Summer 2014

Energy efficient hybrid computing systems using spin devices

Mrigank Sharad
Purdue University

Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Mrigank Sharad

Entitled
Energy Efficient Hybrid Computing Systems Using Spin Devices

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

KAUSHIK ROY
Chair

ANAND RAGHUNATHAN

BYUNGHOO JUNG

DMITRI E. NIKONOV

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): KAUSHIK ROY

Approved by: M. R. Melloch 07-25-2014
Head of the Graduate Program Date

ENERGY EFFICIENT HYBRID COMPUTING SYSTEMS USING SPIN DEVICES

A Dissertation

Submitted to the Faculty

of

Purdue University

by

MrigankSharad

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2014

Purdue University

West Lafayette, Indiana

"When You Really Want Something, the Whole Universe Conspires in Helping You Achieve it" - Paulo Coelho

ACKNOWLEDGEMENTS

I sincerely acknowledge my advisor, Prof. Kaushik Roy, for his constant motivation, insight into the problems, and maintaining the big picture consistently throughout my PhD. This PhD would not be possible without his outstanding support and patience for investigating new problems, which are intricate and interesting simultaneously. I also earnestly thank him for creating a learning atmosphere in the lab through which I acquired knowledge about several related subjects which helped me significantly inpreparing this thesis. I could not wish for a superior and friendlier advisor for pursuing my PhD.

I would also like to acknowledge Prof. Anand Raghunathan, Prof. Jung, and Prof. Nikonov for clarifying my doubts with patience and understanding. They have provided me continuous guidance, intuition and self-confidence in my PhD pursuit

I sincerely thank my parents for their incessant support, love and care which made my PhD very enjoyable. Special thanks to relatives, friends and lab colleagues for their help, whenever needed. Above all, I thank GOD for all the divine grace and blessings.

TABLE OF CONTENTS

Page

Page

Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

ABSTRACT

Sharad, Mrigank. .Ph.D., Purdue University, December 2014.Energy Efficient Hybrid Computing Systems Using Spin Devices. Major Professor: Kaushik Roy.

Emerging spin-devices like magnetic tunnel junctions (MTJ's), spin-valves and domain-wall magnets (DWM) have opened new avenues for spin-based logic design. This work explored potential computing applications which can exploit such devices for higher energy-efficiency and performance. The proposed applications involve hybrid design schemes, where charge-based devices supplement the spin-devices, to gain large benefits at the system level. As an example, lateral spin valves (LSV) involve switching of nano-magnets using spin-polarized current injection through a metallic channel such as Cu. Such spin-torque based devices possess several interesting properties that can be exploited for ultra-low power computation. Analog characteristic of spin current facilitate non-Boolean computation like majority evaluation that can be used to model a neuron. The magneto-metallic neurons can operate at ultra-low terminal voltage of ~20mV, thereby resulting in small computation power. Moreover, since nano-magnets inherently act as memory elements, these devices can facilitate integration of logic and memory in interesting ways. The spin based neurons can be integrated with CMOS and other emerging devices leading to different classes of neuromorphic/non-Von-Neumann architectures. The spin-based

designs involve ‘mixed-mode’ processing and hence can provide very compact and ultra-low energy solutions for complex computation blocks, both digital as well as analog. Such low-power, hybrid designs can be suitable for various data processing applications like cognitive computing, associative memory, and current-mode on-chip global interconnects. Simulation results for these applications based on device-circuit co-simulation framework predict more than ~100x improvement in computation energy as compared to state of the art CMOS design, for optimal spin-device parameters.

1. INTRODUCTION

1.1 Emerging Post-CMOS device technologies

1.1.1 Exploration of novel-technologies for substituting and augmenting CMOS

Over the past three decades silicon MOSFET scaling enabled us to design systems with lowest energy-consumption along with high-performance. However, today, MOSFET scaling faces several impending challenges, like, high leakage-power, high on-chip power density, and device parameter-variations [118-122].Continuing the growth that the semiconductor industry has enjoyed for decades may therefore necessitate exploration of technologies (devices, interconnect, and integration-techniques) beyond the industry mainstays of Silicon and CMOS. To truly leverage the potential of the *new devices*, we may not view them merely as drop-in replacements. Rather, we should seize the unprecedented opportunity to explore novel application-regimes, where the unique characteristics of the emerging-devices can be leveraged to assist and augment CMOS. Future ICs may involve heterogeneous-integration of CMOS with novel device-technologies to achieve specific performance matrices for general-purpose as well as application-specific computing-platforms [123]. For instance, several non-volatile device technologies, like MRAM [125] and PCRAM [126], have been identified as forerunners

for replacing CMOS for on-chip memory, that can overcome the major bottlenecks of SRAM, namely, leakage and scalability [118]. Meeting both, power and performance requirements, in future multi-core processors may require hetero-integration of CMOS with emerging device-technologies like, TFETs [127], for low-power sub-threshold operations. With technology-scaling, degradation of performance and energy-efficiency of global no-chip interconnects has also motivated extensive research for alternate technology solutions. Hetero-integration of on-chip optical links [128]] as well as novel interconnect technologies like graphene and carbon nano-tube (CNT) [129] with CMOS has been extensively studied to explore solution to such design issues. With further innovation and progress in nano-technology and material science, research on such heterogeneous integration for different components of computing hardware is expected to expand further [124].

1.1.2 Search for alternate computing paradigms and their technology solutions

Apart from the research for technology-solutions to the critical design-challenges faced by CMOS based Von-Neumann hardware, recent years have seen growing interest in design-implementation of alternate, non-Boolean computing-models. CMOS transistors, being on/off switches, are an ideal match to the abstractions of switching functions and Boolean logic, which form the underpinnings of modern computing. However, traditional computing models (Boolean logic, von Neumann architecture) are highly inefficient - requiring orders of magnitude more energy-consumption for performing tasks that humans routinely perform, such as visual-recognition, semantic-analysis, and reasoning. Non-Boolean computation models like neural-networks [130],can algorithmically

outperform Von-Neumann architecture for such cognitive computation. Cognitive computing-schemes in general employ feature-extraction from sensory data followed by pattern-matching based on memorized information. ASIC implementation of such non-Boolean computing schemes has gained widespread attention in last few years, especially, for mobile-computing platforms [131]. However, CMOS based implementations for such computing-models prove to be highly inefficient in terms of power and area-complexity, thereby limiting the scale, computing-power and sophistication of actual computing algorithms implemented [132]. These design challenges stems from the inefficiency in modeling the fundamental non-Boolean computing primitives in such schemes using CMOS transistors. The research on energy-efficient hardware for non-Boolean computing models has therefore fueled great interest in emerging device-technologies that can offer operational characteristics more suitable for direct mapping of such non-Boolean computing primitives [142], [143].

1.2 Emerging spin-devices for computing-hardware

Among different post-CMOS device technologies under exploration, different genres of spin-devices, based on nano-scale magnets, have been identified as one of the potential candidates for future memory and logic design [125]. In fact spin-based on-chip memory may be one of the most suitable replacements for SRAM in near future. Although the promise of spin-devices for on-chip memory applications is well accepted, their potential for logic-computation is relatively less established [133]. The target of this research is to explore the potential applications of emerging spin-devices in Boolean as well as Non-Boolean computing hardware. In this work we first discuss the prospects and potential

design-techniques for spin-based Boolean-logic. Following this, our work on spin-based non-Boolean computing is presented, which shows that the specific characteristics of emerging spin-devices can be highly attractive for the implementation of energy-efficient non-Boolean computing-systems. This research also explores the possible advantages of spin devices in the design of low-power global on-chip interconnects. The following sections of this chapter give a brief overview of recent work in the field of spin-based memory and logic devices and links it to the work presented in this thesis.

1.2.1 Spin Devices for on-Chip Memory

Among different potential applications, the prospects of spin-devices for on-chip memory have been found to be most promising [125]. In a nano-scale magnetic-layer, the direction of magnetization vector can be regarded as information similar to charge as information in MOSFETs. Moreover, magnets can retain its magnetization without any external assistance or in other words magnets are non-volatile.

