STT-SNN, showing one to one similarity to biological neural network.

In a biological neural network, '*axons*' are used to transmit electrical-chemical signal between neurons [97][121]. In our proposed ANN hardware (Fig. 5.6), a deep triode current source (DTCS) transistor is used to act as an 'axon' interconnecting the previous stage neuron output (voltage) with MCA synapses. As shown in Fig. 5.7a, the drain to source voltage of DTCS transistor is of the order of few tens of millivolts and it operates in the '*deep-triode*' region where the drain current I_{ds} is linearly proportional to $Vdd-V_T-V_g$, where V_T is the threshold voltage and V_g is the gate voltage. Moreover, the maximum I_{ds} can be tuned by the width of the transistor and V_{ds} as shown in Fig. 5.7a. Therefore, DTCS transistor can be used to transmit the neuron output voltage into synapse current similar to axon [129]. Fig. 5.6 shows the spin-CMOS hybrid ANN (one layer) hardware design using DTCS-axon, MCA-synapses and STT-SNN, which shows one to one similarity to biological neural network. The i^{th} input to the MCA synapses may connect to the j^{th} STT-SNN with either positive, negative or zero weight. This is achieved by programming either g_{ij+} or g_{ij-} to the corresponding weight. For zero weight (i.e. no connectivity), both g_{ij+} and g_{ij-} are driven to high resistance "*off*" state. The input signal to MCA synapses is received through DTCS transistors with source terminals connected to a potential $V+\Delta V$ (for positive weights) and to $V-\Delta V$ (for negative weights), where ΔV can be ~50mV. Ignoring the parasitic resistance of metal cross-bar (for small scale network size), the current going through one synapse can thus be written as $I_{in}(i)\bullet g_{ij}/g_{TR}$, where $I_{in}(i)$ is the current supplied by the i^{th} DTCS transistor, g_{ij} is the synapse weight dependent conductance of the i^{th} input to the j^{th} neuron and g_{TR} is the total conductance of all the memristors connected to the same horizontal bar. Note that, dummy memristors are added such that g_{TR} is equal for all horizontal bars. Thus, the current coming out of each MCA in-plane bar is the total current going into the connected STT-SNN, and can be expressed as $\Sigma I_{in}(i)\bullet(g_{ij+} - g_{ij-})/g_{TR}$, where $I_{in}(i)$ is linearly proportional to the input voltage. The total synapse current determines the STT-SNN output voltage according to the soft-limiting non-linear transfer function shown in Fig. 5.4d.

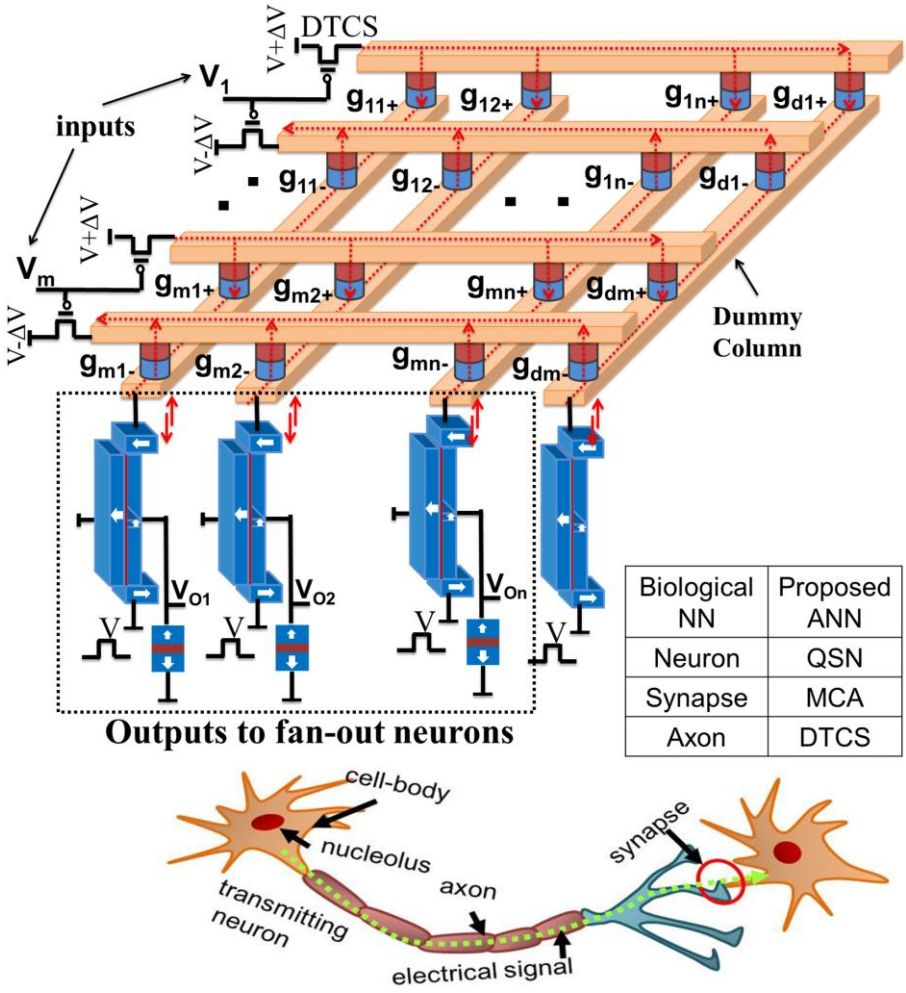


Fig. 5.6 The proposed ANN hardware design using DTCS-axon, MCA-synapse, and STT-SNN

The linearity and source-to-drain current range of DTCS transistor is affected by the fluctuation in drain voltage. As shown in Fig. 5.7b, the non-linearity of DTCS currents can be reduced by using lower range of values for the memristor resistances, hence higher g_{TR}. The other design parameters like the synapse weight resolution, neuron transfer function thresholds etc., are determined by the MCA model [92] and neural network training to ensure the implemented ANN accuracy. The required output current range of DTCS transistor is determined based on the network size, weight resolution of synapses, g_{TR} and neuron threshold. As shown in Fig. 5.7a, the combination of V_{ds} (ΔV) and transistor sizing can tune the DTCS output current range. For a required amount of

DTCS current, the power consumption of MCA is proportional to the voltage across the crossbar (ΔV). Thus, it is desirable to reduce ΔV as much as possible. The minimum ΔV is determined mainly by the non-linearity of DTCS that degrades the output neuron *detection margin* (difference between the highest output to the second highest output) and hence, the matching accuracy. For the benchmark we will describe in the next section, ΔV of 50mV (with regulated DC supply of 1mV prevision [144]) is the minimum voltage to maintain the same matching accuracy as ideal case. Therefore, the MCA-synapses are biased across a small terminal voltage ΔV (between $V+\Delta V$ *and* V), leading to ultra-low power consumption of weighted summation of inputs.

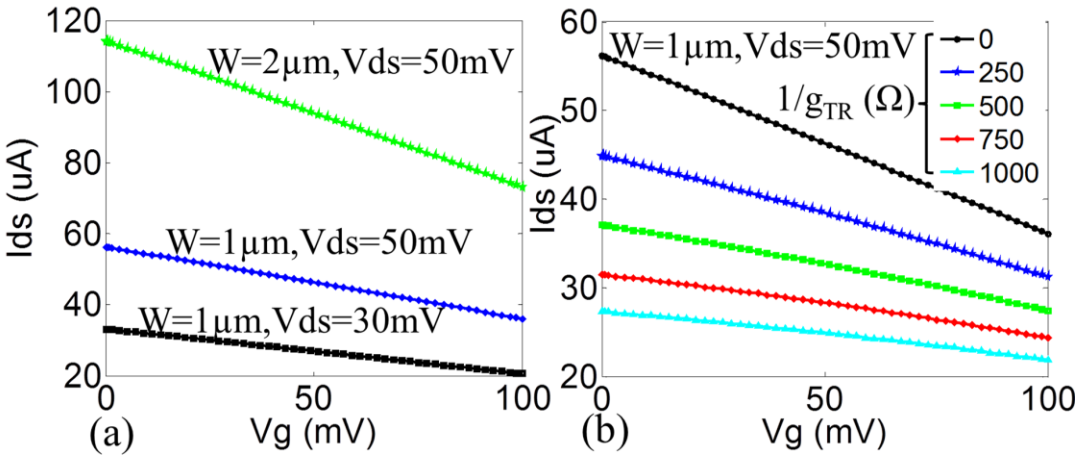


Fig. 5.7 (a) DTCS I_{ds} vs. V_g for different width and V_{ds} (b) non-linearity characteristics of DTCS transistor due to drain terminal memristor load

5.5. Application & Performance Results

In this section, we apply the proposed hybrid Spin-CMOS ANN hardware in a benchmark application (character recognition). We also discuss the performance and its comparison with other CMOS and spin based neuron designs.

In the hybrid Spin-CMOS ANN hardware design, the CMOS peripheral circuits are simulated using IBM 45nm SOI technology. In the character recognition application, the overall process can be divided into two steps - edge extraction and pattern matching. Note that, the edge extraction and ANN training are performed offline. Each alphabet

feature vector is composed of 64 components extracted from four directions: horizontal, vertical and $\pm 45^0$ [129] (Fig. 5.8). Each 64-component feature vector is one test vector to a pre-trained feed-forward ANN composed of hidden layer and output layer as shown in Fig. 5.8. Table-5.2 shows the MATLAB neural network training results using four different neuron transfer functions for the same benchmark and recognition accuracy. It can be seen that the hard-limiting step-function requires much more hidden neurons than the other soft-limiting neurons. It is mainly because the soft-limiting neuron, with a continuous output, has a much larger modeling capacity. Thus, as a soft-limiting neuron model, our proposed STT-SNN can achieve a more compact network size compared to hard-limiting neurons. The mapped hidden layer area can be seen in Fig. 5.10b. For all cases, the number of output neurons is the same, since each output neuron corresponds to one alphabet.

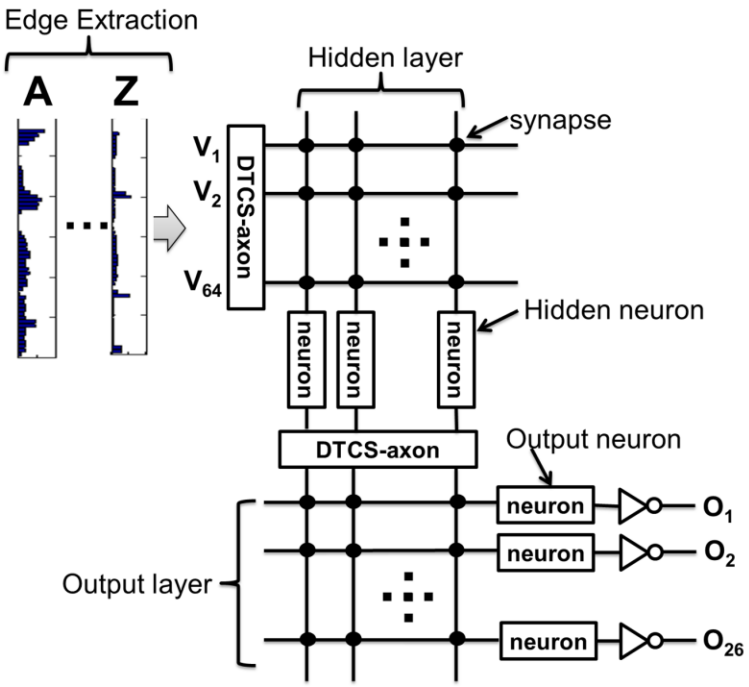


Fig. 5.8 Alphabet feature vectors and two-layer feed-forward ANN architecture. Note, the hardware implementation of each layer can be seen in Fig. 5.6

Table. 5.2 Number of Neurons for Different Neuron Transfer Functions

Transfer functions	Hard-limiting	Soft-limiting		
	Step	Saturated linear	Sigmoid	STT-SNN
# of hidden neuron	24	9	4	5
# of output neuron	26	26	26	26

In the ANN architecture as shown in Fig. 5.8, DTCS-axons in the first (hidden) layer take the analog voltage inputs proportional to input feature vectors and convert them to current going through the MCA-synapses. In all, 64×2 DTCS-axons (positive and negative weights) are required and the MCA (synapse matrix) size is 128×6 (5 hidden neurons and one dummy column). The output layer contains 5×2 DTCS-axons and the MCA size is 10×27 (26 output neurons and one dummy column). Note that, a Gaussian distributed random noise (σ=5%) was added to each memristor conductance value in our simulations to model variations. The simulation results are shown in Fig. 5.9a. The figure shows the normalized output neuron voltages for 26 test alphabets. Pixel (*i, j*) indicates the i^{th} output neuron voltage when the input is the j^{th} alphabet.

During the supervised training of the ANN, the 26 output neurons (O1 to O26) are assigned to indicate 26 alphabets (‘A’ to ‘Z’) respectively. Thus, for each test alphabet (each row in Fig. 5.9a), the diagonal value-(*i, i*) should be the maximum to indicate a correct match. The first (‘A’) and last row (‘Z’) voltage values are separately plotted in Fig. 5.9b. It can be seen that, when the input pattern is ‘A’, output neuron-‘O1’ is the winner. In the case that ‘Z’ is the input pattern, output neuron-‘O26’ is the winner. For the output winner detection, a simple Winner Take All (WTA) circuit described in [143] can be employed. Based on SPICE simulation for this simple alphabet benchmark, we found the voltage difference between the winner and other output neurons is sufficiently large (Fig. 5.9a). Thus, we attached an inverter to each output neuron to sense the output. Only the winner output bit is ‘*0*’, while the others are ‘*1*s’.

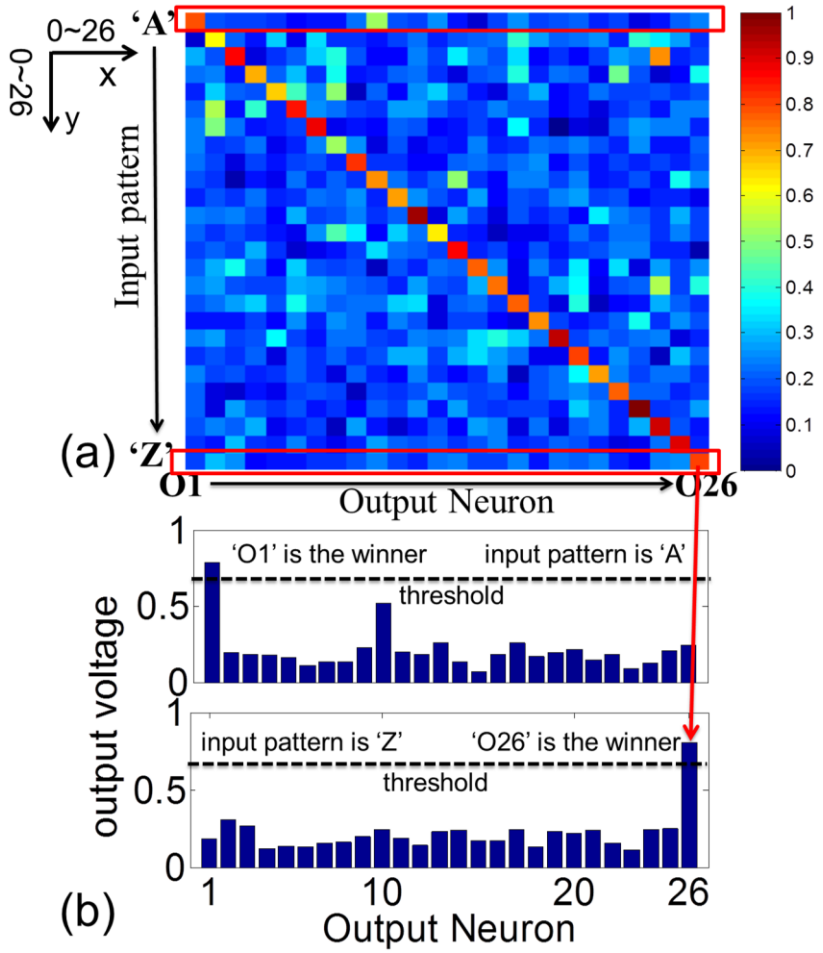


Fig. 5.9 (a) Normalized 26 output neurons' voltages for 26 test input patterns. Note that, pixel (*i*, *j*) indicates i^{th} output neuron voltage for j^{th} input pattern. (b) The 26 output neurons' voltages when the input patterns are 'A' and 'Z'

The energy consumption of a single STT-SNN has three components: programming, sensing and reset energy. For an average of ~40µA of lateral current flowing across the STT-SNN free layer (the total current out of one MCA column/ row), the programming energy is simulated as ~0.5fJ for 1ns clock cycle time. The second component (sensing energy) can be ascribed to the MTJ-based read operation. A read current of ~25 µA (~20% of DW depinning vertical critical current) would lead to ~2.5fJ energy consumption for 1ns read speed. Note that, the sensing current and sensing energy can be reduced by increasing the MTJ MgO thickness (hence, the resistance-area product of MTJ [137]). For the reset operation, a 50µA-1ns current pulse is used in our simulation,

leading to ~0.75fJ reset energy. Thus, the total energy dissipation of one single STT-SNN is ~3.75fJ. Note that, each phase delay is set to be the same (1ns) to make it easy for pipelining the design. We compare the proposed STT-SNN energy with other recent artificial neuron implementations in Fig. 5.10a. Compared with CMOS analog and digital neurons in [114][141], STT-SNN leads to the possibility of more than two orders of magnitude lower energy dissipation. The LSV-based spin-neuron (step function) is around one order of magnitude larger than STT-SNN because of the large hard-axis preset energy [129]. The reasons why the energy consumption of DWM spin-neuron (step function) [114] is smaller than that of STT-SNN is mainly due to *1)* spin-orbital coupling is employed to increase the DW velocity; *2)* a smaller sense current is used; *3)* it implements a step function with hysteresis and no reset operation is required.

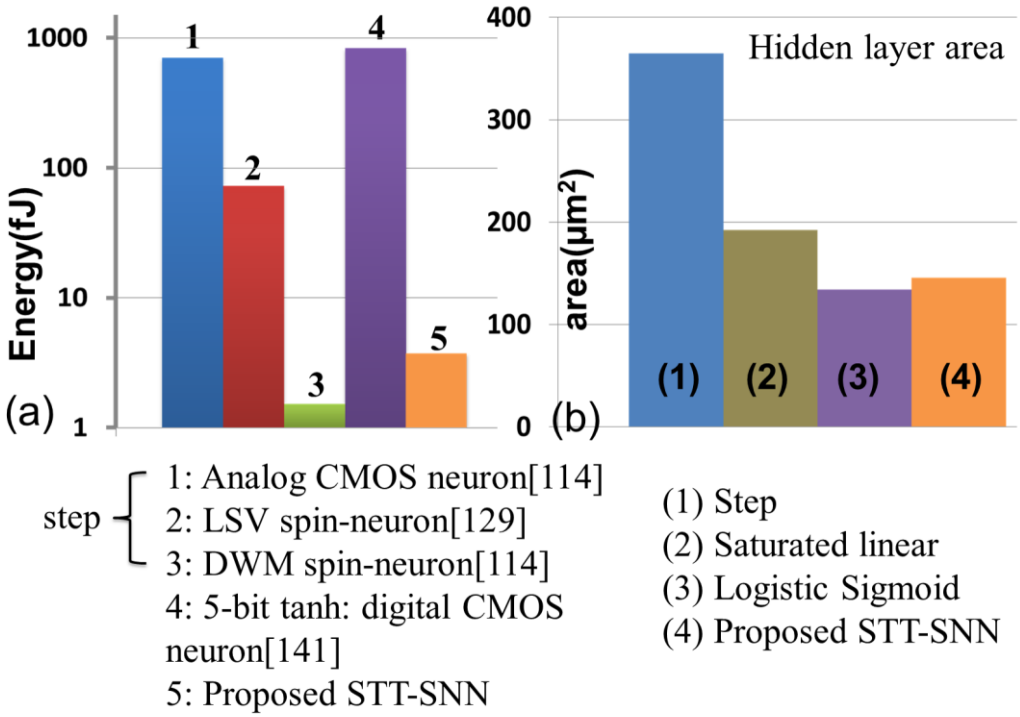


Fig. 5.10 (a) Energy for different single neuron implementations, (b) hidden layer area based on different neuron transfer functions

Apart from the ultra-low energy consumption, the soft-limiting functionality of STT-SNN also leads to reduced number of hidden neurons, and hence smaller hidden layer area for the same benchmark [97][121][142]. The hidden layer areas using four different neuron transfer functions are compared in Fig. 5.10b. It can be seen that the hidden layers using soft-limiting neurons consume much smaller area because of less number of synapses and neurons. STT-SNN leads to ~2.5× lower hidden layer area compared to the hard-limiting step function neuron based ANN. The system level SPICE simulation of our proposed ANN hardware shows the total energy consumption for one alphabet recognition is ~650fJ (Fig. 5.11a), which is ~6.8× lower than that of the LSV neuron (step function) based ANN and more than two orders magnitude lower than the digital/ analog ANN implementation for the same benchmark [129]. Note that, ANN training is performed offline and the programming of MCA-synapses is a one-time operation. Hence, the memristor programming energy is not included in our analysis.