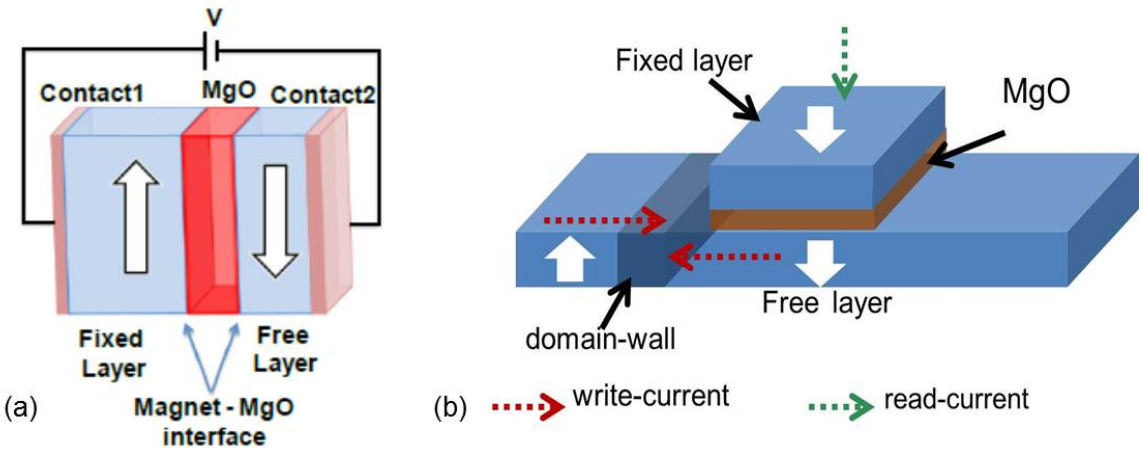


Figure 1.1(a)Magnetic Tunneling Junction (MTJ), (b) Spin-transfer Torque MRAM (STT-MRAM) with access scheme

The basic data-storage device in an MRAM is a magnetic tunnel junction (MTJ), shown in fig. 1a [25]. It constitutes of two nano-magnetic layers separated by a tunneling barrier in the form of MgO. The resistance of an MTJ is high when the layers possess the same spin-polarity and vice-versa. A current-based sense-amplifier is typically used to distinguish between the two resistance states, in order to read the stored data-bit [30].

Magnetization of one of the two magnetic layers in an MTJ is fixed while that of the other can be switched by the application of charge-current of appropriate polarity across the two terminal device [25]. Such current-based switching of nano-magnets is governed by spin transfer torque (STT) effect [32]. The spin-polarization effect of magnets on charge current has been known since long. Charge current in a non-magnetic material has normally zero spin polarization due to random orientation of large number of electron-spins. However, after passing through a nano-magnet, majority of electrons, constituting the charge-current, acquire spin-polarity parallel to that of the magnet. Such a spin-polarized current can exert 'spin-torque' on another magnet, causing it to change its spin-orientation [134]. Spin-torque based magnetization switching offers several advantages, like scalability, energy-efficiency and higher performance, as compared to magnetic-field based switching, that has been widely used for MRAM previously [32, 135].

Spin Transfer torque Magnetic Random Access Memory (STT-MRAM) offers several attractive features like non-volatility, high area-density, zero-leakage and reasonable read-write characteristics in terms of speed and performance [31]. Recent progress in current-induced spin-torque based switching-mechanisms for nano-magnets

have paved the way for further improvements in write performance and energy for MRAM [136], thereby enhancing the suitability of this technology for on-chip memory.

1.2.2 Spin Devices for Boolean Logic

Although promising for memory applications, the promise of spin devices for logic computation is relatively less established [136]. Until recently, nano-magnet logic (NML) was the only predominant spin-based computation scheme under exploration [137], [138]. It employs dipolar coupling between nano-magnets to perform logic computation and offers interesting features like non-volatility, zero leakage and compactness [137]. However, magnetic field based Bennett clocking used in NML requires pulsed current transmission through metal lines that makes it inefficient in terms of computation energy [138]. Theoretical possibility of alternate strategy for Bennett clocking in NML have been proposed recently [139], that makes use of anisotropic strain induced by multiferroic layers to turn magnets to hard axis. If successful, such a scheme could boost up the prospects of NML scheme and would make it attractive for low performance electronics like those used in biomedical implants [139].

Recent experiments on spin torque in device structures like lateral spin valve (LSV) [4], [5] (fig.2a), domain wall magnets (DWM) [6], [7], and magnetic tunnel junctions (MTJ), have opened new avenues for spin-based computation. Several logic schemes have been proposed using such devices. Hybrid design schemes using MTJ have been explored that aim to club memory with logic and can possibly benefit from reduced memory-data traffic [12]. Such schemes can be useful for programmable logic design, where the MTJ based memory cells can offer low-leakage and high density [140].

Although MJT based memory elements may facilitate energy-efficient read-access for programmable-logic cells, the write operation through such a two-terminal, high-resistance vertical spin valve can take order of magnitude more energy than a CMOS-gate [143]. Hence, logic schemes employing switching of 2-terminal MTJs, may not offer any significant advantage over CMOS. Recently, the application of multi-terminal STT-switches have been proposed for digital logic design [16], [18]. Such a device, shown in fig. 1b, offers separate terminals for read and write operations. In the device shown in fig. 1b, the free-layer is a part of an extended nano-magnetic strip, which may have multiple spin-domains of opposite polarities. The transition regions between such opposite-polarity domains is called 'domain-wall' (fig. 1b). A domain-wall can be moved along a nano-strip using STT-effect, resulting from charge-current flow along the strip. Such a current-induced domain-wall motion can be used for 'writing' into the free-layer of an MTJ. The low resistance, magnetic write-path in such a tree terminal device allows much smaller write voltages, as compared to a 2-terminal MTJ. At the same time, it also facilitates efficient sensing through the fixed-layer, similar to a 2-terminal device. Such a low power device can therefore be more efficient for logic design employing magnetization-switching [144].

MTJ-based logic styles involve conversion between spin and charge as state variables for read-write operations. Recently, use of STT in lateral spin valve (LSV) has been proposed to realize All Spin Logic (ASL) that avoids such inter-conversions [9]. An

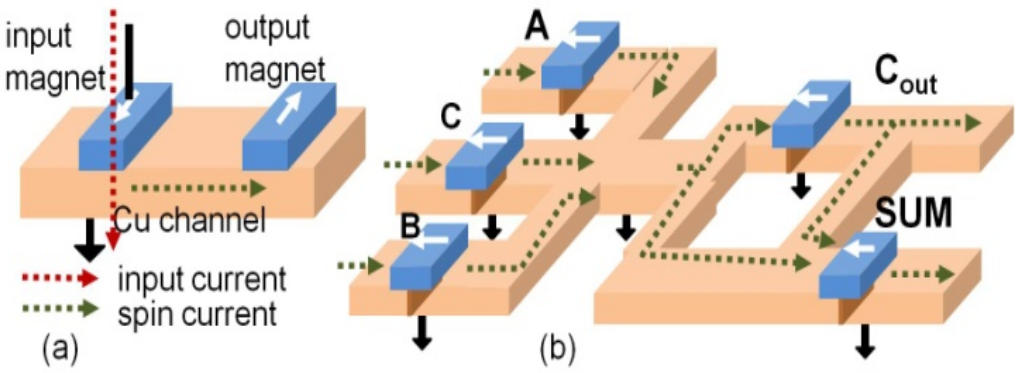


Figure 1.2(a) Lateral spin valve with non-local spin injection, (b) ASL full adder based on spin majority evaluation

LSV (fig. 2a) constitutes of multiple magnets interconnected using non-magnetic channels (like Copper) [3]. In such a multi-terminal device, spin-polarized current injected into the channel through a set of 'input-magnets' can affect the state of one or more 'output-magnets'.Use of STT in LSVs can therefore facilitate higher degree of spin current manipulation for logic-design. ASL employs cascaded LSV's interacting through spin torque, to realize logic gates and larger blocks like compact full adders [9, 12], based on spin majority evaluation (fig. 2b). The ASL gates being fully metallic, can allow ultra-low voltage operation, leading the possibility of energy-efficient switching at gate-level [10]. However, to assess the best performance achievable for a proposed logic style, for a given set of device-parameters, it is important to evaluate the scheme at circuit and system

1.2.3 Spin Devices for Non-Boolean Computing

Most of the spin based computation schemes proposed so far have been centered on modeling digital logic gates using nano-magnetic devices. A wider perspective on

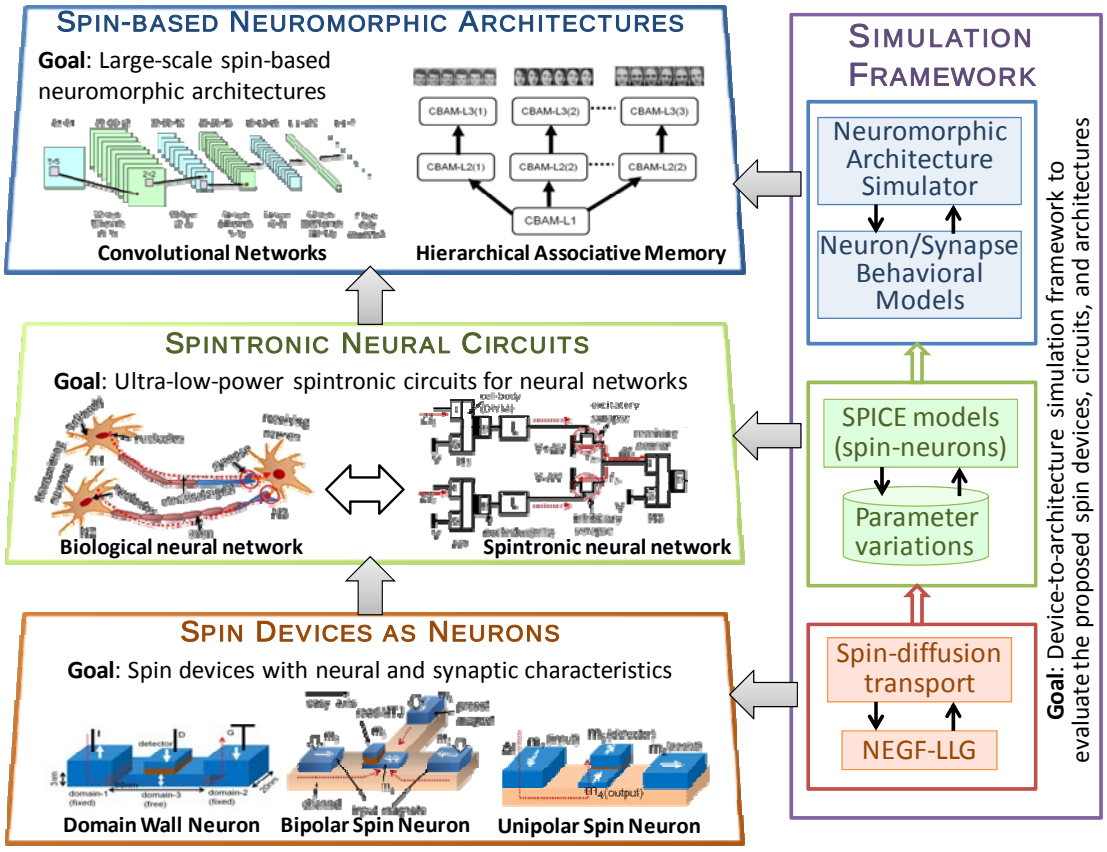


Figure 1.3 Overview of the proposed research – devices, circuits, architectures, and simulation framework

application of spin torque devices, however, would involve, not only exploring possible combination of spin and charge devices but, searching for computation models which can derive maximum benefits from such heterogeneous integration. We noted that ultra-low voltage, current-mode operation of magneto-metallic devices like LSV's and DWM's can be used to realize analog summation/integration and thresholding operations with the help of appropriate circuits, and, can be used to model energy efficient "neurons" [26]-[28]. Such device-circuit co-design can lead to ultra-low power neuromorphic computation architectures, suitable for different data processing applications. The

proposed hybrid design scheme can open a new frontier for spin torque based analog and digital computing. Inspired by this vision, we investigated spin-based non-Boolean/neuromorphic computing (Fig.3) that spans from the device-level to architecture and applications.

As a part of this work we proposed different spin-device-models that can mimic the neuron-functionality and can offer an energy efficient mapping for non-Boolean computing primitives used in such hardware model. Different genres of spin devices, like lateral spin valves, domain wall magnets and spin-torque based oscillators have been explored for such computing models. We explored designs that can maximally leverage the energy benefits of the spin-neurons at the system level. This involved integration of the proposed spin-devices with CMOS and other devices like CMOS-compatible memristors, that can help emulate network-level functionality. Physics-based device simulation has been developed for characterization of the proposed spin-device-models. Device-circuit co-simulation is used for exploring such heterogeneous designs in-order to evaluate system level functionality and performance.

1.2.4 Application of spin-torque switches in the design of global on-chip interconnects

This work identifies global-on-chip-interconnect design as another potentially attractive application of emerging high-speed spin-torque switches. With the scaling of CMOS technology, energy-efficiency and performance of the on-chip global-interconnect degrades [100]. As a result the design of low power and high-speed on-chip global

interconnects can be a major bottleneck for emerging chip-multi-processors (CMP) that employ extensive inter-processor and memory to processor communication.In this work, we explore the potential of low-voltage, magneto-metallic spin-torque (ST) switches for ultra-low energy and high-performance interconnect design [103]. Recently demonstrated high-speed spin-torque switching phenomena based on spin-orbital (SO) coupling effects may be conducive to the design of ultra-low voltage, low-current and high-speed nano-magnetic switches [103]. We present analysis for device and circuit-level optimization of current-mode interconnect design using such switches and compare its performance with conventional CMOS interconnects proposed in literature.

1.3 Thesis Organization

The goals of the proposed research are to (i) establish computing applications (for which CMOS implementations are energy-inefficient) enabled by advances in the physics of spin device technologies, (ii) synergistically explore spin devices, circuit and system design in a regime where the devices are integrated with CMOS to augment its capabilities, (iii) bring together expertise from the device, circuits, architecture, and applications to holistically solve challenges of the beyond-CMOS era.

Rest of the thesis is organized as follows. Spin-torque-based Boolean computing scheme (previously proposed in literature) is provided in chapter-2. All Spin Logic scheme proposed in [9] is discussed in detail along with its limitations and design challenges. Some circuit methods for improving the energy and performance metrics of ASL are also proposed. Chapter 3 introduces the concept of non-Boolean computing with

spin devices. Spin-based device-models proposed in this work for such computing schemes are presented. Chapter 4-6 present different design examples of such spin-based non-Boolean computing systems, which essentially employ hybrid circuit designs with spin-neurons. . Associative computing architecture based on proposed STO-devices and coupling scheme is presented in chapter 7. Chapter 8 describes the potential application of high-speed spin torque switches in global interconnect design. Conclusions and future work are given in chapter 9.

2. BOOLEAN LOGIC WITH SPIN TORQUE: ALL SPIN LOGIC

2.1 All Spin Logic Using Lateral Spin Valves

All Spin Logic (ASL) gates employ multiple nano-magnets interacting through spin-torque using non-magnetic channels. Compactness, non-volatility and ultra-low voltage operation are some of the attractive features of ASL, while, low switching-speed (of nano-magnets as compared to CMOS gates) and static-power dissipation can be identified as the major bottlenecks. In this chapter we explore design techniques that leverage the specific device characteristics of ASL to overcome the inefficiencies and to enhance the merits of this technology, for a given set of device parameters.

Fig. 1a shows a lateral spin valve (LSV) structure, which consists of an 'injecting magnet' (m_1) and a 'receiving' magnet (m_2) connected through a non-magnetic channel. Electrons constituting charge current, after passing through m_1 get left spin-polarized. Spin-polarized charge-current is modeled as a four-component quantity, one charge component I_C and three spin components (Is_x, $Is_y$$Is_z$) [9-11].The charge component relates to the number of electrons constituting the current. The spin components however, denote the effective spin orientation of the current. In a normal charge current, the overall spin current is close to zero, because of random orientations of constituent electron spins. However, due to the effect of spin-momentum exchange, electrons after passing through .