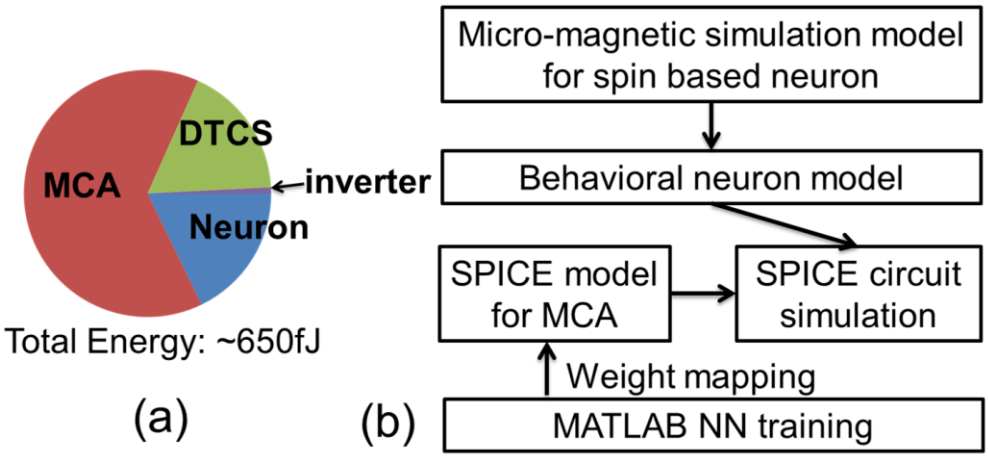


Fig. 5.11 (a) Energy analysis of the proposed ANN hardware for character recognition benchmark, (b) simulation framework

Fig. 5.11b depicts the simulation framework used in this chapter. We employed micro-magnetic simulation for the proposed STT-SNN and it was calibrated with

experimental data from [126]. The MTJ is modeled using NEGF-LLG solution for spin to charge interface [137]. A compact behavioral model of STT-SNN was used in SPICE simulation. The ANN was trained offline using MATLAB Neural Network toolbox [136], which generates the synapse weight matrix for the hidden and output layers from the given training data. The memristor conductance (1kΩ to 32kΩ, [130]) was programmed based on the synapse weight matrix in SPICE. In the system simulation, a Gaussian distributed random noise (σ=5%) was added to each memristor conductance value to account for variations.

5.6. Summary

In this chapter, we presented a domain wall motion based spin-transfer torque device that can efficiently implement a neuron with a soft-limiting non-linear transfer function, operating at ultra-low supply voltage and current. The spin based neuron device allows the peripheral circuits and memristor crossbar array synapses to also operate at very low voltages, thereby leading to ultra-low power consumption for the whole system. The proposed neurons are used to design artificial neural networks that show more than two orders of magnitude lower energy dissipation compared to analog and digital CMOS ANN implementations in 45nm CMOS technology and ~2.5× lower hidden layer area compared with hard-limiting neuron based ANNs. We believe that the proposed spin-transfer torque based soft-limiting non-linear neurons along with MCA-synapses can be used to build energy efficient neuromorphic computing hardware for cognitive computing applications.

6. BRAIN-INSPIRED COMPUTING USING COUPLED SPIN TORQUE OSCILLATORS ARRAY

Spin Torque Oscillator (STO) is based on magnetic spin valves that constitute of a fixed and a free magnetic layer. The magnetization of free layer can be set into sustained oscillations by injecting charge current through the device, under appropriate bias conditions and device configurations. STOs are compact, frequency tunable and CMOS compatible microwave oscillators. They can generate high oscillatory signals using low DC bias current. Moreover, multiple STOs can be frequency/ phase locked through magnetic interaction between free layers, electrical connectivity or external oscillating current/ magnetic field injection. The dynamics of coupled STOs array can be exploited as a robust primitive computational operator for associative computing, image and video analysis, etc. In this chapter, we first discuss the numerical device simulation framework for STOs and different coupling mechanisms, including magnetic coupling, electrical coupling and injection locking. Then, we present an application of injection locked spin Hall induced oscillators in associative computing as a case study. We also discuss CMOS interface circuitries for the design of spin hall induced oscillators based associative module.

6.1. Introduction

The brain-inspired computing models proposed in literatures [83][96][97][107][130][150], constitute of associative pattern matching as the core data processing task. Such associative computing may involve evaluation of conventional distance metrics like, Hamming distance, Gaussian distance or dot product between the template and input patterns. Practical associative computing architectures, like those based on pattern clustering [83][96][97][130][150], may require matching of input patterns with a large number of template patterns, stored in a tree-like hierarchy.

Implementation of such hardware, using the conventional von Neumann digital architecture may incur prohibitively high energy and real estate cost for computing as well as memory.

Recent years have seen growing interest in emerging nano-devices that can provide direct and energy efficient mapping of computing primitives required for such pattern matching tasks, involved in associative computing [83][107][130][150][151]. The pattern matching computations, being inherently variation tolerant, can exploit the "inexact" terminal characteristics of such nano-devices to perform non-Boolean, analog mode operations upon inputs [83][130][150].

Spin Torque Oscillators (STO) are based on magnetic spin valves that constitute of a fixed and a free magnetic layer [151]. The magnetization of the free layer can be set into sustained oscillations by injecting charge current through the device, under appropriate bias conditions and device configurations [48], [164]-[169]. An input dependent shift in the bias state of a set of phase synchronized STOs can be employed for pattern matching applications [143], [151]-[154]. However, the choice of device configuration, synchronization technique and interface circuits can heavily impact the design feasibility and the overall benefits of STO based computing modules.

Recently proposed 3-terminal Spin Hall Effect (SHE) based STO (SHE-STO) offers separate control of frequency and output microwave amplitude, which provides a simple method to tune the output voltage swing without disturbing the frequency. It minimizes the interface circuit overhead for sensing the oscillations [48].

6.2. Spin-Torque Oscillators

In this subsection, we first describe the standard 2-terminal STO (2T-STO) and the basic design conflicts associated with its application to low power computing. Following this, SHE-STO is presented as an alternative device that can overcome the limitations of 2T-STO. The STO numerical simulation model based on Landau-Lifshitz-Gilbert (LLG) equation is also presented.

6.2.1. 2 Terminal STO

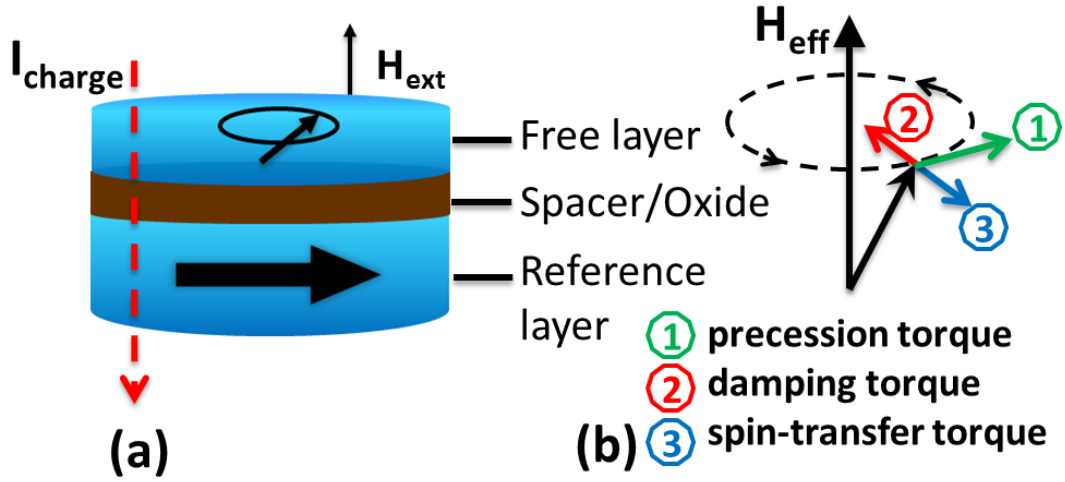


Fig. 6.1 (a) 2-terminal STO device structure, (b) different torque terms acting in the free layer

A standard 2T-STO [164]-[169], shown in Fig. 6.1a, has two ferromagnetic layers separated by either a thin non-magnetic metal (Giant Magneto Resistance -- GMR device) or a thin insulating oxide (Tunneling Magneto Resistance – TMR device). The ferromagnetic layers have two stable magnetization states, depending upon the magnetic anisotropy [168]. One of the magnetic layers has the fixed magnetization, while the magnetization of the other (free layer) one can be influenced by a charge current passing through the device and/or by an applied magnetic field. The fixed magnetic layer spin-polarizes the electrons, which in turn exert spin-transfer torque (STT) in the free layer. The magnetization dynamics of STO free layer can be modeled by Landau-Lifshitz-Gilbert equation with a Slonczewski's term (LLGS equation) [1][185][186] as shown in the followings:

$$\frac{d\boldsymbol{m}}{dt} = -|\gamma|\boldsymbol{m}\times\boldsymbol{H}_{eff} + \alpha\left(\boldsymbol{m}\times\frac{d\boldsymbol{m}}{dt}\right) + |\gamma|\beta\varepsilon\left(\boldsymbol{m}\times\boldsymbol{m}_p\times\boldsymbol{m}\right) - |\gamma|\beta\varepsilon'\boldsymbol{m}\times\boldsymbol{m}_p \tag{6.1}$$

$$\beta = \left|\frac{\hbar}{\mu_0 e}\right| \frac{J}{tM_s} \tag{6.2}$$

$$\varepsilon = \frac{P\Lambda^2}{(\Lambda^2+1)+(\Lambda^2-1)(\boldsymbol{m}\bullet\boldsymbol{m_P})} \tag{6.3}$$

$$H_{eff} = H_{ext} + H_{ani} + H_M + H_{noise} \tag{6.4}$$

where, $\boldsymbol{m}$ is the free layer magnetization, which is a unit vector pointing to the magnetization direction. γ is the Gilbert gyromagnetic ratio, α is the damping constant, $\hbar$ is the Plank's constant, e is the electron charge, J is current density, t is the free layer thickness, Ms is the saturation magnetization of the magnet, P is the polarization constant, $\boldsymbol{m_p}$ is the direction of spin polarization of spin current, Λ is the spin torque asymmetry parameter, $\acute{\varepsilon}$ is the secondary spin transfer term. It includes a precession term induced by effective field H_{eff} (equation-6.4). Here, H_{ext} is the external magnetic field, H_{ani} corresponds to the free layer anisotropy field, H_M represents the magneto-static field which is proportional to the component of the free layer magnetization along its easy axis, and H_{noise} denotes the noise term that models the thermal fluctuations [19][127][166]. As shown in equation-6.1, the first term is the '*precession term*' resulting from magnetic field. The second term denotes the '*damping term*'. The last two terms represent current induced torques that take Slonczewski term and field-like term, respectively. When current is injected through the device shown in Fig. 6.1a (metal spacer or tunneling barrier), it becomes spin-polarized. This flow of spin-polarized current generates spin-transfer torque acting on the magnetic moments. The magnitude of the last two torques is dependent on material and device structures. Note that, for GMR devices, the field like term $\boldsymbol{m}\times\boldsymbol{m_p}$ is typically negligible as transverse spins dephase rapidly [1][184]. While for TMR devices, besides the in-plan torque predicted by Slonczewski [1][185][186], this field-like (output of plane) torque is proven significant in modeling the dynamics of magnet [185][187]. For a given static magnetic field, the free layer magnetization can achieve sustained oscillation when the STT and damping torque balance out each other. (Fig. 6.1b) [164]-[169]. The conductance of STO can be expressed as a function of relative angle (θ) between the magnetizations of the two ferromagnetic layers as:

$$G = \frac{G_P + G_{AP}}{2} + \frac{G_P - G_{AP}}{2}\cos\theta \tag{6.5}$$

where, G_P and G_{AP} denote the conductance when the two layers are parallel (θ = 0º) and antiparallel (θ = 180º). The absolute resistance of a GMR device is much smaller than that of a TMR device (notably, the resistance area product for GMR device can be two orders of magnitude lower than a TMR device [164][165][167][169]). A GMR-STO, being fully metallic, can be operated with very low voltage. However, the sensed signal amplitude is very low, which requires complex sensing circuitry to amplify the signal, leading to high power consumption [164][165] (listed in table-6.2). On the other hand, though the TMR based STO can provide large amplitude output signals, due to the high resistance tunnel junction, it requires a larger bias voltage, leading to energy inefficiency at the device level [167][169] (listed in table-6.2). The standard 2-terminal STO shares the biasing and sensing path, leading to disturbance in tuning frequency and output voltage swing. The recently proposed SHE-STO [48] can overcome the aforementioned bottlenecks.

6.2.2. Spin Hall Effect STO

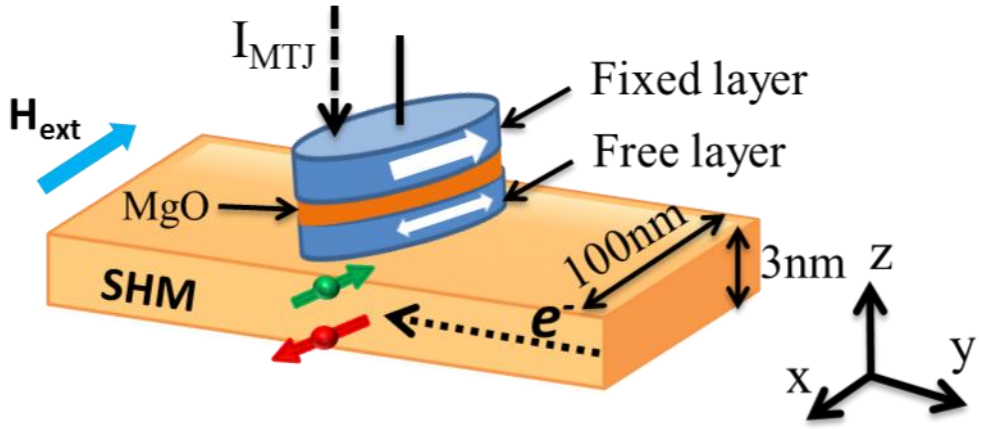


Fig. 6.2 SHE-STO device structure. Spin accumulation at the top and bottom surface of SHM due to SHE. H_{ext} is the applied external magnetic field.

Recently it was experimentally demonstrated that the spin hall effect (SHE) in a thin film with strong spin-orbit coupling can excite magnetic precession in an adjacent ferromagnetic film [48][49][170]. Such a device structure is shown in Fig. 6.2 where a magnetic tunnel junction (MTJ) is milled on the spin hall metal (SHM, β-Ta[49], β-W[65], Pt[63], doped Cu[170]) nano-strip. When a charge current is injected along the SHM strip, the opposite spins accumulate at the top and bottom surface of the SHM strip. Thus, a spin current is generated perpendicular to the SHM strip and is injected to the adjacent MTJ free layer [171]. The spin current generated due to SHE then exerts a spin-transfer torque in the MTJ free layer, leading to a sustained magnetization oscillation of the MTJ free layer. The spin current corresponding to the charge current (*Ic*) can be modeled [171] by:

$$\vec{I_s} = P_{she}(\vec{\sigma} \times \vec{I_c}) \tag{6.6}$$

$$P_{she} = \frac{I_{s-z}}{I_{c-y}} = \frac{A_{MTJ}}{A_{SHM}} \theta_{SHE} (1 - \sec h(\frac{t}{\lambda_{sf}})) \tag{6.7}$$

where $\vec{\sigma}$ is the spin direction, P_{she} denotes the spin hall injection efficiency. The magnitude of P_{she} equals to the ratio of the spin current ($I_{s\text{-}z}$, ±z) to lateral charge current ($I_{c\text{-}y}$, ±y). A_{MTJ} is the area of the MTJ, and A_{SHM} is the cross section area of SHM strip perpendicular to the charge current direction. t is the thickness of SHM, λ_{sf} is the spin flip length, θ_{SHE} is the spin hall angle for the SHM to MTJ free layer interface. In 2T-STO devices we described in previous subsection, spin current is generated by passing charge current through a ferromagnetic layer. Thus, the efficiency of spin current generation is inherently limited by the polarization efficiency of the ferromagnetic layer, less than 1. In SHE-STO, the spin hall injection efficiency can be easily larger than 1. The spin current due to SHE exerts a spin-transfer torque in the adjacent MTJ free layer, which reduces the effective magnetic damping torque. If the STT and magnetic damping torque balance out each other, the MTJ free layer magnetization can achieve sustained oscillation. This dynamics of free layer spins can then be modeled as follows [172]:

$$(1+\alpha^2)\frac{d\vec{m}}{dt} = -|\gamma|\mu_0\vec{m}\times\vec{H} - \alpha|\gamma|\vec{m}\times\vec{m}\times\vec{H} \\ -\vec{m}\times\vec{m}\times\frac{\vec{I}_s}{qN_s} + \alpha\vec{m}\times\frac{\vec{I}_s}{qN_s} \tag{6.8}$$

Where $Ns=M_sV/\mu_B$ is the number of spins comprising the magnet, μ_B is Bohr magneton, V is the volume of the magnet. The $\vec{H}$ acting on the magnet contains anisotropy field ($2Ku/M_s$), demagnetization field ($4\pi M_s$) and thermal noise field [127]. $\vec{I}_s$ is the spin current induced by SHE that is modeled as in equation-6.6. The device parameters used in simulation are listed in table-6.1. The transient simulation of free layer magnetization with $I_{bias}=I_c=320\mu A$ corresponding to output frequency of ~6.6GHz is shown in Fig. 6.3b. SHE-STO output frequency can be tuned by varying the DC bias current as shown in Fig 6.3a.

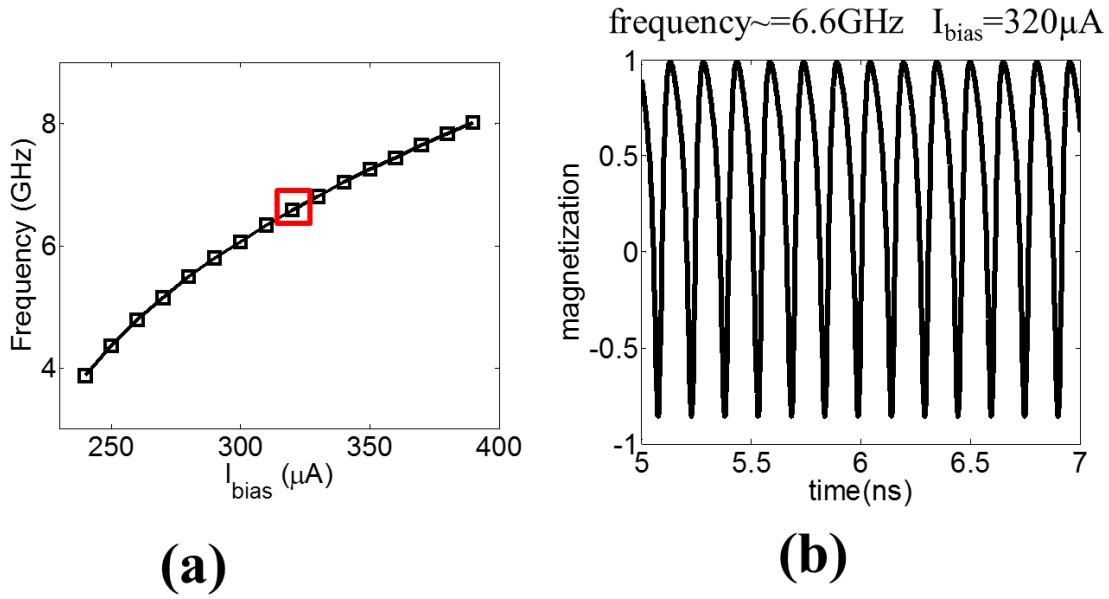


Fig. 6.3 (a) SHE-STO output frequency vs. I_{bias}, (b) transient simulation of SHE-STO free layer oscillation when I_{bias}=320μA.