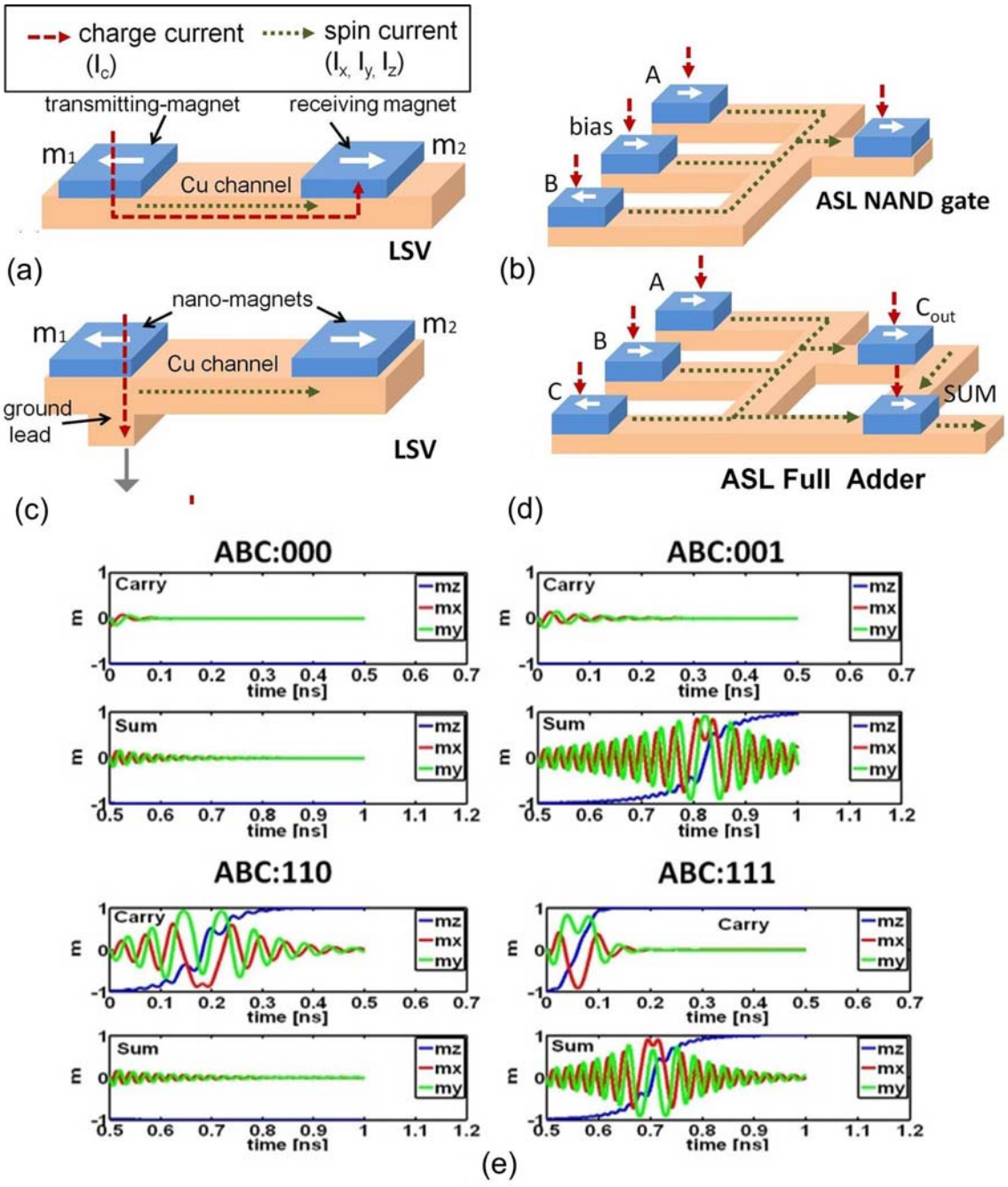


Figure 2.1(a) Lateral spin valve (LSV) with local spin injection, (b) Lateral spin valve (LSV) with non-local spin injection (c) ASL NAND gate (d) ASL full adder(e) Simulation waveforms for FA evaluation.

a magnet acquire a spin polarity parallel to that of the magnet (in this case m_1) [10]. As a result spin-polarized charge current is generated that has a non-zero spin component depending upon the spin orientation of the input magnet m_1. If the spin components of the resulting current are strong enough, they can flip the spin-polarity of the receiving magnet m_2, through which they pass [4]. This effect resultsfrom spin-momentum exchange between the spin-polarized current and the receiving magnet, and, is termed as spin-transfer torque (STT).

Experiments have shown two possible mechanisms for STT induced switching of m_2 under the influence of spin-current injected through m_1. The first method employs direct injection of the spin polarized charge-current into m_2 (fig. 1a). This implies that the charge component I_C, as well as the spin components (Is_x, $Is_y$$Is_z$) of the current injected by $m_1$pass through m_2 and the spin components exert torque on it, causing it to flip.

The second method for STT switching in LSV employs only the spin components of the input current. In this method, the charge component of the input current flows into the ground (fig. 1b). The spin components however result in accumulation of one kind of spin (left spin in this case) under the input magnet m_1. This results in spin-potential difference across the metal channel, causing a spin diffusion current flow, which flips the receiving magnet m_2. Note that, in this case the overall flow of electrons across the channel, i.e., the net charge-current is zero.Owing to the separation of the spin diffusion current responsible for *nano-magnet* switching, from the charge current flow, this phenomena is regarded as ‘non-local’ STT [11].

The use of non-local spin-torque in LSV's facilitates higher degree of spin current manipulation for computing. Non-local spin transport in metal channel can be used to cascade multiple LSV units to realize logic gates. Analog characteristics of current mode switching employed in LSV's can facilitate non-Boolean computation like majority evaluation. Hence, LSV's with multiple input magnets can be used to design spin majority gates. In [9] authors proposed 'all spin logic' (ASL) scheme that employed cascaded LSV's interacting through unidirectional, non-local spin current [3, 4, 9-11]. Interestingly, such spin majority gates can be used to realize compact structures for logic blocks (such as adders, as shown in fig. 2d) that find bulky representation in CMOS circuits. Fig 1c and fig. 1d depict ASL NAND gate [9], and, ASL full-adder (fig. 1d, e) using just five nano-magnets [12]. In the following section we present the 4 component spin-circuit simulation model for ASL.

2.2 For Component Spin Circuit Model for ASL

In order to simulate the neuron model, which is based on the lateral spin valve structure shown in fig. 1a, we need to self-consistently solve both the transport and the magnet dynamics equations. In our model, the channel spin transport is based on the spin diffusion model developed by Valet–Fert [145], The magnet-channel interface is modeled based on the interface model developed by Brataas*et al.* [146]. Both these models are well established and are used for spin transport in long channels [10].The spin diffusion formulation yields four component conductance matrices $G_{magnet,}$ $Gl_{ead,}$$G_{int}$ and G_{ch} for the elements of nano-magnets, supply leads, magnet-channel interface and the non-magnetic channel, respectively. The four components are the charge and the three spin

components. The conductance matrices relate four component voltage drop and current flow between different circuit nodes,

$$\left[I_c, I_c^z, I_c^x, I_c^y\right] = [G]_{4\times4}\left[V_c, V_c^z, V_c^x, V_c^y\right] \tag{2.1}$$

The non-magnetic channel and lead elements are modeled as π-conductance matrices with shunt G_{sh} and G_{se} as shunt and series components, respectively [11].

$$G_{sh} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & g_{sh} & 0 & 0 \\ 0 & 0 & g_{sh} & 0 \\ 0 & 0 & 0 & g_{sh} \end{pmatrix} \quad G_{se} = \begin{pmatrix} \frac{A}{\rho l} & 0 & 0 & 0 \\ 0 & g_{se} & 0 & 0 \\ 0 & 0 & g_{se} & 0 \\ 0 & 0 & 0 & g_{se} \end{pmatrix} \tag{2.2}$$

Here, $g_{sh} = (A/\rho\lambda)\tanh(l/2\lambda)$ and $g_{se}=(A/\rho\lambda)\mathrm{csch}(l/\lambda)$,l is the length of the contact, A is the area of the contact, ρ is the resistivity and λ is the spin-flip length. These conductance matrices are obtained by solving spin-diffusion equation as shown in [11]. Contact-magnet-channel interface can be described through the matrix G_{int}.

$$G_{\text{int}} = \begin{pmatrix} g & gP & 0 & 0 \\ gP & g_{se} & 0 & 0 \\ 0 & 0 & \Gamma+\Gamma^* & i(\Gamma-\Gamma^*) \\ 0 & 0 & -i(\Gamma-\Gamma^*) & \Gamma+\Gamma^* \end{pmatrix} \tag{2.3}$$

where, $g=2-r_l r_l^* - r_r r_r^*$ and $gP=r_r r_r^* - r_l r_l^*$, $\Gamma=1-r_l r_r^*$ and P is the polarization of magnet. r_l and r_r are the reflection coefficients correspond to left and right spin, respectively. The components of the interface matrix are dependent upon the *nano-magnet's* magnetization

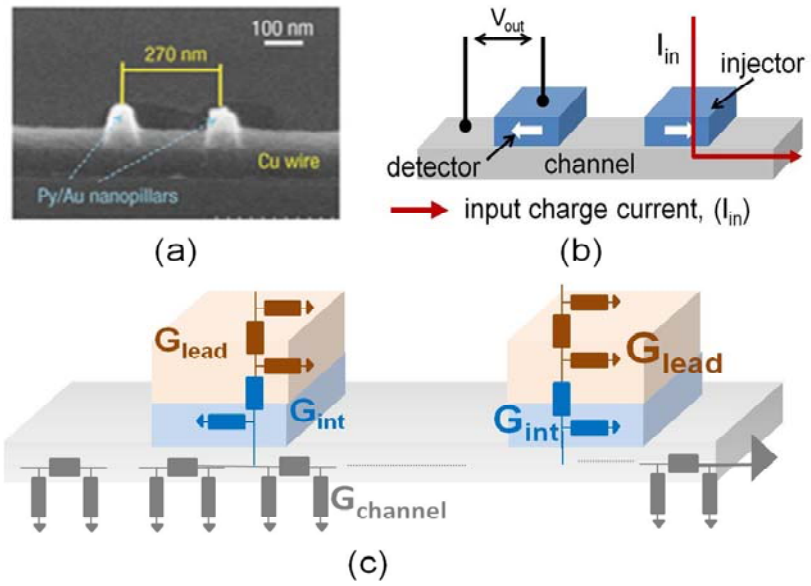


Figure 2.2(a) Fabricated LSV structure in [3], (b) Depiction of structure in fig.2 a, (c) Spin circuit model based on spin diffusion model for the device in fig. 2a

state, to be evaluated self consistently with magnet dynamics. Note that the elements of G_{sh} are responsible for the decay of spin current along the channel due to spin diffuse scattering [11].

The*Nano-magnet* dynamics is captured by solving the Landau-Lifshitz-Gilbert equation (eq.4), self-consistently with spin diffusion.

$$\frac{d\hat{m}}{dt} = -|\gamma|\hat{m}\times\vec{H} + \alpha\hat{m}\times\frac{d\hat{m}}{dt} - \frac{1}{qN_s}\hat{m}\times(\hat{m}\times\vec{I}_s) \qquad (2.4)$$

Here m is the magnetization vector, α is the damping constant, $N_S$is the number of spins in the magnet, γ isgyromagnetic ratio, $I_S$isthespin-current, which is obtained by the transport framework and H is the effective magnetic field given by eq. 5

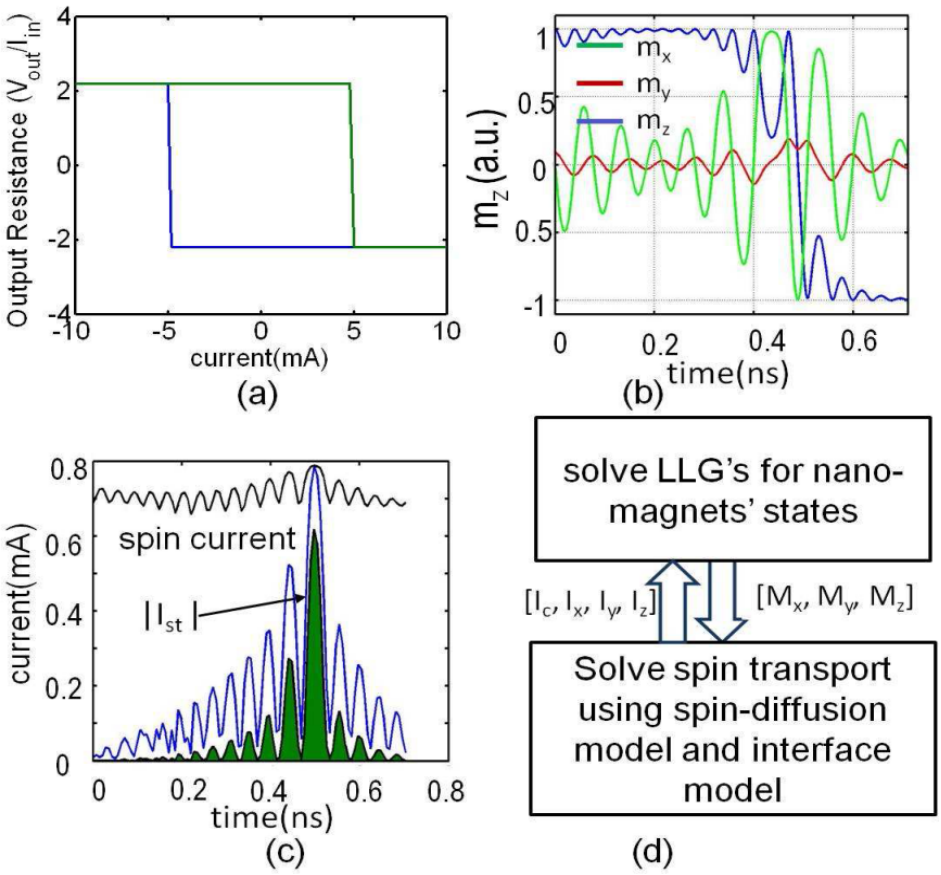


Figure 2.3(a) Calculated spin-valve signal vs input current closely matches the experimental results in [15]. STT induced switching of output *nano-magnet*. (b) Corresponding time evaluation of spin torque acting on the *nano-magnet.* (d)Self consistent solution for spin transport and LLG

$$H_{FREE} = H_{EXT} + H_k z - H_D x \qquad (2.5)$$

Where, H_{EXT} is the external field (normally zero for ASL operation), H_k is the internal uniaxial anisotropy field (along easy axis direction, z) and H_D is the demagnetization fields (along out of the plane direction, x) acting on the free-layer [10]. This simulation-framework has been benchmarked with experimental data on LSV's [10-12]. This approach leads to the mapping of a spin device structure, involving *nano-magnets*

interacting through non-local spin transport, into an equivalent “spin-circuit” [10]. The circuit model for the lateral spin valve is shown in fig. 2c.

Fig.3a is plotted using this model which shows the output voltage per unit input current for the LSV in [3] and matches closely with the experimental data. The device parameters used are as provided in [11]. Fig. 3b shows the switching of output magnet. Fig. 3c depicts the corresponding time evaluation of spin torque acting upon the output magnet. Fig. 3d summarizes the simulation framework used in this work. The spin circuit approach, discussed above.