The biasing and sensing circuit of SHE-STO can be seen in Fig. 6.4. The oscillation dynamics of the free layer can be sensed by injecting a small read current (I_{MTJ}) into the MTJ formed between the free layer and fixed layer (Fig. 6.4), converting the oscillations of the MTJ resistance into an oscillating voltage. The resistance of SHM equals to $\rho L/(wt) = \sim 1k\Omega$, where ρ is the resistivity of SHM ($\rho = \sim 200\mu\Omega \cdot cm$ [63][65]), L is the SHM length (150nm), t is the SHM thickness (3nm). One terminal of the SHM is used as DC biasing and the other one is grounded. Thus, for the sensing of SHE-STO, I_{MTJ} goes through the MTJ and SHM layer to the ground [48].

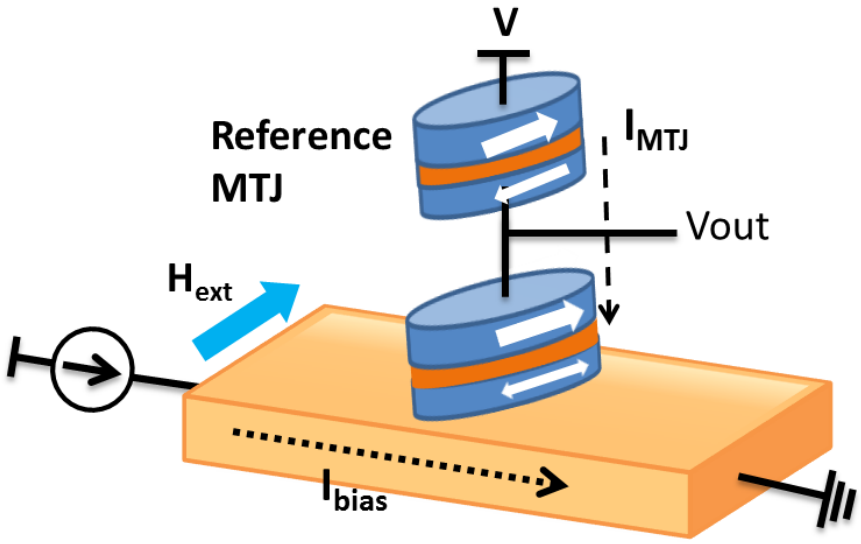


Fig. 6.4 SHE-STO biasing and sensing circuit

Fig. 6.5 shows the peak-to-peak SHE-STO output voltage swing vs. different TMR of the MTJ. Higher TMR may provide higher output voltage swing and hence better robustness. High oxide thickness (t_{ox}) for MTJ provides higher absolute resistance for the voltage divider circuit, minimizing the read current and hence, the static power associated with the sensing operation is reduced. However, a too high MTJ resistance diminishes the output swing for high frequency operation, due to low pass filtering effect. In this work,

we use a TMR of ~200%, supply voltage *V=0.5V*, reference MTJ resistance of 5.2KΩ, which yields a voltage swing of ~0.12V.

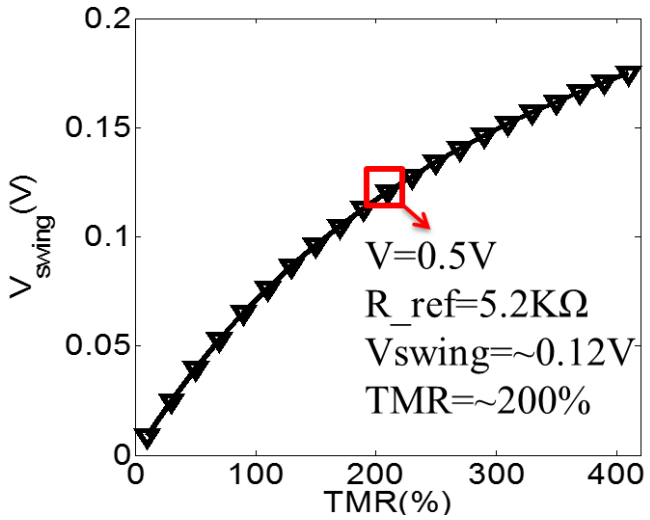


Fig. 6.5 peak-to-peak output voltage swing vs. different TMR

Table. 6.1 SHE-STO Device parameters used in simulation

Symbol	Quantity	Values
W	SHM width	70nm
T	SHM thickness	3nm
θ_{SHE}	Spin hall angle	0.3
λ_{sf}	spin flip length	1.5nm
Ea	Energy barrier	60kT
α	Damping factor	0.03
$\mu_0 Ms$	Saturation magnetization	1T
Hext	External magnetic field	750Oe
ρ_{SHM}	SHM resistivity	~200μΩ•cm

SHE-STO offers separate control of frequency and output microwave amplitude, which provides a simple method to tune the output voltage swing without disturbing the

frequency. It minimizes the interface circuit overhead for sensing the oscillations [48]. Table-6.2 compares the power consumption of SHE-STO with two terminal STOs based on GMR and TMR devices. It shows that, for 2T-GMR STO, the output voltage swing is around 1mV [164][165]. An amplifier is implemented to amplify the output voltage swing to be around 100mV, leading to power consumption of 1.4mW based on our simulation using IBM 45nm CMOS technology. On the other hand, for 2T-TMR STO, the biasing power is larger than that of SHE-STO because of higher biasing resistance as shown in table-6.2. The total STO power consumption of SHE-STO circuit shown in Fig. 6.5 is the lowest compared with TMR or GMR based 2-terminal STO. Furthermore, the 3-terminal SHE-STO device geometry enables independent control of output amplitude and frequency because of the separation of biasing and sensing paths [48].

Table. 6.2 Comparison of power consumption for 2T-STO and SHE-STO

STO type	2T-GMR[165]	2T-TMR[169]	SHE-STO
Device resistance	R_{avg}=10Ω ΔR=100mΩ	R_p=310Ω R_{ap}=620Ω	R_{SHM}=~1kΩ
Device area	π×35nm×35nm	π×35nm×80nm	π×30nm×50nm
Bias current	~10mA	~1mA	~320μA
Bias power	1mW	465μW	103 μW
Sensing voltage	-	-	0.5V
Sensing power	1.4mW (Amplifier)	-	26μW
Output peak-to-peak voltage swing	100mV (Amplified from 1mV)	310mV	120mV
Total power	2.4mW	465μW	129μW
External magnetic field	2.5KOe	11KOe	750Oe
Field line current/ power	62.5mA/ 11.7mW	275mA/ 226.9mW	18.8mA/ 1.1mW

Comments: R_{avg} is the average resistance, ΔR is the resistance change of oscillator, R_p is the parallel MTJ resistance, R_{ap} is the anti-parallel MTJ resistance, R_{SHM} is the spin hall metal resistance. The sensing amplifier of GMR-STO is implemented in IBM 45nm CMOS technology

The external magnetic field requirement and field line power consumption of each STO is also listed at the end of table-6.2. Note that the required external magnetic field can be generated by applying a current flowing through a field line (assuming the distance between the field line to the magnet is 50nm), where the magnitude of current required is computed using Biot-Savart law [188]. The field line power consumption is computed under the assumption that a copper wire (length= 150nm, area= $40\times40nm^2$) is used. It can be easily seen that the power consumption of field line is much higher than STO power (biasing and sensing) for both 2T-STO and SHE-STO. For large scale computing applications, the external magnetic field can be potentially removed by either tilting the ellipse of MTJ in SHE-STO or employing an MTJ structure with a perpendicular magnetic anisotropy (PMA) free layer [59]. For a practical associative pattern matching hardware, integration of a large number of STOs might be essential [143][153][154][158]. SHE-STO can facilitate such large scale integration, due to the simplified CMOS interface and low power operation it offers.

6.3. STO Coupling Mechanisms

Multiple STOs can be frequency and phase synchronized through magnetic coupling [155]-[157], [177][178], electrical coupling [158] or injection locking mechanisms [159][160][161]. In this subsection, we discuss various STO coupling mechanisms, namely magnetic coupling, electrical coupling and injection locking mechanisms. The two terminal IMA STO benchmarked with the experimental data in [169] is used in this subsection.

6.3.1. Magnetic coupling

Two or more STOs can interact with each other via magnetic coupling and can lock to a common frequency if they are located close to each other. Experimentally frequency

locking phenomenon has been demonstrated for two STOs [177][178]. The effect of magnetic coupling is simulated using a coupling field (H_{couple}) term in the effective field (H_{eff}) of LLGS equation [179]. For the case of two STOs (STO1 and STO2), the coupling field acting on STO1 is given by

$$H_{couple1} = \begin{pmatrix} H_{couple\_x1} \\ H_{couple\_y1} \\ H_{couple\_z1} \end{pmatrix} = Cc \begin{pmatrix} m_{x2} \\ m_{y2} \\ m_{z2} \end{pmatrix} \tag{6.9}$$

where, Cc is the coupling coefficient given by $Cc = \frac{M_S A_S}{d^2}$ [172], M_S is the saturation magnetization of the magnet, A_S is the coupling area and d is the distance between STOs. The total effective field acting on the magnetization of free layer of first STO is given by

$$\overrightarrow{H_{eff\_new1}} = \overrightarrow{H_{eff\,1}} + \overrightarrow{H_{couple1}} \tag{6.10}$$

Similarly the second STO experiences a coupling field which depends on the magnetization of the first STO.

Experimentally, as demonstrated in [155], the spin wave propagation rather than field based coupling is shown to be the dominant factor leading to frequency locking when the distance between two STOs is larger than 200nm. However in our work, the distance between two STOs as calculated based on coupling coefficient is less than 100nm. At this distances we show that the field based coupling can be sufficiently strong to lock the STOs. Also the STOs are assumed to be isolated so that there is no spin wave propagation.

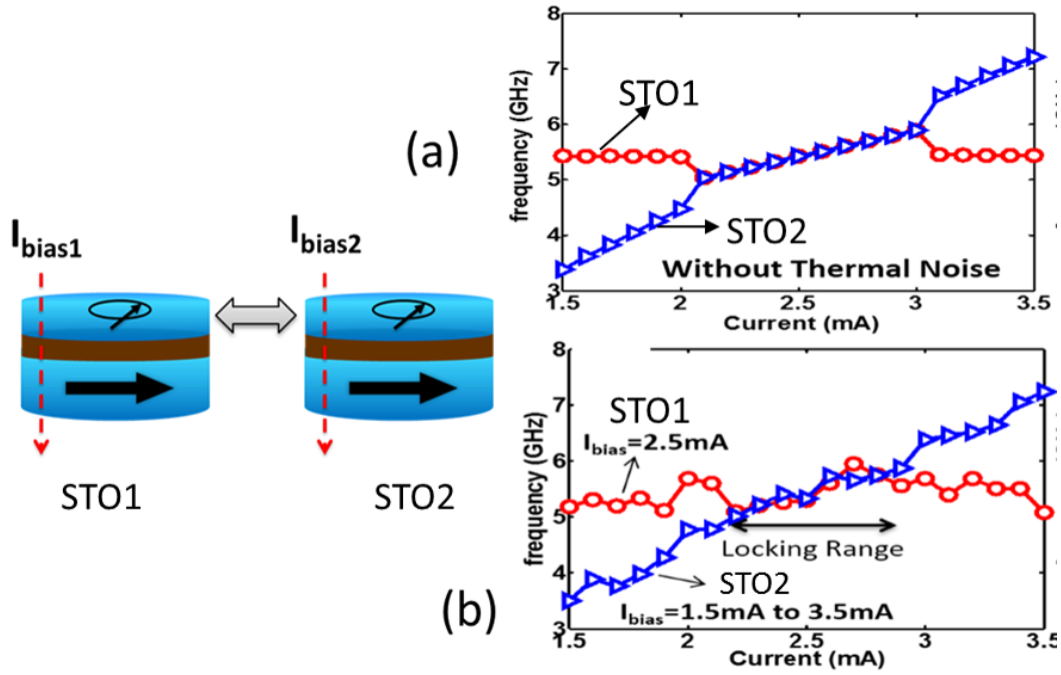


Fig. 6.6 STO frequency vs. DC bias currents in magnetic coupling (a) without thermal noise, (b) with thermal noise at 300K

Fig. 6.6 shows the schematic diagram of two IMA STOs interacting with each other through magnetic coupling. Fig. 6.6a shows the locking range of two IMA STOs without thermal noise. The current (I_{bias1}) through STO1 is kept constant (2.5mA for IMA STO), while the current (I_{bias2}) through the STO2 is increased (from 1.5mA to 3.5mA for IMA STO). When the frequency of STO2 comes close to that of STO1, both STOs get lock to a common frequency. The locking range can be defined as the range of DC input for which the frequencies remain locked. In order to analyze the effect of thermal noise, we modeled thermal effects using a Gaussian random magnetic field H_{noise}=($H_{noise\text{-}x}$, $H_{noise\text{-}y}$, $H_{noise\text{-}z}$,). The mean of the Gaussian distribution is zero, while the standard deviation is $\sqrt{2\alpha K_B T/\gamma M_s V \Delta t}$ [127], where α is Gilbert damping factor, K_B is Boltzmann's constant, γ is the gyromagnetic ratio, M_s is the saturation magnetization, V is the volume of free layer and Δt is the time step used in solving LLG equation. Fig. 6.6b shows the locking range of IMA STOs with thermal noise included in simulations. Note that, frequency vs current

plot is not smooth due to thermal noise in these plots. It can be seen that the locking range is reduced with the thermal noise at room temperature (300K).

6.3.2. Electrical Coupling

Multiple STOs can also be coupled through electrical connectivity as shown in Fig. 6.7 [158]. Each STO has an independent current bias (I_{bias1} and I_{bias2}), leading to independent oscillations. The oscillation of the STO is sensed via tunneling magneto-resistance (TMR) and combined into a broadcast signal:

$$I_{broadcast} = \frac{1}{M_s N} \sum_{i=1}^{N} C_i M_i \tag{6.11}$$

where, N is the total number of STOs, C_i is the coupling constant that can be set by the coupling circuit, M_S is the saturation magnetization, M_i is the i^{th} STO free layer magnetization. This broadcast current is fed back to the network and is superposed with the bias current of each STO. The combined current is then used to drive STO.

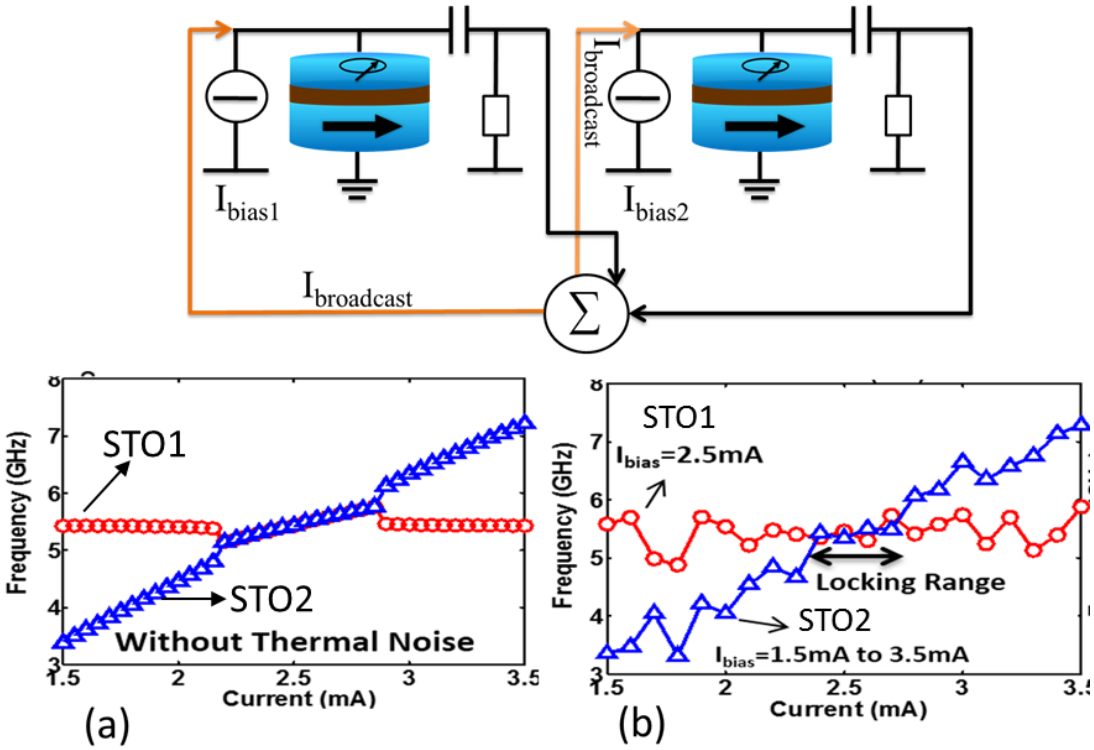


Fig. 6.7 STO frequency vs. DC bias currents in electrical coupling (a) without thermal noise, (b) with thermal noise

For a given non-zero coupling constant, the two STOs are frequency locked when their independent frequencies generated by the DC bias are located in a certain locking range. Fig. 6.7 shows the two electrical coupled IMA STOs, where C_1=C_2=0.3mA, I_{bias1} is kept constant at 2.5mA and I_{bias2} is swept from 1.5mA to 3.5mA. It can be seen that the frequencies of the two STOs get locked when the two DC biases (hence, frequencies) are close to each other (within the locking range). Fig. 6.7a shows the plot without thermal noise and Fig. 6.7b shows the plot with thermal noise for IMA STO respectively. The locking range can be improved by increasing the coupling constant, mainly because of the larger broadcast current amplitude, hence stronger feedback. Compared with magnetic coupling, a large number of STOs can be coupled through such electrical connectivity [158].

6.3.3. Injection Locking

Synchronization of STOs to an external Radio Frequency (RF) oscillating signal (injection locking) was experimentally studied as a function of STO intrinsic parameters [160][161][180][181]. If the frequency of the injected signal is close to the STO free-running frequency, the STO gets frequency locked to this injected reference signal. The injected signal can be either oscillating current (current based injection locking) or oscillating magnetic field (field based injection locking), discussed in detail in following subsections.

6.3.3.1. Current Injection Locking

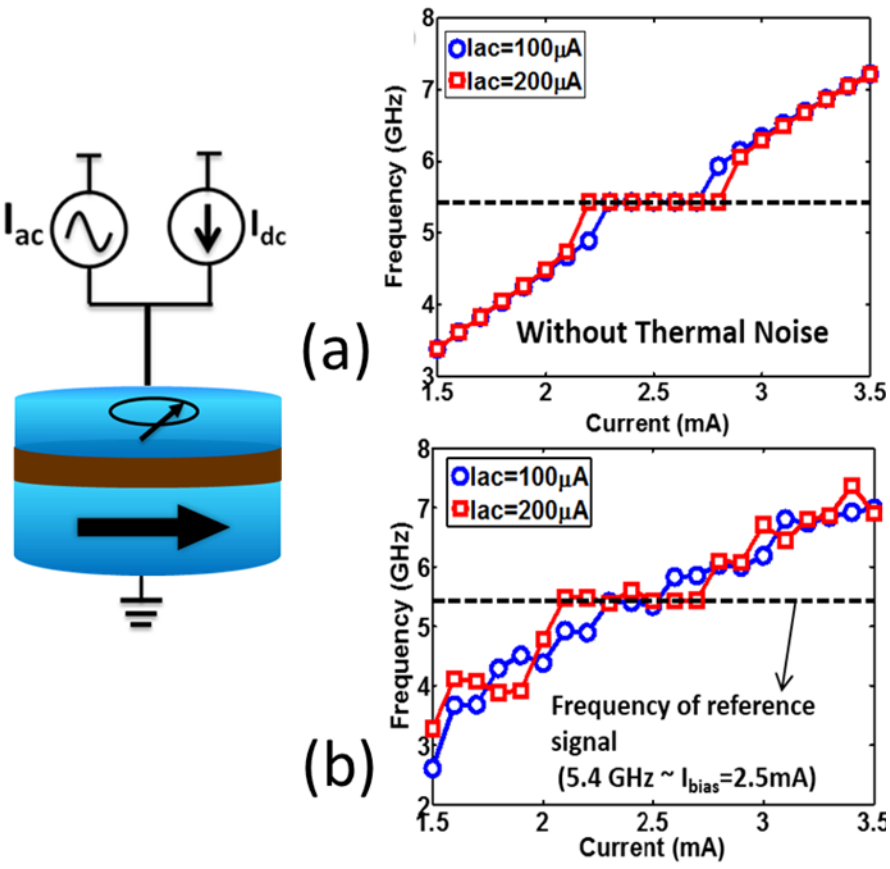


Fig. 6.8 STO frequency vs. DC bias currents in current injection locking mechanism (a) without thermal noise, (b) with thermal noise

In current injection locking, an oscillating current (I_{ac}) is injected into STO along with the bias current (I_{dc}) as shown in Fig. 6.8. In the presence of AC injected current (I_{ac}), the β term (equation-6.2) in LLGS contains both DC and AC currents as shown below.

$$\beta = \left| \frac{\hbar}{\mu_0 e} \right| \frac{J_{ac} + J_{dc}}{t M_s} \tag{6.12}$$

where, $J_{ac} = J_0 \cos(2\pi f_{ac} t)$ is the AC current density. Here J_0 is the injected current density amplitude and f_{ac} is the frequency of injected current. Fig. 6.8a shows the simulation results of the IMA STO output frequencies, with varying injected AC current amplitude. It can be observed that both STO outputs lock to the injected signal when the DC bias is in the locking range. Fig. 6.8a depicts the IMA STO locking behavior when the DC bias is swept from 1.5mA to 3.5mA along with a constant injected current signal of frequency 5.4GHz. IMA STO remains locked to the injected current oscillating signal for the locking range [2.3mA-2.7mA] when the injected current amplitude is 100 μA. This locking range can be increased by increasing the strength of injected signal, which also conforms to the experiments on injection locked STO [161][160][180][181]. If the injected current amplitude is increased to 200μA, the locking range is extended to [2.2mA-2.8mA] correspondingly. Fig. 6.8a shows the plots without thermal noise, and Fig. 6.8b is the plot with thermal noise for IMA STOs. It can be seen that the thermal noise can degrade the locking range.