2.3 Prospects and Challenges of All Spin Logic

The critical current required for STT induced switching scales down with magnet dimensions. As a result the ASL scheme could potentially benefit from aggressive device scaling in terms of computation energy as well as area density [11]. Analysis presented in [12] suggested the use of clocking in ASL circuits for lower computation energy. It was also shown that current-mode Bennett clocking in ASL along with hard-axis switching could achieve speed-performance comparable to CMOS. The design challenges associated with ASL can be broadly classified into categories : first, the issues related to material and device fabrication, and the second, those related to circuit techniques. Among the first, the limited spin-diffusion-length (λ) of metal channels connecting the nano-magnets can be identified as a major bottle neck for ASL. It restricts the distance over which spin signal, can be reliably transmitted [11]. Spin polarization strength of the current, decays exponentially with the distance travelled along the non-magnetic

channel.The strength of input spin signal (V_{IN}) after propagating a distance 'x' (V_X) is given by eq. 6.

$$V_x = \frac{P^2\left[\frac{2R}{R_S(1-P^2)}\right]^2 V_{IN} e^{\frac{-x}{\lambda}}}{\left[1+\frac{2R}{R_S(1-P^2)}\right]^2 - e^{\frac{-x}{\lambda}}} \tag{2.6}$$

Where, λ is the spin-diffusion length (SDL) that indicates the distance over which spin signals decay, R is the magnet resistance, P is the input-interface polarization and R_S is the channel resistance over a unit spin-flip length. Note that, Copper andgraphene have been shown to offer relatively high spin-flip length (SFL, λ) (~1μm and ~5μm respectively) [9], [145].

The other important concern is the quality of interface between the nano-magnets and the non-magnetic channel, which determines the efficiency of spin-injection. An ideal magnet-channel interface would spin-polarize all the electrons injected through it into the channel, leading to ~100% spin injection. However, an imperfect interface can significantly lower the spin-injection efficiency due to spin-flip processes associated with different sources of impurities at the interface. In this work we have assumed ideal spin-channel interfaces for the input and the output magnet-channel interface. As stated earlier, high spin-polarization constant-P (close to 1) is favorable for the input interface and vice-versa [11].

This work targets to explore the second category of design issues with stated above, namely circuit design techniques, assuming the optimal device and technology

parameters available. Although, ultra-low voltage operation for an ASL device may seem conducive to low energy switching, the overall energy-efficiency for complex ASL logic blocks may be limited by the large static current requirement for each gate and relatively low switching speed of magnets as compared to CMOS [33]. In this work we propose some design techniques to enhance the performance, energy-benefits and density of ASL, exploiting specific features of the device.

2.4 Pipelined, stacked ASL for low power high density and high performance

Performance for a large ASL block can be enhanced by the use of two-phase pipelining. Since a *nano-magnet* preserves its state upon removal of supply voltage, ASL can facilitate fine-grained pipelining, without the need of additional latches. However, this comes at the cost of power and area overhead, resulting from the clocking transistors that are used to turn the supply on and off for a given logic stage. In this work be analyze the pros and cons of pipe-line ASL design. We propose a 3-D integration scheme for ASL which can exploit a pipelined ASL design to realize ultra-high density and low power computational blocks. In the following sections these concepts are elaborated.

2.4.1 Two Phase Piplelined ASL

In this section operation of 2-phase pipelined ASL is described. Following this a brief discussion on device level optimization is given.

2.4.1.1 Device Operation

Fig. 4 shows three ASL stage connected using 2-phase pipelined scheme. The magnets m_1 and m_3 are driven by clock, whereas m_2 is driven by an inverted clock. When the clock

is high, m_1 and m_3 act as transmitting magnets. They receive charge current from a clocked transistor (not shown in the figure). The injected charge current induces a spin-

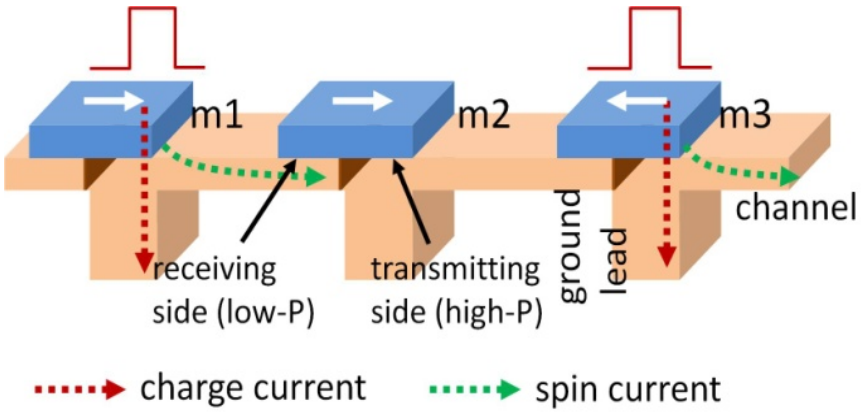


Figure 2.4Three ASL stages connected using 2-phase pipelined scheme

diffusion current on the transmitting (high-*P*) side of m_1 and m_3, which in turn, is absorbed by the receiving magnets' low-*P* side (low-*P* side of m_2 as shown in the figure). Thus, the data stored in m_1 is transferred to m_2 and that stored in m_3 is transferred to the next magnet (not shown in the figure). During this phase, m_2 is not connected to the supply and hence does not receive any charge current. When the clock goes low, the magnets m_1 and m_3 turn into receivers and are kept in the floating state. For m_3, m_2 acts as the transmitter. Thus, the information stored in m_2 during the high-clock phase is transferred to m_3 during the low clock phase.

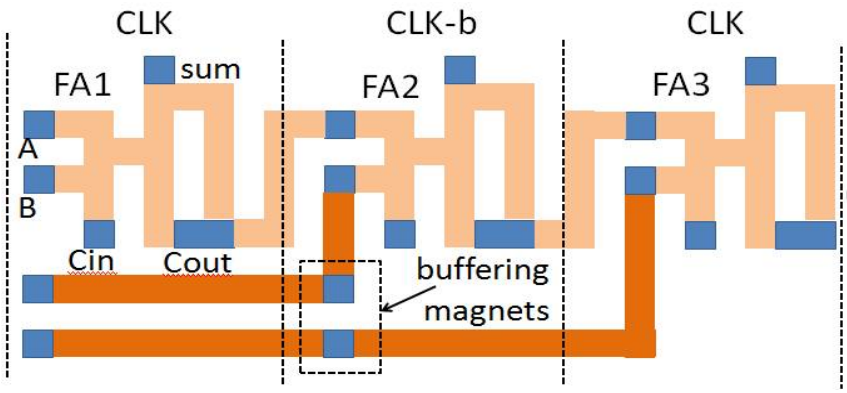


Figure 2.5ASL full adders connected using 2-phase pipelining scheme

The same scheme can be extended to an arbitrary pipelined logic design. Fig. 5 shows two ASL full adders connected using the pipelining scheme described above. Here, pipelining-granularity has been taken as a single FA. Note that a single ASL-FA evaluation corresponds to two magnet switching delays. Current is supplied simultaneously to all the five magnets in a FA. First C_{out} evaluates, based on the values of the inputs A, B and C_{in}. This is followed by the evaluation of SUM, which depends upon the state of C_{out} and the three inputs. In the same clock phase, the result of C_{out} and SUM is transmitted to the next stage FA, which is in a floating state. In this example, the clock pulse width must be at least as wide as three magnet switching delays. A finer pipe-line granularity would involve decomposing one FA evaluation two steps, namely, C_{out}evaluation and SUM evaluation.

Note that, in contrast to pipelines CMOS, a 2-phase pipelined ASL does not require any additional latch, as each of the *nano-magnets* itself acts as a latch and hence, facilitates fine-grained pipelining of a large logic block.

2.4.1.2 Device Operation

The device operation for 2-phase pipelining can be optimized by appropriate choice of device structure and operating conditions. Fig. 6a depicts the spin-diffusion model for ASL device. The device elements, namely, the metal channel, the *nano-magnets* and the magnet-channel interface are modelled as four component conductance elements, one charge conductance and three spin conductance's (G_{se}, G_{sh}: series and shunt conductance of metal channel, G_{int} : magnet-interface conductance) [9-11]. As mentioned earlier, the charge component of the spin-polarized current injected into the channel through the high-P side of the transmitter magnet passes into the ground lead. In the pipelined scheme, since the receiving magnet is in floating state, there is no charge current flow in the channel and through the receiving magnet. A part of the spin component of the input current is also lost to the ground, whereas the rest is absorbed by the receiving magnet. Increasing the length of the ground lead increases its charge resistance, as well as, its spin resistance. Hence, for a given input current I_{charge}, the spin current I_{spin}, absorbed by the receiving magnet increases with increasing ground resistance R_g (fig. 6b). The ratio of $I_{spin}$and I_{charge}can be defined as the non-local spin-injection efficiency (NLSE). Note that, experimentally ~20% efficiency for non-local spin injection has been demonstrated (with P~0.5) [10]. In this work we used 25% NLSE in simulations.