6.3.3.2. Field Injection Locking

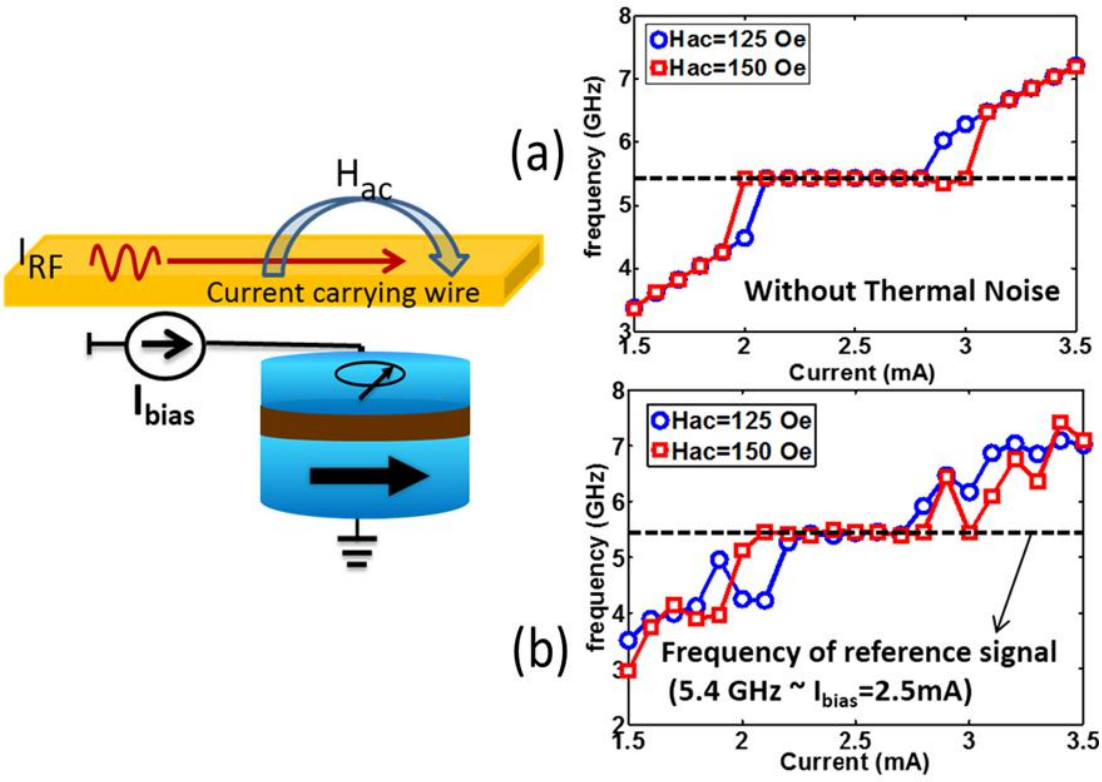


Fig. 6.9 STO frequency vs. DC bias currents in field injection locking mechanism (a) without thermal noise, (b) with thermal noise

In this method, an oscillating field (H_{ac}) is used as an injected reference signal to which the STO is locked. The oscillating field can be generated by a wire carrying oscillating current. The effective field in the presence of H_{ac} is given by

$$\overrightarrow{H_{eff\_new}} = \overrightarrow{H_{eff}} + \overrightarrow{H_{ac}} \tag{6.13}$$

where, $\overrightarrow{H_{ac}} = H_0 \cos(2\pi f_{ac} t)$. Here H_0 is the amplitude and f_{ac} is the frequency of reference field respectively. H_{eff} is the effective field. Fig. 6.9a shows the output frequencies of IMA STO with varying reference oscillating field amplitude. It can be seen that STO frequencies lock to that of reference field when the DC bias is in the locking range. Fig. 6.9a shows the IMA STO locking behavior when the DC bias is swept

from 1.5mA to 3.5mA with a reference magnetic field of frequency 5.4GHz. IMA STO remains locked to the reference oscillating field signal within the locking range [2.1mA-2.8mA] when the reference field amplitude is 125Oe. This locking range can be extended by increasing the strength of injection. If the injected field amplitude is increased to 150Oe, the locking range is extended to [2mA-3mA] correspondingly. Fig. 6.9a shows the plots without thermal noise and Fig. 6.9b shows the same plot with thermal noise for IMA STO, respectively.

Magnetic coupling involves spin wave interaction through a shared magnetic substrate or dipolar field exchange of physically isolated STOs lying in close proximity [155]-[157]. Thus, the number of STOs can be synchronized through magnetic coupling is strongly dependent upon geometrical constraints of a physical design. The maximum number of STOs in a magnetically coupled cluster may, therefore, be limited. For electrical coupling [158], complex interface circuits are required to generate feedback current for each STO, which may dominate the power consumption of STO coupling cluster [158]. Thus, in the next few subsections, we employ injection locking as a robust and energy efficient locking scheme in the STO based associative module design, which essentially provides several advantages over other locking schemes: 1) large number of STOs can be locked in one cluster; 2) immunity to thermal noise and parameters variations; 3) simpler interface circuits design.

6.4. Injection Locked SHE-STO Cluster

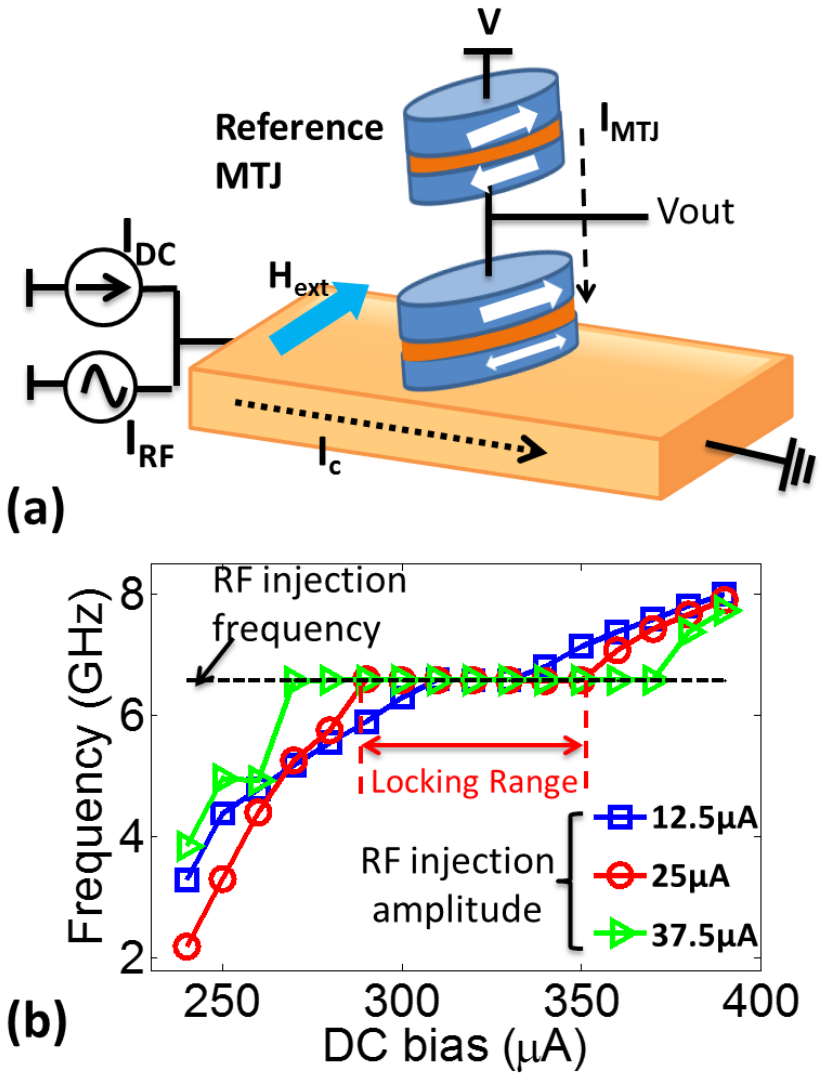


Fig. 6.10 (a) SHE-STO locked to an external microwave current, (b) SHE-STO frequency vs. different RF current amplitude, showing SHE-STO locks to external RF signal and DC locking range increases with higher RF amplitude

In this subsection, we simulate the injection locked SHE-STO array. In Fig. 6.10, the bias current of SHE-STO is the superposition of an external RF and DC currents. In order

to simulate SHE-STO injection locking phenomena, we add the external RF component into our numerical SHE-STO model:

$$I_c(t) = I_{DC} + I_{RF}(\mathrm{t}) \tag{6.14}$$

where, $I_c(t)$ is the superposition of DC bias and external RF current at time *'t'*. We substitute this new $I_c(t)$ to equation-6.6. Fig. 6.10 shows the simulation results of the SHE-STO output frequencies, varying the DC bias and the RF amplitude. It can be seen that the SHE-STO output lock to the external RF signal when the DC bias is in the *DC locking range*. Fig. 6.10b depicts the SHE-STO behavior when the DC bias is swept from 240μA to 390μA along with a constant RF signal of frequency 6.6GHz. SHE-STO remains lock to the injected RF signal for the DC locking range of [290μA-350 μA] when the external RF amplitude is 25 μA. This locking range can be improved by increasing the strength of RF injection (Fig. 6.10b), which conforms to the experiments on injection locked STO [159][160][161].

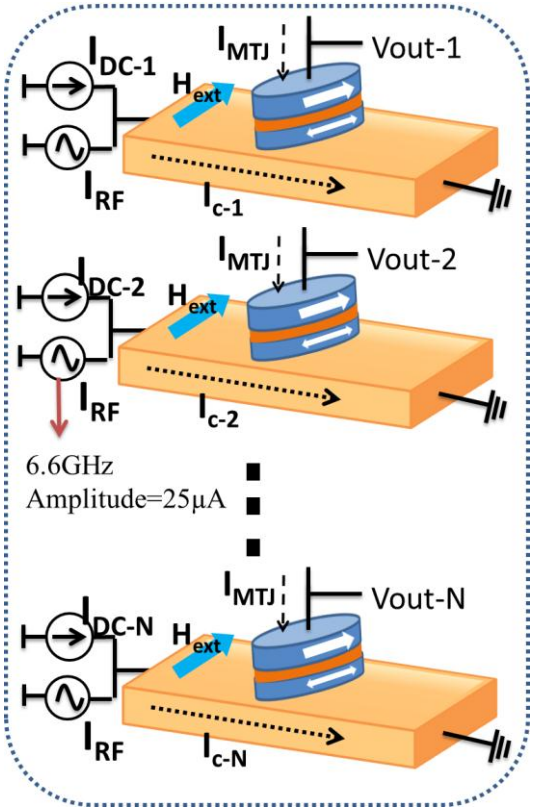


Fig. 6.11 *N*-number of SHE-STOs can be locked to a common external RF signal

A cluster containing multiple SHE-STOs can be locked to a common RF signal as shown in Fig. 6.11. If the external RF frequency is close to that of the free running frequency of the SHE-STOs (determined by the DC bias), the SHE-STOs can get phase locked to the injected RF current signal. If the DC bias of each SHE-STO is close enough (within the DC locking range), all of the SHE-STOs are found to be locked to the common external RF signal as shown in Fig. 6.12a, where RF amplitude-$|I_{RF}|$ = 25μA, RF frequency-f_{RF}=6.6GHz, and the DC bias of each SHE-STO is [I_{DC-1}, I_{DC-2}, …, I_{DC-8}] = [324, 330, 326, 328, 332, 328, 324, 326]μA. If some of the DC biases are distinct enough (out of DC locking range), they are found unlocked to the common external RF signal as shown in Fig. 6.12b ($|I_{RF}|$ = 25μA, f_{RF}=6.6GHz, [I_{DC-1}, I_{DC-2}, …, I_{DC-8}] = [330, 346, 354, 372, 355, 341, 335, 368]μA). Thus, injection locking can be effective for mutual synchronization and phase locking among SHE-STOs.

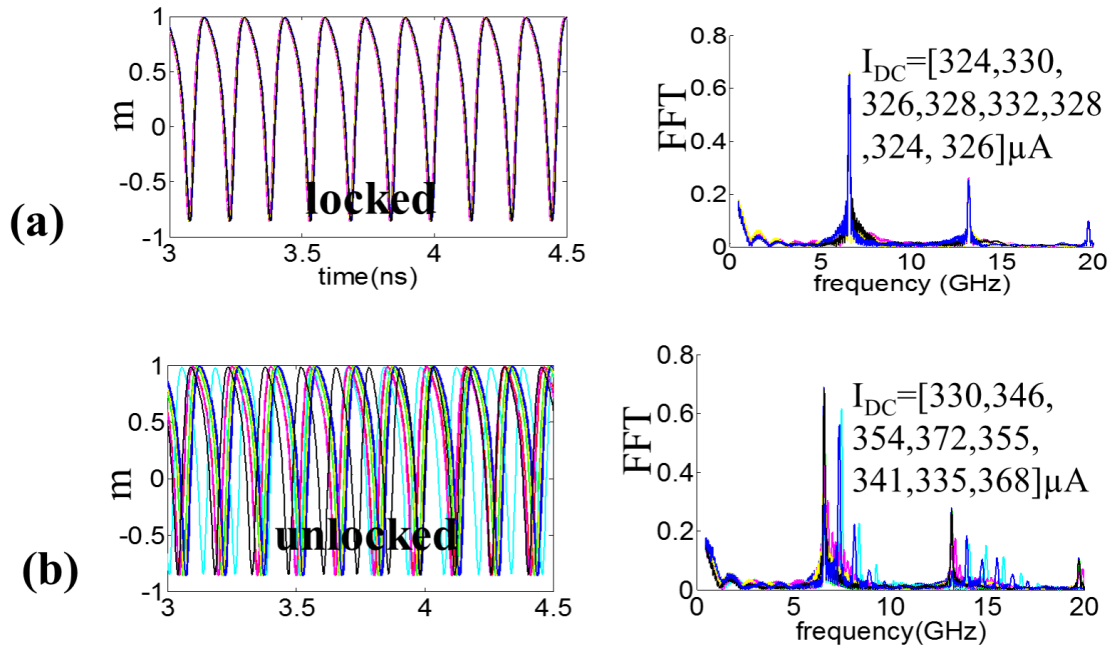


Fig. 6.12 transient waveforms and FFT of 8 SHE-STOs when they are (a) locked or (b) unlocked with different SHE-STO DC biases

We analyze the impact of parameter variations by introducing Gaussian spread (σ=5%) in the critical STO device parameters like the saturation magnetization (M_s) and the Gilbert damping constant (α). These parameters can have significant spread across multiple device samples [162][163], and hence it is important to evaluate the impact of the spread in these parameters upon the dynamics of coupled STOs. Thermal effects are modeled using a stochastic Gaussian magnetic field, H_{noise}=($H_{noise\text{-}x}$, $H_{noise\text{-}y}$, $H_{noise\text{-}z}$,). The mean of the Gaussian distribution is zero, while the standard deviation is $\sqrt{2\alpha K_B T/\gamma M_s V \Delta t}$ [127], where α is Gilbert damping factor, K_B is Boltzmann's constant, γ is the gyromagnetic ratio, M_s is the saturation magnetization, V is the volume of free layer and Δt is the time step used in solving LLG equation. Fig. 6.13a shows the output signals for 8 injection locked SHE-STOs respectively biased with [$I_{DC\text{-}1}$, $I_{DC\text{-}2}$, …, $I_{DC\text{-}8}$] = [324, 330, 326, 328, 332, 328, 324, 326]μA and f_{RF}=6.6GHz, where all of the SHE-STOs get phase locked without considering the parameter variations and thermal noise.

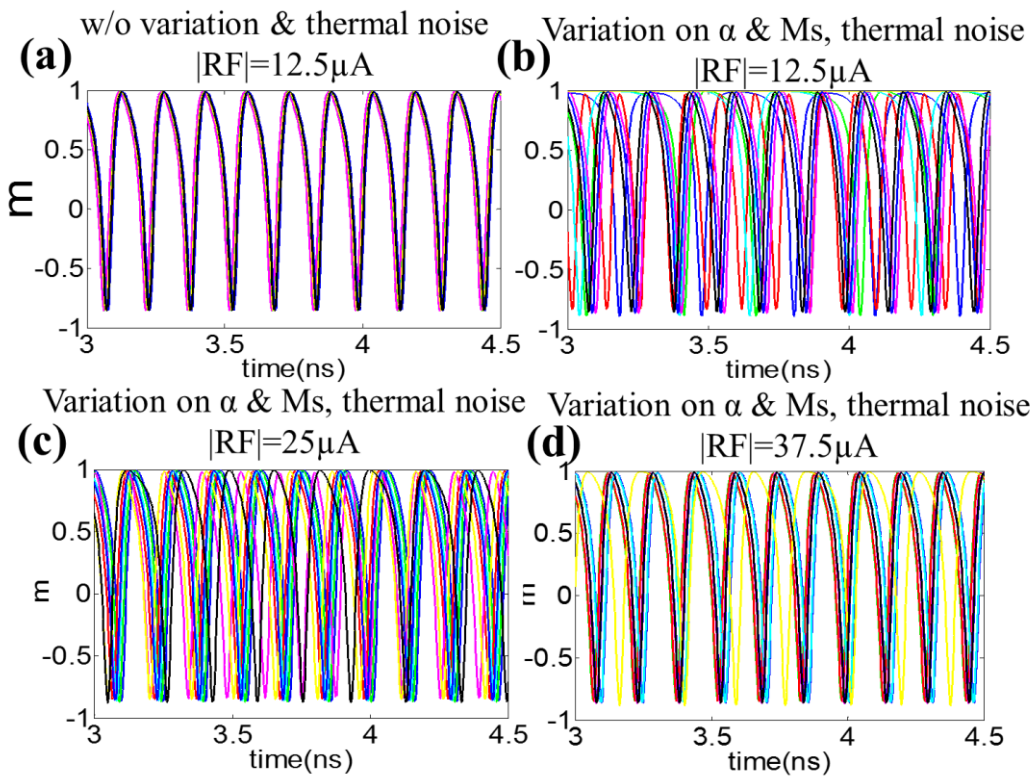


Fig. 6.13 transient plots for 8 injection locked SHE-STOs (a) without parameter variations and thermal noise, (b) with parameter variations and thermal noise when RF amplitude is 12.5μA, (c) 25μA, (d) 37.5μA. Note: the DC inputs of each SHE-STO are [330, 346, 354, 372, 355, 341, 335, 368]μA, external RF frequency is 6.6GHz

When the parameter variations and thermal noise are included, they introduce some jitters and phase noises as shown in Fig. 6.13b, which reduces the degree of synchronization. This side effects of parameter variations and thermal noise can be suppressed by applying stronger RF bias to the injection locked SHE-STOs as shown in Fig. 6.13b-d. The plots show reduction in the jitter and the phase noise with increase in the amplitude of RF signal, thereby leading to stronger phase synchronization [159]-[161]. It can be explained that the stronger injected I_{RF} results in stronger locking strength and this global RF signal is not affected by the noise of individual magnet. However, higher RF amplitude may also cause higher reactive power. In this work, RF amplitude of 37.5μA is used in the associative module design.

6.5. Associative Computing Using Injection Locked SHE-STO Cluster

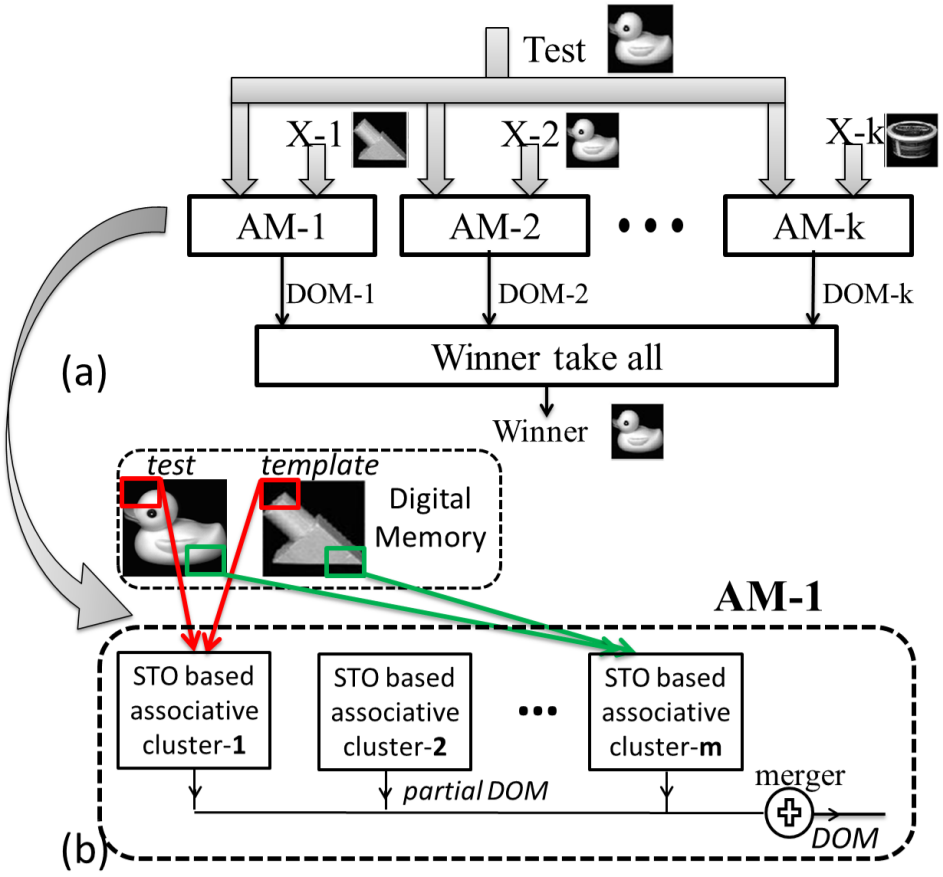


Fig. 6.14 (a) The architecture of associative computing for pattern matching, (b) the architecture of individual associative module design

The architecture of associative computing for pattern recognition is shown in Fig. 6.14a [143][176], [151]-[154]. An image data set consisting of *k* images are stored in the memory, and *k* parallel associative modules (AM) compute the degree of match (DOM) between the test image and each stored template image. The winner take all (WTA) circuit identifies the maximum DOM and outputs the winner index. The architecture of individual AM is shown in Fig. 6.14b, where the test and template images are partitioned into *m* fragments. Each STO based associative cluster takes the corresponding image fragments as inputs and computes the DOM between these two image fragments. The

outputs of individual STO associative clusters are combined through an analog merger to generate the overall DOM for the entire image.

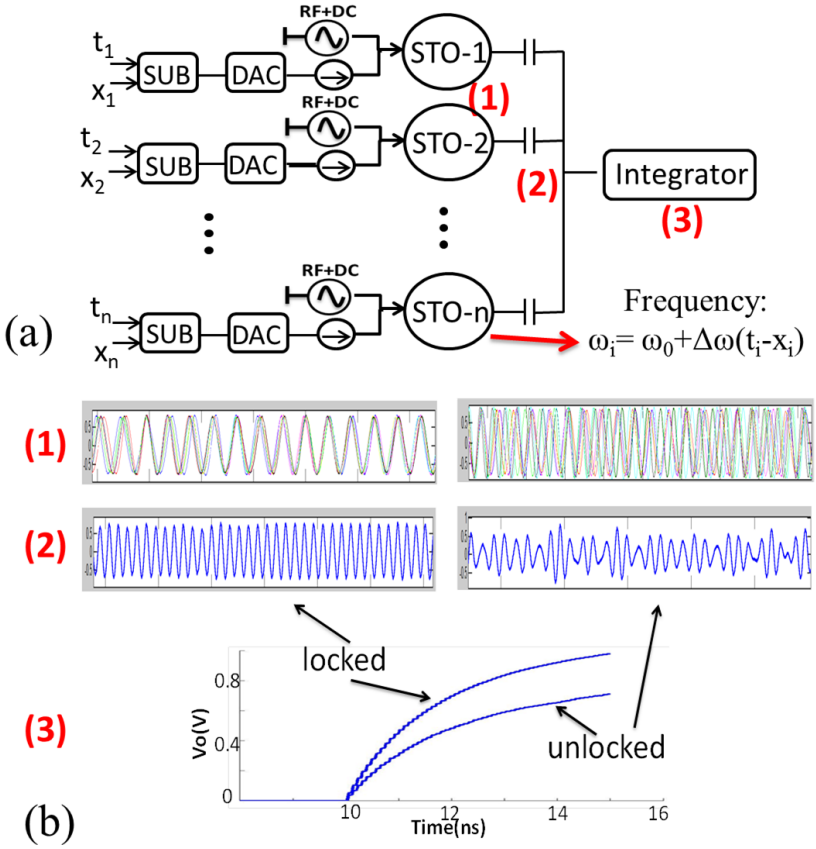


Fig. 6.15 (a) Circuit blocks of STO based associative cluster (b) transient simulation waveform of (1) STO outputs (2) capacitive addition outputs (3) integrator outputs

Fig. 6.15a shows the circuit blocks of STO based associative cluster using *frequency shift keying* [143], [151]-[154]. All the STOs are initially biased with the same DC and RF currents (DC+RF), which enforces phase locked oscillation of all the STOs in the cluster. To compute the associative matching between two vectors of *n* elements ($[t_1,\ldots,t_n]$ and $[x_1,\ldots,x_n]$, (t_i and x_i are digital values), a digital subtractor (SUB) computes the difference and a digital to analog converter (DAC) converts this difference into

analog DC current that can shift the frequency of STO. Thus, each STO frequency (ω_i) is shifted by the difference between the test (t_i) and template (x_i) vector elements

$$\omega_i = \omega_0 + \Delta\omega(t_i - x_i) \tag{6.15}$$

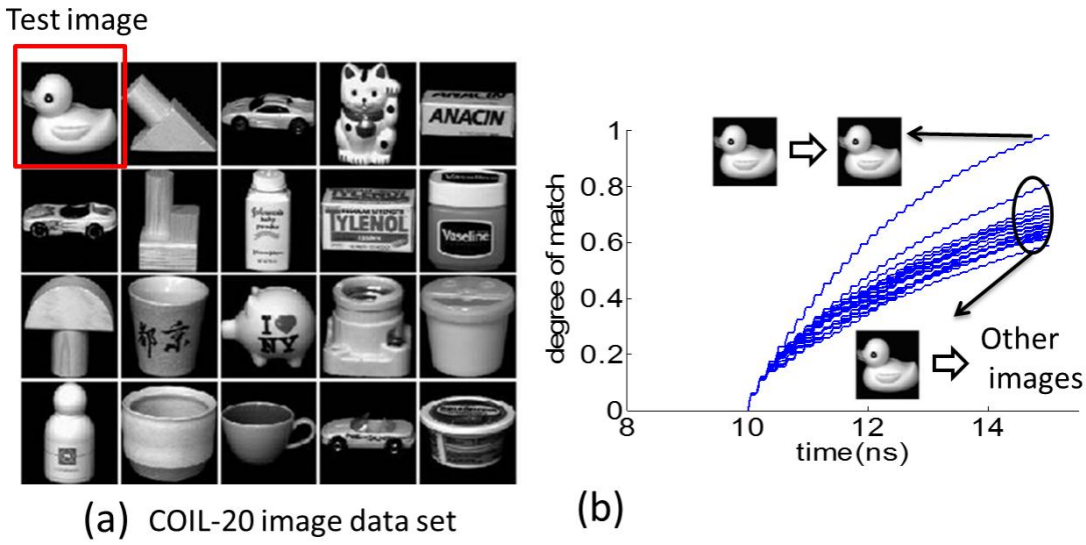


Fig. 6.16 (a) COIL-20 image data set [118] used in simulation: pixel values corresponding to the individual images were stored as 1-D analog templates, (b) merger outputs for a particular test (duck) image compared with all the other template images.

If the two vectors closely match each other, the inputs to the STOs are too small to bring them out of the locking state. The STOs, therefore, retain phase and frequency locking (Fig. 6.15b-(1) left). On the other hand, if the two vectors are significantly different, the inputs to the STOs are large in magnitude resulting in loss of locking (Fig. 6.15b-(1) right). The STO cluster circuit shown in Fig. 6.15a performs a capacitive summation of the individual STO waveforms and applies the sum to an integrator [143]. In the case of phase locked waveform, the summation results in a regular sinusoidal waveform which leads to fast charging of the integrator output (Fig. 6.15b-(2) & (3)). On the other hand, in the case of unlocked STO cluster, the summation is an irregular and low amplitude waveform, which leads to lower or negligible charging of the output (Fig.

6.15b-(2) & (3)). Thus, for a constant integration time, the DOM between a test vector and a template vector can be identified by comparing the integrator output voltage.

We apply the above mentioned architecture to pattern recognition application using COIL-20 image data set [118] (image compressed into 16×16 pixels, 5-bit grayscale). DC currents that are proportional to the elemental difference between the test and the template images are injected into the SHE-STOs (each cluster contains 8 injection locked SHE-STOs, totally 256/8 = 32 clusters). The integrator outputs (partial DOMs) of the SHE-STO clusters are summed and the result is the overall DOM. Higher value of the integrator (merger) output implies closer match and vice-versa. The merger output shown in Fig. 6.16b is for the case of a 'duck' image as input, which results in the template image for the 'duck' to be identified as the best match. The merger outputs of all other template images are significantly lower than the best matching template, as shown in the plot. Note that, the effects of parameter variations and thermal noise are not included in this plot. These effects are analyzed in the next subsection.

6.6. CMOS Interface Circuits and System Performance

In this section, we will present the design of CMOS interface circuitry for SHE-STO based AM and the energy analysis. The Monte-Carlo simulation of the implemented SHE-STO based AM design will also be discussed.

6.6.1. CMOS Interface Circuits Design

Fig. 6.15a shows the circuit block diagram for associative computing module with the coupled STO's as distance measuring block. The key CMOS circuit blocks consist of digital subtractor, DAC, integrator, analog merger and winner take all circuits. We will explain each circuit block and the SPICE simulation in the following subsections. Note that, all of the circuits are implemented and simulated in IBM 45nm technology.

6.6.1.1. Absolute Digital Subtractor

The test and template images are stored in memory as digital values. A digital subtractor is required to compute the elemental difference. We implemented a 5-bit

absolute digital subtractor consisting of a comparator and transmission gate logic based Brunt-Kung adder [174]. The simulation results are given in table-6.3.

6.6.1.2. Digital to Analog Converter (DAC)

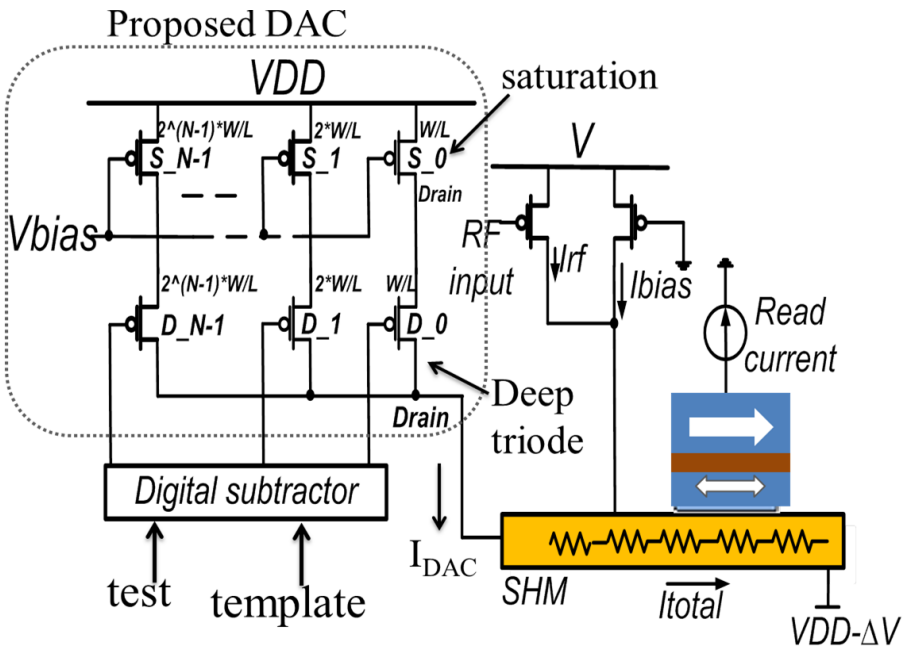


Fig. 6.17 Proposed DAC circuit for SHE-STO

Following digital subtraction, a DAC is used to convert the digital difference between the test and template images into analog current that acts as a DC input for generating a frequency shift in the STOs. In order to fully leverage the benefits of ultra-low power SHE-STO based AM, we propose a new DAC design as shown in Fig. 6.17. In our proposed DAC, two sets of binary weighted transistors are stacked, where the top transistors (S_N-1, …, S_0) operate in saturation region to provide a constant current and the bottom transistors (D_N-1, …, D_0) operate in the Deep Triode (DT) region. These DT transistors can be controlled by the binary inputs. In the stack, saturated transistor provides constant current flow and large channel length is used for accurate matching. The DT transistors control the speed of conversion and a small channel length is

preferred. Note that, voltage fluctuation at the drain terminals of DT transistors does not impact the DAC linearity because constant current is maintained by the saturated transistor (table-6.3). The total SHE-STO bias current can be expressed as:

$$I_{total} = I_{bias} + I_{RF} + I_{DAC} \tag{6.16}$$

where I_{total} is the total bias current for SHE-STO, I_{DAC} is the DAC output current corresponding to the elemental difference between the test and template image pixels. The proposed DAC performance is shown in table-6.3. Compared with conventional current steering DAC [175], our proposed DAC consumes ~25× lower energy as shown in table-6.3.

6.6.1.3. Integrator

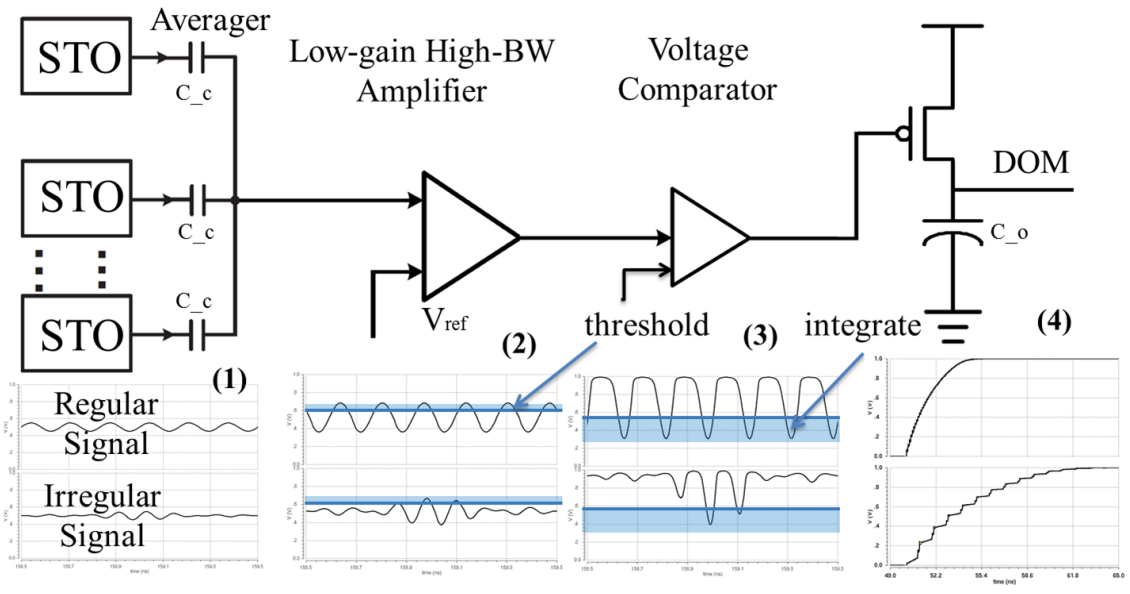


Fig. 6.18 Integrator circuit design and the transient waveforms. Note, regular signal corresponds to locked case. Irregular signal corresponds to unlocked case

Fig. 6.18 shows our circuit design of integrator and the transient simulated waveforms. This circuit performs a capacitive averaging of the individual STO waveforms. In the case of a good match between test and template images, the STO

cluster is locked and the “Averager” results in a regular (Fig. 6.18-(1)&(2)) sinusoidal waveform which leads to fast charging of the integrator output. On the other hand, if two images do not match, the STO cluster is unlocked and the “Averager” output is an irregular (Fig. 6.18-(1)&(2)) and low amplitude waveform which leads to lower or negligible charging of the output. In Fig. 6.18, a low-gain high-BW amplifier is used to amplify the oscillation signals to higher amplitude (~300mV in our simulation). Then, a voltage comparator generates pulsed signals with different on/off ratio based on the input waveform, as shown in Fig. 6.18-(3). The last circuit component of the integrator consists of a PMOS transistor and a charging capacitor. The “low” output voltage of previous stage turns on the PMOS transistor and charges the capacitor. Therefore, the regular oscillation signal charges the capacitor faster than the irregular oscillation waveform. Equivalently, for a constant integration time, the higher integrator output voltage corresponds to a higher DOM between the test and template images.

6.6.1.4. Analog Merger

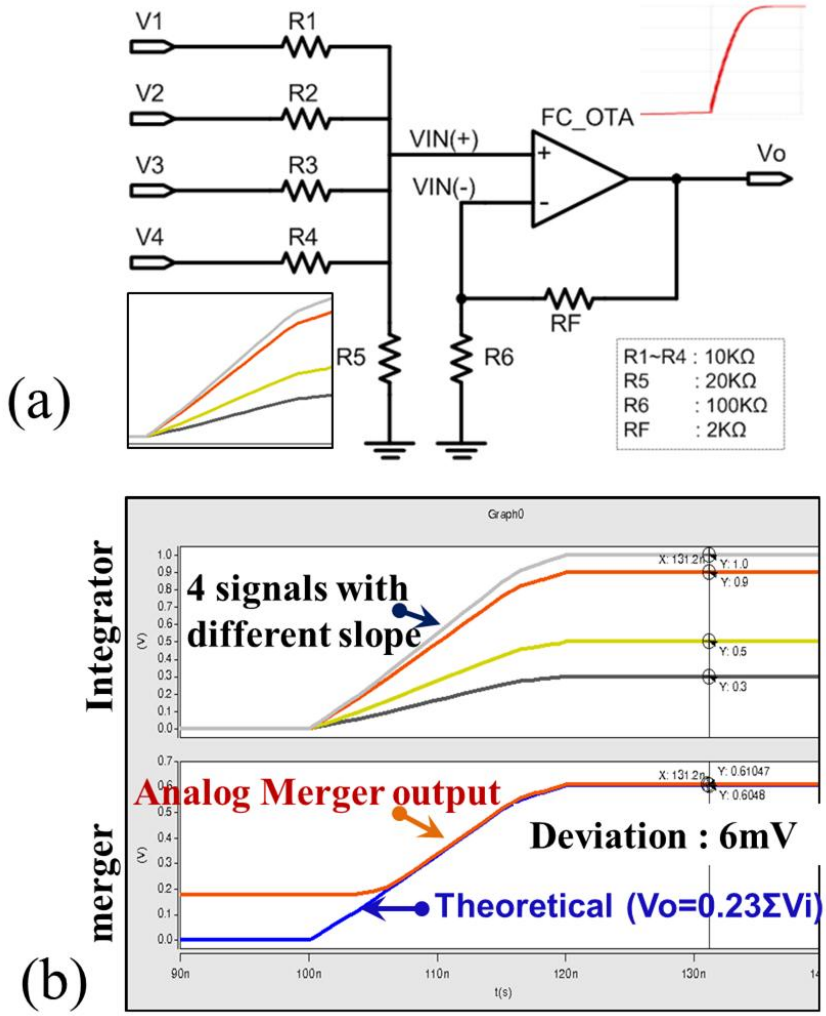


Fig. 6.19 (a) Analog merger circuit, (b) Simulation results

As we described in the previous section, an analog merger is required to combine the partial DOMs of individual STO clusters into an overall DOM for the entire image. Fig. 6.19a shows our circuit design of an analog merger based on a non-inverting summer with operational amplifier. For example, if 4 inputs (for simplicity) are assumed, the output voltage can be calculated as follows:

$$V_o = (\frac{V_1}{R_1} + \frac{V_2}{R_2} + \frac{V_3}{R_3} + \frac{V_4}{R_4})(R_1 \bullet R_2 \bullet R_3 \bullet R_4)(1 + \frac{R_F}{R_6}) \quad (6.17)$$

In our design, $R_1=R_2=R_3=R_4=10k\Omega$, $R_5=20k\Omega$, $R_6=100k\Omega$ and $R_F=2k\Omega$. Therefore, $Vo=0.23(V_1+V_2+V_3+V_4)$. The SPICE simulation of the proposed analog merger with 4 inputs is shown in Fig. 6.19b. It can be seen that the merger output matches well with the theoretical outputs ($V_o=0.23\Sigma V_i$). The performance of each of the interface circuit blocks are tabulated in table-6.3.

6.6.1.5. Winner Take All Circuit

A simple winner-take-all (WTA) circuit based on NOR gate described in [143] is employed in our work. As described in the previous subsection, the regular oscillation yields faster voltage rise, while irregular oscillation yields slower voltage rise. In the WTA circuit, all of the integrator outputs are connected with a NOR gate through buffers. When one of the integrator (merger) output voltage reaches the threshold voltage of the inverter in the buffer, it upsets the NOR circuit and stops the capacitor from charging further. The first upsetting inverter in the WTA circuit is identified as the winner.

Table. 6.3 CMOS interface circuit simulation results

SUB (5 bit)	Power	speed	Critical path		
	13.22μW	400MHz	185ps		
Current steering DAC	Power	speed	DNL	INL	FOM
	251μW	100MHz	0.24 LSB	0.49 LSB	2.51 pJ/conv
Proposed DAC	98 μW	1GSPS	0.14 LSB	0.27 LSB	98 fJ/conv
Integrator	Power	Integrate time	C_c	C_o	W_P
	400.8μW	5ns	50fF	150fF	654nm
Analog merger	Power	DC gain	f3dB	Phase margin	
	191μW	31dB	322 MHz	51^0	

Comments: All circuits are simulated in IBM 45 nm technology; voltage supply=1V; Both DACs are 5 bit. DNL: differential non-linearity; INL: integral non-linearity; FOM: figure of merit; C_c: coupling capacitor; C_o: charging capacitor; W_P: PMOS width;

6.6.2. System Performance and Variation Analysis of SHE-STO based AM

Based on the simulation of each circuit block shown in table-6.3, the total energy consumption of one single injection locked SHE-STO based AM is 259pJ as listed in table-6.4. It can be seen that the total energy consumption of AM based on proposed DAC can achieve more than 3× lower than that of AM based on conventional current steering DAC.

Table. 6.4 Energy analysis of associative module

Element	Energy for one	Number of units in each AM	Energy per AM
subtractor	33.05fJ	256	8.46pJ
CS-DAC[1]	2.5pJ	256	640pJ
P-DAC[2]	98fJ	256	25.09pJ
SHE-STO	627fJ	256	160.5pJ
integrator	2pJ	32	64pJ
merger	1pJ	1	1pJ
Total-1[3]	**874pJ**	**Total-2[4]**	**259pJ**

Comments: image size is 16×16 pixels, 5-bit grayscale, each STO cluster contains 8 STOs. WTA circuit is shared by all of the AMs, it is not included here. Integration time is 5ns

1: CS-DAC is 5-bit current steering DAC

2: P-DAC is our proposed 5-bit DAC

3: total-1 is the total energy consumption of AM based on current steering DAC

4: total-2 is the total energy consumption of AM based on proposed DAC

Fig. 6.20a shows the normalized outputs of SHE-STO based AM for all 20 patterns shown in Fig. 6.16a. Pixel-(*i, j*) indicates the SHE-STO AM output when j^{th} pattern compared with i^{th} pattern. It can be seen that the value of pixel-(*i, i*) is the maximum in i^{th} row (i.e. i^{th} pattern compares with itself, *i*=1,2,..,20), which indicates a correct match. In this work, we define the *detection margin* as (DOM(1^{st})-DOM(2^{nd}))/DOM(1^{st}), where DOM(1^{st}) is the best DOM and DOM(2^{nd}) is the second best DOM. A larger detection margin is required to maintain a high recognition accuracy under device parameter variations, thermal noise and interface circuit variations. Fig. 6.20b depicts the detection margin for all 20 patterns. It can be seen that most of the detection margin (except pattern #3, 6 and 19) are above ~10%, which can be easily detected by WTA circuit. The

detection margins of pattern #3, 6 and 19 are relatively small (~5%) due to the fact that these three patterns are very close to each other as shown in Fig. 6.20b.

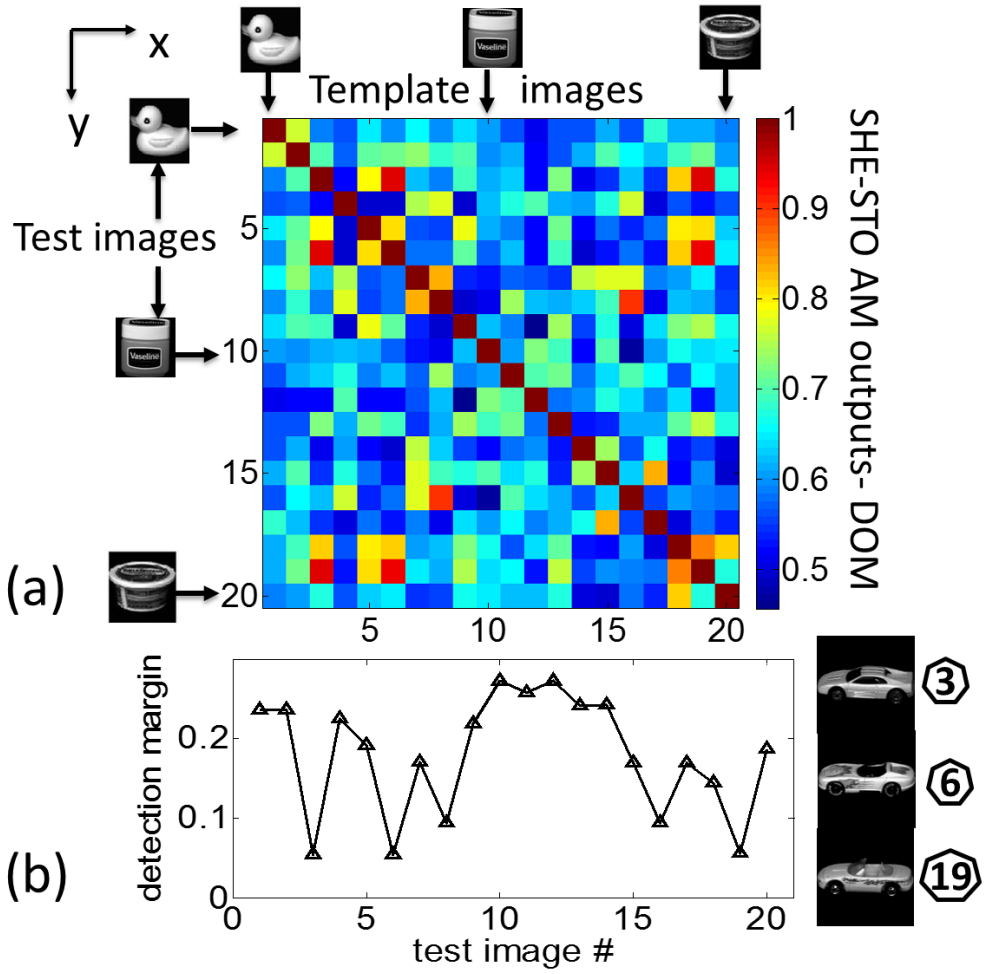


Fig. 6.20 (a) Normalized outputs of SHE-STO based AM for all 20 patterns shown in Fig. 6.16a. Note that, pixel (i, j) indicates the SHE-STO AM output when ith pattern compared with jth pattern (b) detection margin for all 20 patterns. Pattern #3, 6 and 19 are shown in the right. Note that detection margin=(DOM(1st)-DOM(2nd))/DOM(1st)

In order to analyze the effects of device variations, thermal noise and interface circuit variations on the detection margin, we have carried out Monte-Carlo simulation (100 simulation runs). During Monte-Carlo simulation, the Gaussian distributed (σ=5%) variations are added on STO physical parameters (damping factor, saturation

magnetization) and the thermal effects (room temperature, 300K) are modeled using a randomly fluctuating field drawn from a Gaussian distribution of zero mean and standard deviation of $\sqrt{2\alpha K_B T/\gamma M_s V \Delta t}$ [127]. The interface circuit variations (including transistor size, capacitance, etc.) are also included in the Monte-Carlo simulation. Fig. 6.21 shows the comparison of the AM outputs without and with variations. The test image is the 'duck' image, which compares with the 20 template images shown in Fig. 6.16a. For simplicity, we only show the best match (blue line) and second best match (red line) cases. It can be seen that the detection margin is reduced from ~20% (without variations) to ~16% (with variations, worst case). The reduction of detecion margin is also observed in the simulations using other patterns as test images. 17 out of 20 patterns (except pattern #3, 6 and 19) can be correctly identified in Monte-Carlo simulations (100 simultaion runs). Our results indicate the injection locked SHE-STO based AM is relatively immune to interface circuitry variations, device parameter variations and thermal noise.

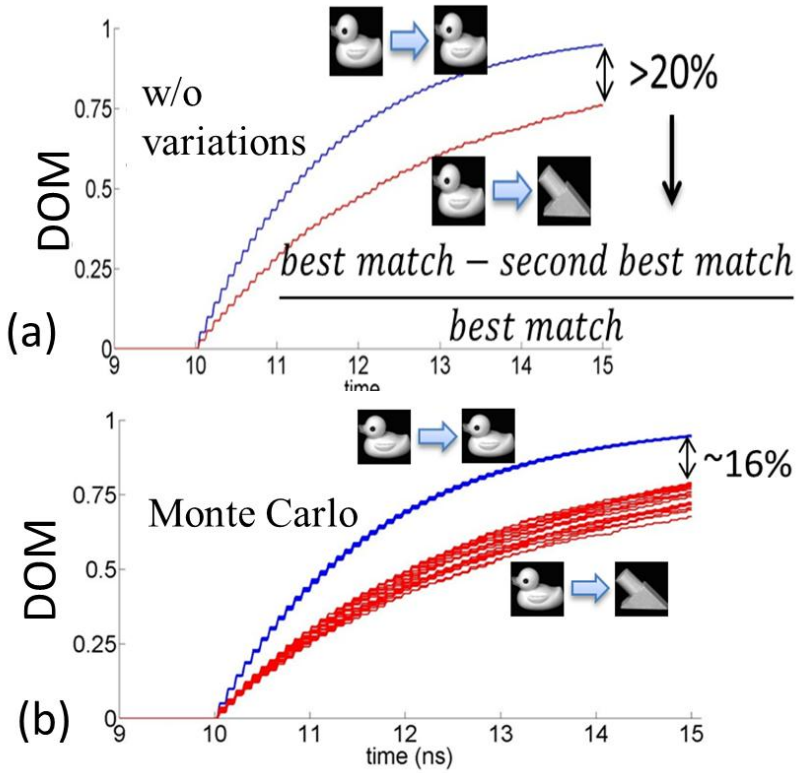


Fig. 6.21 transient AM output (a) without variation, (b) Monte-Carlo simulation on interface circuits, device parameters and thermal noise. Note that, only the best match and second best match outputs are shown for simplicity. Blue line is the best match, and red line is the second best match.

6.7. Summary

We proposed a variation tolerant injection locked Spin Hall induced oscillator array for associative computing. The numerical simulation framework for injection locked SHE-STO cluster was described and the results show robust oscillations under parameter variations and thermal noise. Our results show that the proposed system architecture with coupled SHE-STOs along with associated CMOS interface circuitries can be suitable for robust and energy efficient associative computing and pattern matching.

7. SUMMARY

Spin-transfer torque devices are unlikely to be drop-in replacements for CMOS. They may be integrated with CMOS and other charge based devices to model energy efficient computing systems. In this dissertation, we have explored new Boolean and brain-inspired computing models that are inherently suited to the characteristics of STT devices, thereby attaining performance that CMOS cannot achieve.

First, we show that non-volatile STT devices can be combined with CMOS compatible memristors for designing energy efficient configurable Boolean threshold logic gate. In such a design, the memristive cross-bar array is used to implement current mode summation of binary inputs, and the second step of threshold logic - thresholding operation is performed by the ultra-low power spintronic threshold device. The proposed field programmable spin-memristor threshold logic gate arrays can work at a small terminal voltage of ~50mV, leading to ultra-low power dissipation in both gates and programmable interconnect networks. Such hardware can achieve more than 100× improvement in energy and 1000× improvement in energy-delay product, as compared to state of the art CMOS FPGA based TLG.

Next, for brain-inspired computing, we have exploited different spin-transfer torque device structures that can implement the hard-limiting and soft-limiting artificial neuron transfer functions respectively. As cases studies, we apply these STT based neurons (‘spin-neuron’) in various neural network architectures, such as hierarchical temporal memory and feed-forward neural network, for performing “human-like” cognitive computing. In hierarchical temporal memory design, the low power, current mode spin-neurons combined with MCA are explored in the dot product based pattern matching, which is the core computing block in the design of HTM hardware. Such a direct mapping of the core-computing primitive of the cortical computing system can be very

attractive for large-scale and energy efficient design. The simulated spin based HTM computing block results in ~200× lower energy consumption compared to the CMOS based HTM node design.

However, in brain-inspired computing, soft-limiting neurons are greatly preferred to hard-limiting neurons due to their much improved modeling capacity, which leads to higher network accuracy and lower network complexity. Thus, we propose a domain wall motion based STT device that can efficiently implement a neuron with a soft-limiting non-linear transfer function, operating at ultra-low supply voltage and current. The spin based neuron device allows the peripheral circuits and memristive cross-bar array synapses to also operate at very low voltages, thereby leading to ultra-low power consumption for the whole system. This proposed soft-limiting spin-neuron is used to design artificial neural networks that show more than two orders of magnitude lower energy dissipation compared with analog and digital CMOS ANN implementations in 45nm CMOS technology and ~2.5× lower hidden layer area compared with hard-limiting neuron based ANNs. Moreover, the proposed spin-transfer torque based soft-limiting non-linear neurons along with MCA-synapses can be used to build large scale energy efficient neuromorphic computing hardware for cognitive computing applications.

In the final part of the dissertation, we discuss the numerical device simulation framework for spin-torque oscillators and different coupling mechanisms for an STO array, including magnetic coupling, electrical coupling and injection locking. We show the dynamics of coupled spin-torque oscillators array can be exploited to estimate multi-dimensional distance metric for associative computing, image and video analysis, etc. We also presented an application of injection locked spin hall induced oscillators for associative computing as a case study. Our results show that the proposed system architecture with coupled SHE-STOs and the associated CMOS interface circuitries can be suitable for robust and energy efficient associative computing/ pattern matching.

LIST OF REFERENCES

LIST OF REFERENCES

[1] J. Slonczewski, “Current-driven excitation of magnetic multi-layers”, *J. of Magn. and Magn. Mater.*, vol. 159, no. 1-2, pp. L1–L7, Jun. 1996.

[2] L. Berger, “Emission of spin waves by a magnetic multilayer traversed by a current”, *Phys. Rev. B*, vol. 54, no. 13, pp. 9353–9358, Oct. 1996.

[3] E. Y. Tsymbal and D. G. Pettifor, "Perspectives of giant magneto-resistance", *Solid State Phys. – Adv. Res. Appl*., vol. 56, pp. 113–237, 2001.

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

[5] Y. Huai, F. Albert, P. Nguyen, M. Pakala, and T. Valet, "Observation of spin-transfer switching in deep submicron-sized and low-resistance magnetic tunnel junctions", *Appl. Phys. Lett*., vol. 84, no. 16, p. 3118, 2004

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

[7] M. N. Baibich, J. M. Broto, A. Fert, F. N. Van Dau, and F. Petroff, "Giant Magnetoresistance of (001)Fe/(001)Cr Magnetic Superlattices", *Phys. Rev. Lett*., vol. 61, no. 21, pp. 2472–2475, Nov. 1988.

[8] K. K. Bhuwalka, S. Sedlmaier, A. K. Ludsteck, C. Tolksdorf, J. Schulze, and I. Eisele. "Vertical tunnel field-effect transistor", *IEEE Transactions* on *Electron Devices, vol.* 51, no. 2, pp. 279-282, 2004

[9] K. Boucart and A. M. Ionescu. "Double-gate tunnel FET with high-κ gate dielectric", *IEEE Transactions on Electron Devices,* vol. 54, no. 7, pp.1725-1733, 2007

[10] G. Autès, J. Mathon, and a. Umerski, "Strong Enhancement of the Tunneling Magnetoresistance by Electron Filtering in an Fe/MgO/Fe/GaAs(001) Junction", *Phys. Rev. Lett*., vol. 104, no. 21, p. 217202, May 2010

[11] P. Mavropoulos, M. Leˇzaíc, and S. Blügel, "Half-metallic ferro-magnets for magnetic tunnel junctions by ab initio calculations", *Phys. Rev. B,* vol. 72, no. 17, p. 174428, Nov. 2005.

[12] M. Bowen, M. Bibes, a. Barthelemy, J.P. Contour, a. Anane, Y. Lematre, and a. Fert, "Nearly total spin polarization in La2/3Sr1/3MnO3 from tunneling experiments," *Appl. Phys. Lett*., vol. 82, no. 2, p. 233, 2003

[13] 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. of Solid-State Circuits*, vol. 43, no. 1, pp. 109–120, Jan. 2008.

[14] K. Ono, T. Kawahara, R. Takemura, K. Miura, H. Yamamoto, M. Yamanouchi, J. Hayakawa, K. Ito, H. Takahashi, S. Ikeda, H. Hasegawa, H. Matsuoka, and H. Ohno, "A disturbance-free read scheme and a compact stochastic-spin-dynamics-based MTJ circuit model for Gbscale SPRAM", in 2009 *IEEE Int. Electron Devices Meeting (IEDM)*. IEEE, Dec. 2009, pp. 1–4.

[15] C. Lin, S. Kang, Y. Wang, K. Lee, X. Zhu, W. Chen, X. Li, W. Hsu, Y. Kao, M. Liu, M. Nowak, and N. Yu, "45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell," in *2009 IEEE Int. Electron Devices Meeting (IEDM)*. IEEE, Baltimore, MD, pp. 1–4, Dec. 2009,

[16] Y. M. Lee, C. Yoshida, K. Tsunoda, S. Umehara, M. Aoki, and T. Sugii, "Highly scalable STT-MRAM with MTJs of top-pinned structure in 1T/1MTJ cell", in *Proc. Symp. on VLSI Technol.*, Honolulu, Jun. 2010, pp. 49–50.

[17] A. Driskill-Smith, S. Watts, D. Apalkov, D. Druist, X. Tang, Z. Diao, X. Luo, A. Ong, V. Nikitin, and E. Chen, "Non-volatile Spin-Transfer Torque RAM (STT-RAM): An analysis of chip data, thermal stability and scalability", in 2010 *IEEE Int. Memory Workshop (IMW*), vol. 1, no. 408, pp. 1–3, 2010

[18] T. Kawahara, "Challenges toward gigabit-scale spin-transfer torque random access memory and beyond for normally off, green information technology infrastructure (Invited)", *J. of Appl. Phys*., vol. 109, no. 7, 07D325, 2011.

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

[20] B. Behin-Aein, A. Sarkar, S. Srinivasan, and S. Datta, "Switching energy-delay of all spin logic devices", *Appl. Phys. Lett*. vol. 98, iss. 12, pp. 123510, Mar. 2011.

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

[22] 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. of IEEE/ACM Int. Symp. Nanoscale Arch*., San Diego, CA, Jun. 2011, pp. 129-136.

[23] J. Z. Sun, "Spin-current interaction with a monodomain magnetic body: a model study", *Phys. Rev. B* vol. 62, pp. 570--578, 2000.

[24] M. Sharad, K. Yogendra, K. Kwon, and K. Roy, "Design of ultra high density and low power computational blocks using nano-magnets", *In* IEEE *14th International Symposium on Quality Electronic Design*, Santa Clara, CA, pp. 223-230, Mar. 2013.

[25] Charles Augustine, "Spintronic Memory and Logic: From Atoms to Systems", Ph.D. dissertation, Purdue University, 2011

[26] B. Behin‐Aein, S. Salahuddin and S. Datta, "Switching energy of ferromagnetic logic bits", *IEEE Trans. Nanotech*., vol. 8, no. 4, pp.505‐514, 2009.

[27] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L. Wang and Y. Huai, "Spin-transfer torque switching in magnetic tunnel junctions and spin‐transfer torque random access memory," *Journal of Physics* : Condensed Matter, vol. 19, no. 16, pp. 165209 (1‐13), April, 2007.

[28] S. Salahuddin and S. Datta, "Interacting systems for self-correcting low power switching", *Appl. Phys. Lett*., vol. 90, pp. 093503(1‐3), 2007.

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

[30] T. Miyazaki, T. Yaoi, and S. Ishio, "Large magnetoresistance effect in 82ni-fe/al-al 2 o 3/co magnetic tunneling junction", *Journal of magnetism and magnetic materials*, vol. 98, no. 1, pp. L7–L9, 1991.

[31] T. T. Miyazaki and N. Tezuka, "Giant magnetic tunneling effect in Fe/AI₂O₃/Fe junction", *Journal of Magnetism and Magnetic Materials*, vol. 139, no. 3, pp. L231–L234, 1995.

[32] W. Butler, X.-G. Zhang, T. Schulthess, and J. MacLaren, "Spin dependent tunneling conductance of fe-mgo-fe sandwiches," *Physical Review B*, vol. 63, no. 5, p. 054416, 2001.

[33] S. S. 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 materials*, vol. 3, no. 12, pp. 862–867, 2004.

[34] J.-i. Inoue, "Gmr, tmr, bmr, and related phenomena", *Nanomagnetism and Spintronics*, p. 15, 2013.

[35] T. Sasaki, T. Suzuki, Y. Ando, H. Koike, T. Oikawa, Y. Suzuki, and M. Shiraishi, "Local magnetoresistance in fe/mgo/si lateral spin valve at room temperature", *Applied Physics Letters*, vol. 104, no. 5, p. 052404, 2014.

[36] Y. Ji, A. Hoffmann, J. Jiang, and S. Bader, "Spin injection, diffusion, and detection in lateral spin-valves", *Applied physics letters*, vol. 85, no. 25, pp. 6218–6220, 2004.

[37] Y. Fukuma, L. Wang, H. Idzuchi, S. Takahashi, S. Maekawa, and Y. Otani, "Giant enhancement of spin accumulation and long-distance spin precession in metallic lateral spin valves", *Nature materials*, vol. 10, no. 7, pp. 527–531, 2011.

[38] C.-C. Lin, A. V. Penumatcha, Y. Gao, V. Q. Diep, J. Appenzeller, and Z. Chen, "Spin transfer torque in a graphene lateral spin valve assisted by an external magnetic field", *Nano letters*, vol. 13, no. 11, pp. 5177– 5181, 2013.

[39] J. Samm, J. Gramich, A. Baumgartner, M. Weiss, and C. Sch¨onenberger, "Optimized fabrication and characterization of carbon nanotube spin valves", *Journal of Applied Physics*, vol. 115, no. 17, p. 174309, 2014.

[40] G. Beach, M. Tsoi, and J. Erskine, "Current-induced domain wall motion", Journal of magnetism and magnetic materils, vol. 320, no.7, pp.1272-1281, 2008

[41] T. Valet and A. Fert, "Theory of the perpendicular magnetoresistance in magnetic multilayers", *Physical Review B*, vol. 48, no. 10, p. 7099, 1993.

[42] T. Kimura, T. Sato, and Y. Otani, "Temperature evolution of spin relaxation in a nife/cu lateral spin valve", *Physical review letters*, vol. 100, no. 6, p. 066602, 2008.

[43] H. Aurich, A. Baumgartner, F. Freitag, A. Eichler, J. Trbovic, and C. Sch¨onenberger, "Permalloy-based carbon nanotube spin-valve", *Applied Physics Letters*, vol. 97, no. 15, p. 153116, 2010.

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

[45] K.-S. Ryu, L. Thomas, S.-H. Yang, and S. Parkin, "Chiral spin torque at magnetic domain walls", *Nature nanotechnology*, vol. 8, no. 7, pp. 527–533, 2013.

[46] S. Emori, U. Bauer, S.-M. Ahn, E. Martinez, and G. S. Beach, "Current driven dynamics of chiral ferromagnetic domain walls", *Nature materials*, vol. 12, no. 7, pp. 611–616, 2013.

[47] I. M. Miron, T. Moore, H. Szambolics, L. D. Buda-Prejbeanu, S. Auffret, B. Rodmacq, S. Pizzini, J. Vogel, M. Bonfim, A. Schuhl et al., "Fast current-induced domain-wall motion controlled by the rashba effect", *Nature Materials*, vol. 10, no. 6, pp. 419–423, 2011.

[48] L. Liu, O. Lee, T. Gudmundsen, D. Ralph, and R. Buhrman, "Currentinduced switching of perpendicularly magnetized magnetic layers using spin torque from the spin hall effect", *Physical review letters*, vol. 109, no. 9, p. 096602, 2012.

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

[50] A. Brataas and K. M. Hals, "Spin-orbit torques in action", *Nature nanotechnology*, vol. 9, no. 2, pp. 86–88, 2014.

[51] G. Finocchio, M. Carpentieri, E. Martinez, and B. Azzerboni, "Switching of a single ferromagnetic layer driven by spin hall effect", *Applied Physics Letters*, vol. 102, no. 21, p. 212410, 2013.

[52] G. Yu, P. Upadhyaya, Y. Fan, J. G. Alzate, W. Jiang, K. L. Wong, S. Takei, S. A. Bender, L.-T. Chang, Y. Jiang et al., "Switching of perpendicular magnetization by spin-orbit torques in the absence of external magnetic fields", *Nature nanotechnology*, vol. 9, pp. 548-554, May 2014.

[53] I. M. Miron, G. Gaudin, S. Auffret, B. Rodmacq, A. Schuhl, S. Pizzini, J. Vogel, and P. Gambardella, "Current-driven spin torque induced by the rashba effect in a ferromagnetic metal layer", *Nature materials*, vol. 9, no. 3, pp. 230–234, 2010.

[54] T. Suzuki, S. Fukami, N. Ishiwata, M. Yamanouchi, S. Ikeda, N. Kasai, and H. Ohno, "Current-induced effective field in perpendicularly magnetized ta/cofeb/mgo wire", *Applied Physics Letters*, vol. 98, no. 14, p. 142505, 2011.

[55] X. Fan, J. Wu, Y. Chen, M. J. Jerry, H. Zhang, and J. Q. Xiao, "Observation of the nonlocal spin-orbital effective field", *Nature communications*, vol. 4, p. 1799, 2013.

[56] P. Haazen, E. Mur`e, J. Franken, R. Lavrijsen, H. Swagten, and B. Koopmans, "Domain wall depinning governed by the spin hall effect", *Nature materials*, vol. 12, no. 4, pp. 299–303, 2013.

[57] R. Liu, W. Lim, and S. Urazhdin, "Spectral characteristics of the microwave emission by the spin hall nano-oscillator", *Physical Review Letters*, vol. 110, no. 14, p. 147601, 2013.

[58] V. E. Demidov, S. Urazhdin, H. Ulrichs, V. Tiberkevich, A. Slavin, D. Baither, G. Schmitz, and S. O. Demokritov, "Magnetic nanooscillator driven by pure spin current", *Nature materials*, vol. 11, no. 12, pp. 1028–1031, 2012.

[59] Z. Zeng, G. Finocchio, B. Zhang, P. 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," Scientific Reports, vol. 3, 1426, Mar. 2013

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

[61] A. Hoffmann, "Spin hall effects in metals", *IEEE Transactions on Magnetics*, vol. 49, pp. 5172–5193, 2013.

[62] L. Liu, T. Moriyama, D. Ralph, and R. Buhrman, "Spin-torque ferromagnetic resonance induced by the spin hall effect", *Physical review letters*, vol. 106, no. 3, p. 036601, 2011.

[63] H. Ulrichs, V. Demidov, S. Demokritov, W. Lim, J. Melander, N. Ebrahim-Zadeh, and S. Urazhdin, "Optimization of pt-based spin hall-effect spintronic devices", *Applied Physics Letters*, vol. 102, no. 13, p. 132402, 2013.

[64] D. Bhowmik, L. You, and S. Salahuddin, "Spin hall effect clocking of nanomagnetic logic without a magnetic field", *Nature nanotechnology*, vol. 9, pp. 59-63, Nov. 2013.

[65] C.-F. Pai, L. Liu, Y. Li, H. Tseng, D. Ralph, and R. Buhrman, "Spin transfer torque devices utilizing the giant spin hall effect of tungsten", *Applied Physics Letters*, vol. 101, no. 12, p. 122404, 2012.

[66] Y. Niimi, Y. Kawanishi, D. Wei, C. Deranlot, H. Yang, M. Chshiev, T. Valet, A. Fert, and Y. Otani, "Giant spin hall effect induced by skew scattering from bismuth

impurities inside thin film cubi alloys", *Physical review letters*, vol. 109, no. 15, p. 156602, 2012.

[67] S.H. Jo, K-H Kim, and W. Lu. "High-density crossbar arrays based on a Si memristive system." *Nano letters*, vol. 9, no. 2, pp.870-874, Jan. 2009

[68] S. Jo and W. Lu, "CMOS Compatible Nanoscale Nonvolatile Resistance Switching Memory", *Nano Letters*, vol. 8, no. 2, pp. 392-397, Jan. 2008.

[69] L. Gao, F. Alibart, and D. Strukov, "Analog-input analog-weight dot-product operation with Ag/a-Si/Pt memristive devices", in *Proceedings of the 20th International Conference on VLSI and System-on-Chip (VLSI-SoC)*, Santa Cruz, CA, Oct. 2012, pp. 88-93.

[70] S. Shin, K. Kim, and S. Kang, "Memristor-based fine resolution programmable resistance and its applications", in *Proceedings of the IEEE International Conference on Communications, Circuits and Systems*, Milpitas, CA, July 2009, pp. 948-951.

[71] K. Likharev, "Biologically inspired computing in CMOL CrossNets", in *Proceedings of the AAAI Fall Symposium Series*. Arlington, VA, Nov. 2009, pp. 90.

[72] Ö. Türel, J. Lee, X. Ma and K. Likharev, "Neuromorphic architectures for nanoelectronic circuits", *International Journal Circuits Theory and Application*, vol. 32, no. 5, pp. 277-302, Sept. 2004.

[73] J. Rajendra, H. Manem, R. Karri and G.S. Rose, "An Energy-Efficient Memristive Threshold Logic Circuit." *IEEE Trans. On Computers*, vol. 61, No. 4, pp. 474-487, Apr. 2012

[74] T. Tran, A. Rothenbuhler, E.H.B. Smith, and V. Saxena, "Reconfigurable Threshold Logic Gates Using Memristive Devices." *IEEE SubVT*, Waltham, MA, Oct. 2012, pp.1-3

[75] D. Chabi, W. Zhao, D. Querlioz and J.O. Klein, "Robust neural logic block (NLB) based on memristor cross-bar array." IEEE/ACM international symposium on Nanoscale Architecture, San Diego, CA, June 2011, pp. 137-143

[76] L.Gao, F. Alibart, and D.B. Strukov, "Programmable CMOS/Memristor Threshold Logic." *IEEE Trans. Nanotech*, vol. 12, no.2, pp.115-119, Mar. 2012.

[77] C.K. Lim, T. Devolder, C. Chappert, J. Grollier, V. Cros, A. Vaurès, A. Fert and G. Faini, "Domain wall displacement induced by subnanosecond pulsed current", *App. Phy. Lett*., vol. 84, no. 15, pp.2820-2822, Apr. 2004

[78] D-T. Ngo, N. Watanabe and H. Awano, “CoB/Ni-based multilayer nanowire with high-speed domain wall motion under low current control”, *Japnese Journal of Applied Physics,* vol. 51, no. 9R, pp.093002, Sept. 2011

[79] S. Fukami, 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, T. Sugibayashi, "Low-current perpendicular domain wall motion cell for scalable high-speed MRAM," *VLSI Tech. Symp*, Honolulu, HI, June 2009, pp. 230-231

[80] A.V. Khvalkovskiy, V. Cros, D. Apalkov, V. Nikitin, M. Krounbi, K. A. Zvezdin, A. Anane, J. Grollier, and A. Fert, "Matching domain-wall configuration and spin-orbit torques for efficient domain-wall motion", *Physical Review B* , vol. 87, no. 2, pp. 020402, Jan. 2013

[81] D. Morris, D. Bromberg, J.G. Zhu, and L. Pileggi, "mLogic: Ultra-low voltage non-volatile logic circuits using STT-MTJ devices", In Proceedings of the 49th *Annual Design Automation Conference*, San Francisco, CA, June 2012, pp. 486-491.

[82] M. Sharad, C. Augustine, and K. Roy, “Boolean and non-Boolean Computing Using Spin Devices”, *IEDM*, San Francisco, CA, Dec. 2012, pp. 11.61.1-11.6.4

[83] M. Sharad, D. Fan, and K. Roy, “Ultra Low Power Associative Computing with Spin Neurons and Resistive Crossbar Memory,” *IEEE Design Automation Conference*, Austin, TX, May 2013, pp. 1-6

[84] R. Zhang, P. Gupta, L. Zhong, and N.K. Jha, "Synthesis and optimization of threshold logic networks with application to nanotechnologies." *Design, Automation and Test in Europe Conference and Exhibition,* Feb. 2008, pp. 904-909

[85] Y. Lin, F. Li, and L. He, "Routing track duplication with fine-grained power-gating for FPGA interconnect power reduction." *ASPDAC*, Jan. 2005, pp. 645-650

[86] J.A. Currivan, Y. Jang ; M.D. Mascaro, M.A. Baldo, et al., "Low Energy Magnetic Domain Wall Logic in Short, Narrow Ferromagnetic Wires", *IEEE Mag. Lett*., vol. 3, Apr. 2012

[87] Q. Xia, W. Robinett, M.W. Cumbie, N. Banerjee, T.J. Cardinali, J.J. Yang, W. Wu, X. Li, W.M. Tong, D.B. Strukov, G.S. Snider, G.Medeiros-Ribeiro and R.S. Williams, "Memristor−CMOS hybrid integrated circuits for reconfigurable logic." *Nano letters* , vol.9, no.10 pp.3640-3645, Sept. 2009

[88] P.J. Metaxas, J. Sampaio, A. Chanthbouala, et al. "High domain wall velocities via spin transfer torque using vertical current injection." *Scientific reports* vol. 3, May 2013.

[89] C. Augustine, A. Raychowdhury, B. Behin-Aein, S. Srinivasan, J. Tschanz, V.K. De, and K. Roy, "Numerical Analysis of Domain Wall Propagation for Dense Memory Arrays", *IEDM*, Washington, DC, Dec. 2011, pp17.6.1-17.6.4

[90] H. Manem, and G.S. Rose. "A read-monitored write circuit for 1T1M multi-level memristor memories." *In Circuits and systems (ISCAS), 2011 IEEE international symposium on*, Rio deJaneiro, May 2011, pp. 2938-2941

[91] C-M. Jung, J-M. Choi, and K-S. Min, "Two-Step Write Scheme for Reducing Sneak-Path Leakage in Complementary Memristor Array", *IEEE Trans. on Nanotech*, vol. 1. No.3, pp. 611-618, May 2012.

[92] K-H. Kim, G. Siddharth, D. Wheeler, J.M. Cruz-Albrecht, T. Hussain, N. Srinivasa, and W. Lu. "A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications." *Nano letters* vol. 12, no. 1, pp.389-395, 2011

[93] S. Fukami, M. Yamanouchi, T. Koyama, K. Ueda, et. al., "High-Speed and Reliable Domain Wall Motion Devices: Material Design for Embedded Memory and Logic Application", *VLSI Technology Digest of Technical papers*, Honolulu, HI, June 2012, pp. 61-62

[94] G. Zhang, Q. Zhang, C-T. Bui, G.Q. Lo, and B. Li, "Thermoelectric performance of silicon nanowires", *Applied Physics Letter*, vol. 94, no. 21, pp. 213108, 2009

[95] http://math.nist.gov/oommf/

[96] D.C. Van Essen, CH. Anderson, and DJ. Felleman, "Information Processing in the Primate Visual System: An Integrated Systems Perspective", *Science*, vol. 255, pp. 419-423, Jan. 1992

[97] R. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, vol. 4, pp. 4-22, Apr. 1987.

[98] P. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proceedings of the 12th IEEE International Conference on Document Analysis and Recognition*, Edinburgh, UK, Aug. 2003, pp. 958-958.

[99] D. George and J. Hawkins, "A hierarchical Bayesian model of invariant pattern recognition in the visual cortex," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, Montreal, QC, Canada, Aug. 2005, pp. 1812-1817.

[100] D. George and J. Hawkins, "Towards a Mathematical Theory of Cortical Micro-circuits," *PLoS Computational Biology*, vol. 5, issue. 10, pp. 1-26, Oct. 2009.

[101] D. Maltoni, "Pattern Recognition by Hierarchical Temporal Memory," University of Bologna, Bologna, Italy, Technical Report, Apr. 2011.

[102] W. Melis, S. Chizuwa, and M. Kameyama, "Evaluation of the hierarchical temporal memory as soft computing platform and its VLSI architecture," in *Proceedings of the 39th International Symposium on Multiple-Valued Logic*, Naha, Okinawa, May 2009, pp. 233-238.

[103] J. Hawkins, S. Ahmad, and D. Dubinsky, "Hierarchical temporal memory including HTM cortical learning algorithms," Numenta. Inc. Palto Alto, CA, Tech. Rep., 2010.

[104] R. Berdan, T. Prodromakis, and C. Toumazou, "High precision analogue memristor state tuning," *Electronics Letters*, vol. 48, pp. 1105-1107, Aug. 2012.

[105] F. Miao, W. Yi, I. Goldfarb, J. Yang, M. Zhang, M. Pickett, J. Strachan, G. Medeiros-Ribeiro, and R. Williams, "Continuous Electrical Tuning of the Chemical Composition of TaO x -Based Memristors," *ACS Nano*, vol. 6, no. 3, pp. 2312-2318, Feb. 2012.

[106] B. Mouttet, "Proposal for memristors in signal processing," in *Proceedings of the 3rd Int. ICST Conf. Nano-Net*, Revised selected papers, Boston, MA, Sept. 2008, pp. 11-13.

[107] M. Hu, H. Li, Q. Wu, and G. S. Rose, "Hardware realization of BSB recall function using memristor crossbar arrays," in *Proceedings of the 49th Annual Design Automation Conference*, San Francisco, CA, June 2012, pp. 498-503.

[108] J. Vogel, M. Bonfim, N. Rougemaille, O. Boulle, I. Miron, S. Auffret, B. Rodmacq, G. Gaudin, J. Cezar, F. Sirotti, and S. Pizzini, "Direct Observation of Massless Domain Wall Dynamics in Nanostripes with Perpendicular Magnetic Anisotropy," *Physical Review Letter*, vol. 108, no. 24, pp. 247202, June 2012.

[109] D. Ngo, K. Ikeda, and H. Awano, "Direct Observation of Domain Wall Motion Induced by Low-Current Density in TbFeCo Wires," *Applied Physics Express*, vol. 4, no. 9, pp. 093002, Sept. 2011.

[110] P. Kinget, "Device mismatch and tradeoffs in the design of analog circuits," *IEEE Journal Solid-State Circuits*, vol. 40, pp. 1212-1224, June 2005.

[111] A. Demosthenous, S. Smedley, and J. Taylor, "A CMOS analog winner-take-all network for large-scale applications," *IEEE Trans. Circuits Syst. I,* vol. 45, pp. 300-304, Mar. 1998.

[112] R. DŁugosz, and T. Talaśka, "Low power current-mode binary-tree asynchronous Min/Max circuit," *Microelectronics Journal,* vol. 41, no. 1 pp. 64-73, Jan. 2010

[113] M. Sharad, G. Panagopoulos, and K. Roy, "Spin neuron for ultra-low power computational hardware," in *Proceedings of the 70th Annual Device Research Conference (DRC),* University Park, TX, June 2012, pp. 221-222.

[114] M. Sharad, D. Fan, and K. Roy, "Spin-neurons: A possible path to energy-efficient neuromorphic computers," *Journal of Applied Physics*, vol. 114, no. 23, pp. 234906, Dec. 2013.

[115] I. M. D'Aquino, "Nonlinear magnetization dynamics in thin-films and nanoparticles," Ph.D. dissertation, University of Naples Federico II, 2004.

[116] K. Kim, S. Seo, J. Ryu, K. Lee, and H. Lee, "Magnetization dynamics induced by in-plane currents in ultrathin magnetic nanostructures with Rashba spin-orbit coupling," *Physical Review B,* vol. 85, no. 18, pp. 180404, May 2012.

[117] Yann.lecun.com, "MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges," [Online]. Available: http://yann.lecun.com/exdb/mnist/.

[118] Cs.columbia.edu, "CAVE | Software: COIL-20: Columbia Object Image Library," [Online] Available: http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php.

[119] Hpl.hp.com, "HP Labs: CACTI," 2015. [Online]. Available: http://www.hpl.hp.com/research/cacti/.

[120] C. Wang, R. Kuo, and T. Tsai, "A high precision low dropout regulator with nested feedback loops," *Microelectronics Journal,* vol. 42, no. 7, pp. 966-971, July. 2011.

[121] PJ. Braspenning, F. Thuijsman, and AJMM. Weijters, "Artificial neural networks: an introduction to ANN theory and practice", Vol. 931. Springer Science & Business Media, 1995.

[122] L. Chua and L. Yang, "Cellular neural networks: Applications," *IEEE Transactions on Circuits and Systems*, Vol. 35, no. 10, pp.1273-1290, Oct. 1988

[123] M. Ueda, Y. Kaneko, Y. Nishitani, and E. Fujii, "A neural network circuit using persistent interfacial conducting heterostructures," *Journal of Applied Physics* Vol. 110, no. 8 pp. 086104, Oct. 2011

[124] R. Dlugosz, T. Talaska, and W. Pedrycz, "Current-mode analog adaptive mechanism for ultra-low-power neural networks," *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 58, no. 1 pp.31-35, Jan. 2011

[125] A. Bermak, "A highly scalable 3D chip for binary neural network classification applications," In Proceedings of the 2003 International Symposium on Circuits and Systems, Bangkok, Thailand, May 2003, pp. V-685

[126] S. Fukami, M. Yamanouchi, K. J. Kim, T. Suzuki, N. Sakimura, D. Chiba, S. Ikeda, et. al., "20-nm magnetic domain wall motion memory with ultralow-power operation," in *Proceedings of the IEEE International Electron Devices Meeting (IEDM)*, Washington, DC, Dec. 2013, pp. 3.5.1-3.5.4.

[127] Ikeda, W. Brown, "Thermal Fluctuations of a Single-Domain Particle," *Physical Re- view*, vol. 130, no. 5, pp. 1677–1686, Jun. 1963.

[128] A. Chanthbouala, R. Matsumoto, J. Grollier, V. Cros, A. Anane, A. Fert, A. V. Khvalkovskiy, et al., "Vertical-current-induced domain-wall motion in MgO-based magnetic tunnel junctions with low current densities," *Nature Physics* vol. 7, no. 8, pp. 626-630, Apr. 2011

[129] M. Sharad, C. Augustine, G. Panagopoulos, and K. Roy, "Spin-based neuron model with domain-wall magnets as synapse," IEEE Transactions on Nanotechnology, Vol. 11, no. 4, pp. 843-853, June 2012

[130] M. Sharad, D. Fan, K. Aitken, and K. Roy, "Energy-Efficient Non-Boolean Computing With Spin Neurons and Resistive Memory," *IEEE Transactions on Nanotechnology*, vol. 13, pp. 23-34, Jan. 2014.

[131] D. Fan, Y. Shim, A. Raghunathan and K. Roy, "STT-SNN: A Spin-Transfer-Torque Based Non-Linear Soft-Limiting Neuron for Low-Power Artificial Neural Networks," I*EEE Transactions on Nanotechnology:* vol. PP, no. 99, June 2015

[132] K. Roy, D. Fan, X. Fong, Y. Kim, M. Sharad, S. Paul, S. Chatterjee, S. Bhunia, and S. Mukhopadhyay "Exploring Spin Transfer Torque Devices for Unconventional Computing", IEEE *Journal on Emerging and Selected Topics in Circuits and Systems*, Vol. 5, No. 1, March 2015

[133] D. Fan, M. Sharad, and K. Roy, "Design and synthesis of ultra low energy spin-memristor threshold logic," *IEEE Transactions on Nanotechnology*, vol. 13, no. 3, pp. 574–583, May 2014.

[134] M. Sharad, D. Fan and K. Roy, "Energy-Efficient and Robust Associative Computing with Injection-Locked Dual Pillar Spin-Torque Oscillators", *IEEE Transactions on Magnetics*, vol. 51, no. 7, Jan. 2015.

[135] [online] mumax.github.io

[136] [online] HTTP://WWW.MATHWORKS.COM/PRODUCTS/NEURAL-NETWORK/

[137] X. Fong, S.K. Gupta, N.N. Mojumder, S.H. Choday, C. Augustine, and K. Roy, "KNACK: A hybrid spin-charge mixed-mode simulator for evaluating different genres of spin-transfer torque MRAM bit-cells," In *Proceedings of the 16th International Conference on Simulation of Semiconductor Processes and Devices*, Osaka, Japan, Sept. 2011, pp. 51-54.

[138] S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tanabe, and J. Yamada, "A 1-V high-speed MTCMOS circuit scheme for power-down application circuits," *IEEE Journal of Solid-State Circuits*, Vol. 32, no. 6 pp.861-869, June 1997

[139] X. Wang, Y. Chen, H. Xi, H. Li, and D. Dimitrov, "Spintronic memristor through spin-torque-induced magnetization motion," IEEE Electron Device Letters, Vol. 30, no. 3, pp.294-297, Feb. 2009

[140] SH. Jo, T. Chang, I. Ebong, BB. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano letters*, Vol. 10, no. 4, pp.1297-1301, Apr. 2010

[141] SG. Ramasubramanian, R. Venkatesan, M. Sharad, K. Roy, and A. Raghunathan, "SPINDLE: SPINtronic deep learning engine for large-scale neuromorphic computing," In *Proceedings of the 2014 international symposium on Low power electronics and design*, La Jolla, CA, Aug. 2014, pp. 15-20

[142] W. Duch and N. Jankowski, "Survey of neural transfer functions," *Neural Computing Surveys* Vol. 2, no. 1, pp.163-212, 1999

[143] T. Shibata, R. Zhang, SP. Levitan, D.E. Nikonov, and G.I. Bourianoff, "CMOS supporting circuitries for nano-oscillator-based associative memories," In *Proceedings of the 13th International Workshop on Cellular Nanoscale Networks and Their Applications*, Turin, 2012, pp. 1-5.

[144] A. Saberkari, E. Alarcón, and S.B. Shokouhi, "Fast transient current-steering CMOS LDO regulator based on current feedback amplifier," *the VLSI journal Integration*, Vol. 46, no. 2, pp.165-171, Mar. 2013

[145] S. Fukami, T. Suzuki, N. Ohshima, K. Nagahara, and N. Ishiwata, "Micromagnetic analysis of current driven domain wall motion in nanostrips with perpendicular magnetic anisotropy," *Journal of Applied Physics* Vol.103, no. 7, pp. 07E718, Apr. 2008

[146] M. Kläui, P.-O. Jubert, R. Allenspach, A. Bischof, J. A. C. Bland, G. Faini, U. Rüdiger, C. A. F. Vaz, L. Vila, and C. Vouille, "Direct Observation of Domain-Wall Configurations Transformed by Spin Currents," *Physics Review Letter*s, Vol. 95, pp.026601, July 2005

[147] M.C. Hickey and J. S. Moodera, "Origin of Intrinsic Gilbert Damping", *Physical review letters*, vol. 102, pp.137601, Apr. 2009

[148] Y. Togawa, T. Kimura, K. Harada, T. Akashi, T. Matsuda, A. Tonomura, and Y. Otan, "Current-excited magnetization dynamics in narrow ferromagnetic wires," *Japanese journal of applied physics* Vol.45, no. pp.L683-L685, July 2006

[149] G. Meier, M. Bolte, R. Eiselt, B. Krüger, DH Kim, and P. Fischer, "Direct imaging of stochastic domain-wall motion driven by nanosecond current pulses," *Physical review letters*, Vol.98, no.18, pp. 187202, May 2007

[150] D. Fan, M. Sharad, A. Sengupta and K. Roy, "Hierarchical Temporal Memory Based on Spin-Neurons and Resistive Memory for Energy-Efficient Brain-Inspired Computing," *IEEE Transactions on Neural Networks and Learning Systems*, 2015 (accepted, in proceeding of publication)

[151] G. Csaba, M. Pufall, W. Rippard and W. Porod, "Modeling of Coupled Spin Torque Oscillators for Application in Associative Memory," *in Proceedings of IEEE conf. on Nanotechnology*, Birmingham, Aug. 2012, pp1-4

[152] G. Csaba, M. Pufall, D.E. Nikonov, G.I. Bourianoff, A. Horvath, T. Roska andW. Porod, "Spin torque oscillator models for applications in associative memories," *in Proceedings of 13th International Workshop on Cellular nanoscale Networks and Their Applications*, Turin, Aug. 2012, pp. 1-2

[153] D. E. Nikonov, G. Csaba, W. Porod, T. Shibata, D. Voils, D. Hammerstrom, I. Young and G.I. Bourianoff, "Coupled-Oscillator Associative Memory Array Operation," arXiv:1304.6125 [cond-mat.mes-hall]

[154] S. P Levitan, Y. Fang, D.H. Dash, T. Shibata, D.E. Nikonov, and G.I. Bourianoff, "Non_Boolean Associative Architecture Based on Nano-Oscillators", *in Proceedings of 13th International Workshop on Cellular nanoscale Networks and Their Applications*, Turin, Aug. 2012, pp. 1-6

[155] M.R Pufall, W.H. Rippard, S.E. Russek, S. Kaka and J.A. Katine, "Electrical Measurement of Spin-Wave Interactions of Proximate Spin Transfer Nanooscillators," *Physical Review Letters*, vol. 97, no.8, pp. 087206/1-4, Aug. 2006

[156] A. N. Slavin, and V. S. Tiberkevich, "Nonlinear self-phase-locking effect in an array of current-driven magnetic nanocontacts," *Physical Review B*, vol. 72, no. 9, pp. 092407, Sep. 2005.

[157] T. Moriyama, G. Finocchio, M. Carpentieri, B. Azzerboni, D.C. Ralph, and R. A. Buhrman, "Phase locking and frequency doubling in spin-transfer-torque oscillators with two coupled free layers," *Physical Review B*, vol. 86, no. 6, pp.060411, Aug. 2012

[158] G.Csaba and W. Porod, "Computational study of spin-torque oscillator interactions for non-Boolean computing applications," *IEEE trans. On Mag.* Vol. 49, No. 7, pp. 4447-4451, July 2013

[159] B. Razavi, "A study of injection locking and pulling in oscillators", *IEEE J. Solid-State Circuits*, vol. 39, no. 9, pp.1415-1424, Sept. 2004

[160] B. Georges, J. Grollier, M. Darques, V. Cros, C. Deranlot, B. Marcilhac, G. Faini, and A. Fert, "Coupling Efficiency for Phase Locking of A Spin Transfer Nano-Oscillator to A Microwave Current," *Phys. Rev. Lett.* Vol. 101, no. 1, pp. 017201, Jul. 2008

[161] W.H. Rippard, M.R.Pufall, S.Kaka, T.J.Silva, S.E.Russek, and J.A.Katine, "Injection Locking and Phase Control of Spin Transfer Nano-oscillators," *Phys. Rev. Lett.* Vol. 95, no. 6, pp.067203, Aug. 2005.

[162] R. P. Cowburn, "Property variation with shape in magnetic nanoelements," *Journal of Physics D: Applied Physics*, vol. 33, no. 1, pp. R1-R16, Jan. 2000.

[163] G. D. Fuchs1, J. C. Sankey, V. S. Pribiag, L. Qian1, P. M. Braganca, A. G. F. Garcia, E. M. Ryan, Zhi-Pan Li, O. Ozatay, D. C. Ralph and R. A. Buhrman, "Spin-torque ferromagnetic resonance measurements of damping in nanomagnets," *Applied Physics Letters*, vol. 91, no. 6, pp.062507-062507, Aug. 2007.

[164] K. Kodama, T. Furubayashi, H. Sukegawa, T. M. Nakatani, K. Inomata and K. Hono1, "Current-perpendicular-to-plane giant magnetoresistance of a spin valve using Co2MnSi Heusler alloy electrodes," *Journal of Applied Physics* vol.105.no.7, pp.07E905-07E905, Apr. 2009

[165] W. Rippard, A.M. Deac, M.R. Pufall, J.M. Shaw, M.W. Keller, S.E. Russek, G.E. Bauer, and C. Serpico, "Spin-transfer dynamics in spin valves with out-of-plane magnetized CoNi free layers," *Phys. Rev. B*, vol.81, no. 1, pp.014426, Jan. 2010

[166] J. C. Sankey, I. N. Krivorotov, S. I. Kiselev, P. M. Braganca, N. C. Emley, R. A. Buhrman, and D. C. Ralph, "Mechanisms limiting the coherence time of spontaneous magnetic oscillations driven by dc spin-polarized currents," *Phys. Rev. B*, vol.72, no. 22, pp.224427, Dec.2005

[167] K. Ohashi, K. Hayashi, K. Nagahara, K. Ishihara, E. Fukami, J. Fujikata, S. Mori, M. Nakada, T. Mitsuzuka, K.Matsuda, H. Mori, A. Kamijo, H. Tsuge, "Low-resistance tunnel magnetoresistive head," *IEEE Transactions on Magnetics*, vol.36, no. 5, pp.2549-2553, Aug. 2002

[168] F.A. Shah, V.K. Sankar, P. Li, G. Csaba, E. Chen and G.H. Bernstein "Compensation of orange-peel coupling effect in magnetic tunnel junction free layer via shape engineering for nanomagnet logic applications," *Journal of Applied Physics*, vol.115, no.17, pp.17B902, May. 2014.

[169] T. Wada, T. Yamane, T. Seki, T. Nozaki, Y. Suzuki, H. Kubota, A. Fukushima, S. Yuasa, H. Maehara, Y. Nagamine, K. Tsunekawa, D. D. Djayaprawira, and N. Watanabe, "Spin-transfer-torque-induced rf oscillations in CoFeB/MgO/CoFeB magnetic tunnel junctions under a perpendicular magnetic field," *Phys. Rev. B*, vol. 81, no.10, pp.104410, Mar. 2010.

[170] A. Fert, and P. M. Levy. "Spin Hall Effect induced by resonant scattering on impurities in metals," *Physical Review Letters*, vol. 106, no.15, pp.157208, Apr. 2011

[171] S. Manipatruni, D.E. Nikonov, and I.A. Young, "Energy-delay performance of giant spin Hall effect switching for dense magnetic memory," *Applied Physics Express*, vol. 7, no. 10, pp. 103001, Oct. 2014

[172] S.Datta, S. Salahuddin, and B. Behin-Aein, "Non-volatile spin switch for Boolean and Non-Boolean Logic," *Appl. Phys. Lett.* Vol.101, no. 25, pp.252411, Dec. 2012

[173] T. L. Gilbert, "A Lagrangian formulation of the gyromagnetic equation of the magnetization field," *Phys. Rev.* vol.100, 1243 (1955)

[174] D. Yogain, V. Krishna, and A. Baliga, "Design of High Speed Adders of Efficient Digital Design Blocks," ISRN Electronics, vol.2012, 2012

[175] J. Deveugele and M. Steyaert, "A 10-bit 250-MS/s binary-weighted current-steering DAC", *IEEE Journal of Solid-State Circuits*, vol. 41, issue. 2, Feb. 2006

[176] N. Srinivasa, et. al. "Probabilistic inference devices for unconventional processing of signals for intelligent data exploitation," *40th Annual GOMAC Tech Conference*, March 23-26, St. Louis, MO, 2015

[177] S. Kaka, M.R. Pufall, W.H. Rippard, T.J. Silva, S.E. Russek and J.A. Katine, "Mutual phase-locking of microwave spin torque nano-oscillators", *Nature Letters*, vol. 437, pp. 389-392, Sept. 2005

[178] F.B. Mancoff, N.D. Rizzo, B.N. Engel and S. Tehrani, "Phase-Locking in double-point-contact spin-transfer devices", *Nature Letters*, vol. 437, pp. 393-395, Sept, 2005

[179] A. Horvath, F. Corinto, G. Csaba, W. Porod and T. Roska, "Synchronization in cellular spin torque oscillator arrays", *13th International workshop on cellular nanoscale networks and their applications*, Turin, Aug. 2012, pp. 1-6

[180] Z. Li, Y. Charles Li, and S. Zhang, "Dynamic magnetization states of a spin valve in the presence of dc and ac currents: Synchronization, modification and chaos", *Physical Review B*, vol. 74, pp. 054417, Aug. 2006.

[181] W. Rippard, M. Pufall, and A. Kos, "Time required to injection-lock spin torque nanoscale oscillators", *Applied Physics Letters*, vol.103, no.18, pp. 182403, Oct. 2013

[182] W.S. Zhao, S. Chaudhuri, C. Accoto, J-O. Klein, C. Chappert, and P Mazoyer, "Cross-Point architecture for spin transfer torque magnetic random access memory ", *IEEE Trans. on Nanotechnology*, vol.11, no.5, pp.907-917, Sept. 2012.

[183] X. Yao, J. Harms, A. Lyle, F. Ebrahimi, Y. Zhang and J-P. Wang, "Magnetic Tunnel Junction-Based Spintronic Logic Units Operated by Spin Transfer Torque", *IEEE Trans. on Nanotechnology*, vol.11, no. 1, pp.120-126, June 2011.

[184] K. Xia, P.J. Kelly, G.E.W. Bauer, A. Brataas, and I. Turek, "Spin torques in ferromagnetic/normal-metal structures," *Phys. Rev. B*, vol. 65, no. 22, pp. 220401, June 2002.

[185] J. C. Slonczewski and J. Z. Sun, "Theory of voltage-driven current and torque in magnetic tunnel junctions," *J. Magn. Magn*. Mater. vol. 310, no. 2, pp. 169–175, Mar 2007.

[186] J. Xiao, A. Zangwill and MD. Stiles, "Boltzmann test of Slonczewski's theory of spin-transfer torque," *Phys. Rev. B*, vol.70, no.17, pp.172405, Nov. 2004

[187] S. Salahuddin, D. Datta, and S. Datta, "Spin Transfer Torque as a Non-Conservative Pseudo-Field,' *arXiv:0811.3472* [cond-mat.mes-hall]

[188] S. Ramo, J. R. Whinnery, and T. V. Duzer, “Fields and Waves in Communication Electronics”, 3rd ed. New York: Wiley, 1994.

[189] R. D. McMichael and M. J. Donahue, "Head to head domain wall structures in thin magnetic strips," *Magnetics, IEEE Transactions on*, vol. 33, no. 5, pp. 4167–4169, 1997.

[190] Y. Nakatani, A. Thiaville, and J. Miltat, "Head-to-head domain walls in soft nano-strips: a refined phase diagram," *Journal of Magnetism and Magnetic Materials*, vol. 290, pp. 750–753, Apr. 2005.

[191] I. Dzyaloshinsky, "A thermodynamic theory of ‘weak’ ferromag-netism of antiferromagnetics," *Journal of Physics and Chemistry of Solids*, vol. 4, no. 4, pp. 241–255, Jan. 1958.

[192] T. Moriya, "Anisotropic Superexchange Interaction and Weak Ferromagnetism," *Phys. Rev.*, vol. 120, no. 1, pp. 91–98, Oct. 1960.

[193] M. Heide, G. Bihlmayer, and S. Blügel, "Dzyaloshinskii-Moriya interaction accounting for the orientation of magnetic domains in ultrathin films: Fe/W(110)," *Phys. Rev. B*, vol. 78, no. 14, p. 140403, Oct. 2008.

VITA

VITA

Deliang Fan received his Bachelor of Science degree in Electronic Information Engineering from Zhejiang University, China, in 2010 and Master of Science degree in Electrical and Computer Engineering from Purdue University, West Lafayette, IN, USA, in 2012.Currently he is a graduate research assistant of Professor Kaushik Roy and pursuing Ph.D. degree in Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA.

His primary research interest lies in cross-layer (algorithm/architecture/circuit) co-design for low-power Boolean, non-Boolean and neuromorphic computation using emerging technologies like spin-transfer torque devices. His past research interests include cross-layer digital system optimization and imperfection-resilient scalable digital signal processing algorithms and architectures using significance driven computation.