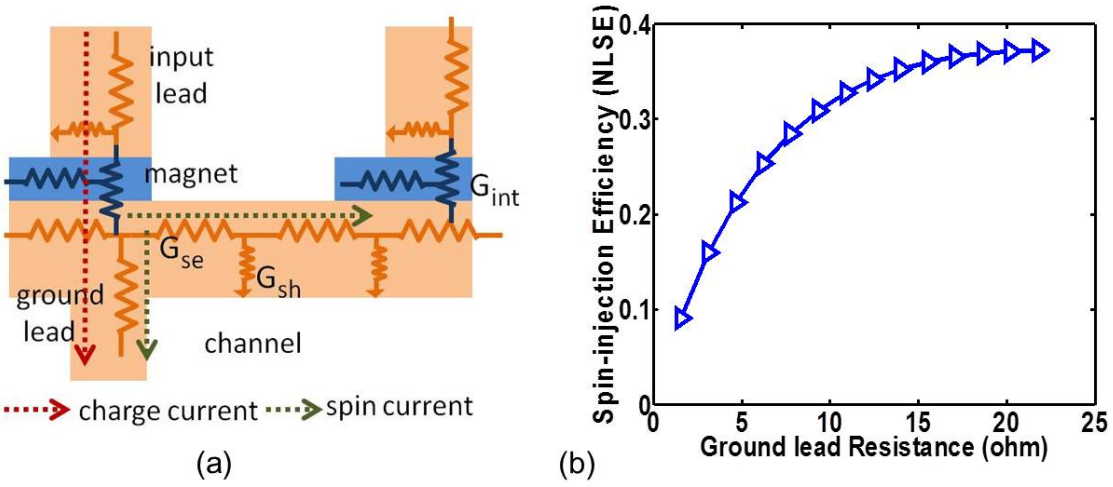


Figure 2.6 (a) Spin diffusion model for an ASL device showing an input magnet and an output magnet connected using a metal channel, (b) Plot showing increase in non-local spin injection efficiency with increasing ground resistance for 15x30x1nm^3 output magnet.

Scalability of *nano-magnets* in ASL is also tightly coupled to the value of R_g. For a given input current, scaling down the area of a receiving magnet, lowers its conductive interface with the metal channel, thereby resulting in lower spin current absorption. However, as the density of spin-current absorption remains constant, a constant switching speed is maintained (fig. 7b).Benefit of *nano-magnet* scaling however, can be obtained by simultaneously scaling of R_g. As explained above, by reducing R_g along with the magnet area, the spin injection efficiency is maintained, and hence value of spin injection in enhanced, leading to faster switching (fig 7b).

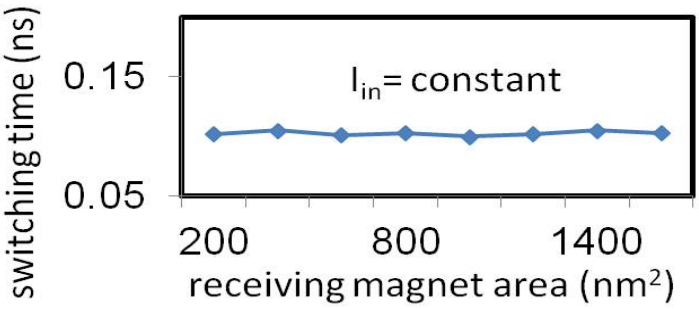


Figure 2.7 (a) Scaling of magnet area maintains a constant switching speed for a given input current and a fixed R_g

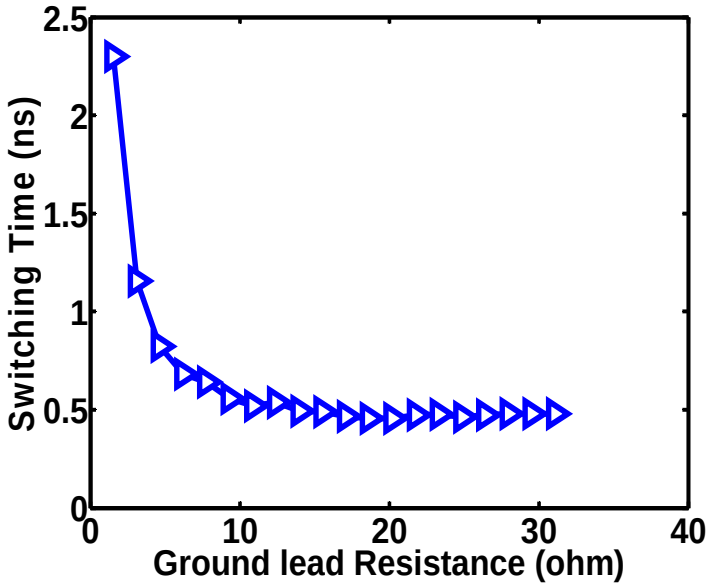


Figure 2.7b (b) Up-scaling of R_g with reducing magnet area leads to faster switching

The ultimate scaling of *nano-magnets* in an ASL device will be therefore governed by the scalability of the ground lead (or via).

An important consideration for a pipelined ASL design is the choice of clock period. For a *nano-magnet*, switching energy, E_{sw}, can be expressed as in eq. (7)

$$E_{sw} = T_{sw} \text{ x } I_{sw} \text{ x } V \qquad (2.7)$$

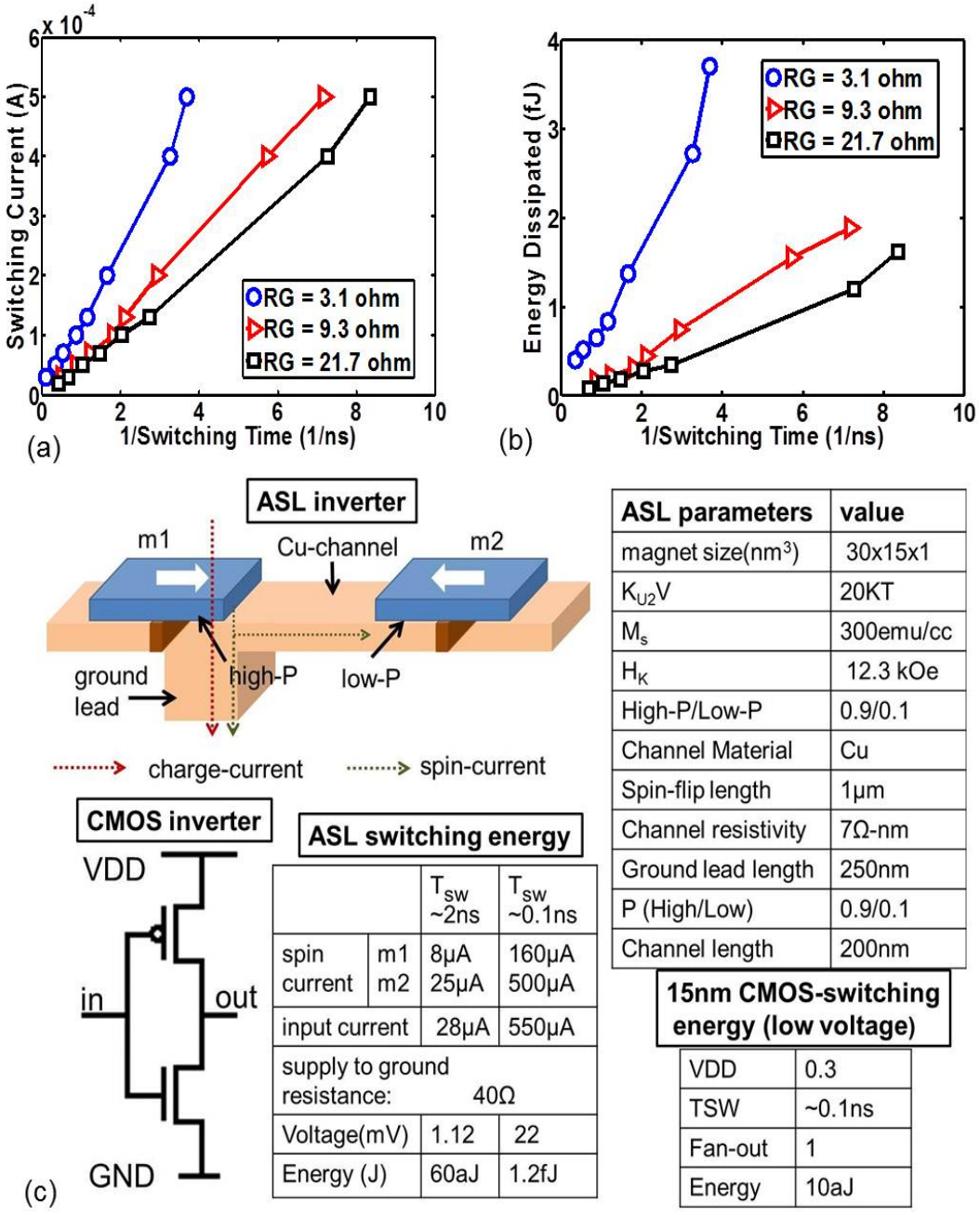
Where,T_{sw} is the switching time, $I_{sw}$is the input current and V is the terminal voltage. The minimum spin-polarized current required to switch the state of the magnet from one of its stable state to another is determined by the critical field H_C , of the magnet given by eq. 8:

$$H_C = 2\frac{K_u}{M_s} \tag{2.8}$$

Here, $K_u$is the uniaxial anisotropy constant and $M_s$is the saturation magnetization of the magnet [10]. For the spin-current I_{SW} to be able to switch the magnet of volume V, the spin-torque equivalent field,H_{sw}(eq. 9) must be greater than H_c.

$$H_{SW} = I_{SW}\ 2\ \frac{\mu_B\ q}{M_S V|\gamma|} \tag{2.9}$$

Where, $\mu_B$is the Bohr Magneton, $q$is the electron charge,γ is the gyromagnetic ratio [145]. Beyond the critical switching current, the *nano-magnet* switching timeT_{sw} is inversely proportional to the switching current to the first order (fig. 8a) [11]. Since, higher I_{sw} requires higher V, faster switching speed incurs linearly higher switching energy, as shown in fig. 8b. Thus, for low-energy operation it is desirable to operate the pipelined ASL with a low frequency clock. However, in presence of thermal noise, the probability of correct switching of *nano-magnets* reduces steeply with reducing current. Fig. 9 shows the plot for the probability of correct evaluation vs. I_{sw}/I_{cr}, where I_{cr} is the critical current required to switch the *nano-magnet* in a long enough time. Stochastic-LLG has been employed to determine this trend [9]. It involves an additional thermal noise field $h f_{x,y,z}(t)$ in the LLG equation (eq. 5).



ASL switching energy

		T_{sw} ~2ns	T_{sw} ~0.1ns
spin current	m1	8μA	160μA
	m2	25μA	500μA
input current		28μA	550μA
supply to ground resistance: 40Ω			
Voltage(mV)		1.12	22
Energy (J)		60aJ	1.2fJ

ASL parameters	value
magnet size(nm^3)	30x15x1
$K_{U2}V$	20KT
M_s	300emu/cc
H_K	12.3 kOe
High-P/Low-P	0.9/0.1
Channel Material	Cu
Spin-flip length	1μm
Channel resistivity	7Ω-nm
Ground lead length	250nm
P (High/Low)	0.9/0.1
Channel length	200nm

15nm CMOS-switching energy (low voltage)

VDD	0.3
TSW	~0.1ns
Fan-out	1
Energy	10aJ

Figure 2.8 (a) Nano-magnet switching current increases linearly with switching frequency, (b) Switching energy for ASL device increases linearly with switching speed, (c) comparison of ASL switching energy at two different switching speeds with low voltage 15nm CMOS switching energy

$hf_{x,y,z}$(t) is Gaussian distributed, with zero mean and standard deviation (σ) given by eq. 10 [9].

$$\sigma^2 = \frac{\alpha}{1+\alpha^2}\frac{2K_B T}{|\gamma|M_S V} \tag{2.10}$$

This observation implies that, apart from the performance requirement, thermal noise related bit-error rate plays critical role in determining the lower limit of switching energy achievable for pipelined ASL.

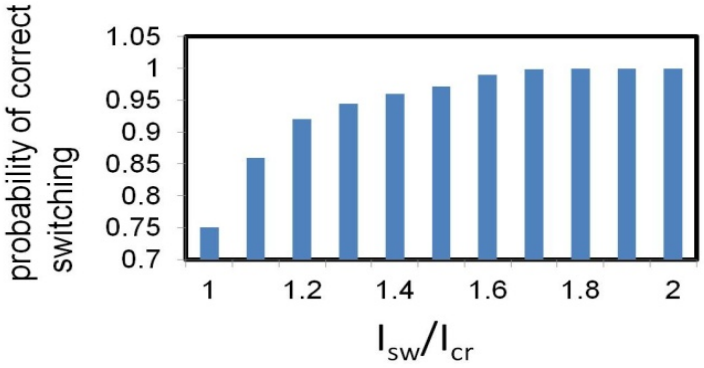


Figure 2.9Switching probability vs. I_{sw}/I_{cr}

Fig. 8c shows the comparison of switching energy for ASL inverter at two different switching speeds with low voltage 15nm CMOS inverter. The charge-current input into the input magnet of the ASL inverter (m1) must produce sufficient spin-current injection into the output magnet (m2) , required for a specific switching speed. For the ASL device parameters given in fig. 8c, the spin-injection efficiency (the ratio of spin-current input into m2 to charge current input into m1) was found to be ~25% (as discussed earlier, further improvement in this value might be possible by employing larger ground resistance, which would reduce the loss of spin current into the ground lead). For a 500MHz 8x8 multiplier, the switching delay for each full adder is required to be ~100ps.

From the results given in fig. 8c, it is evident that, for such a computing block, standard, non-clocked ASL gates operating at ~100ps gate delays may consume up to two orders of magnitude larger switching energy as compared to a low voltage 15nm CMOS design.

2.4.2 Pipelined Multiplier Design

We analyzed 8-bit carry save multiplier (fig. 10) design with ASL, using the 2-phase pipelining scheme described above. The corresponding layout is shown in fig. 10. The layout has been done using just two metal layers. This is because, due to limited spin diffusion length of metal channels, routing of spin signal though longer via's (more than two metal layers) becomes challenging(fig. 11), and may require insertion of additional buffer magnets.

Each pipelined stage constitutes of a parallel bank of full adders. Each such stage receives current from a clocked CMOS transistor. The transistors belonging to the alternate stages are driven by complementary clock phases (clk and clkb) in order to implement 2-phase pipelining, as depicted in fig. 12. Owing to comparatively large resistance of the transistors, for a given switching delay, the drain to source voltage required is significantly larger than the voltage directly applied to the nano-magnets in the non-pipelined case.

Minimum area for the ASL multiplier is obtained when the area of the clocked transistors (Area_Tx) equals that of the ASL array (Area_spin). For this case the area for the 8-bit multiplier was estimated to be ~50x lower than that of a 15nm CMOS design. But for the minimum area case, the drain to source voltage required was found to be around ~160mV for 500MHz operation, and resulted in ~9x higher power consumption

as compared to CMOS. The supply voltage and hence the power consumption can be reduced by increasing the size (number of fingers) of the clocked transistors (fig. 12a). Note that, scaling down the supply voltage and scaling up the transistor widths by the same factor maintains the level of supplycurrent per gate.As a result static power involved in ASL computation is lowered (fig .12b). However, due to

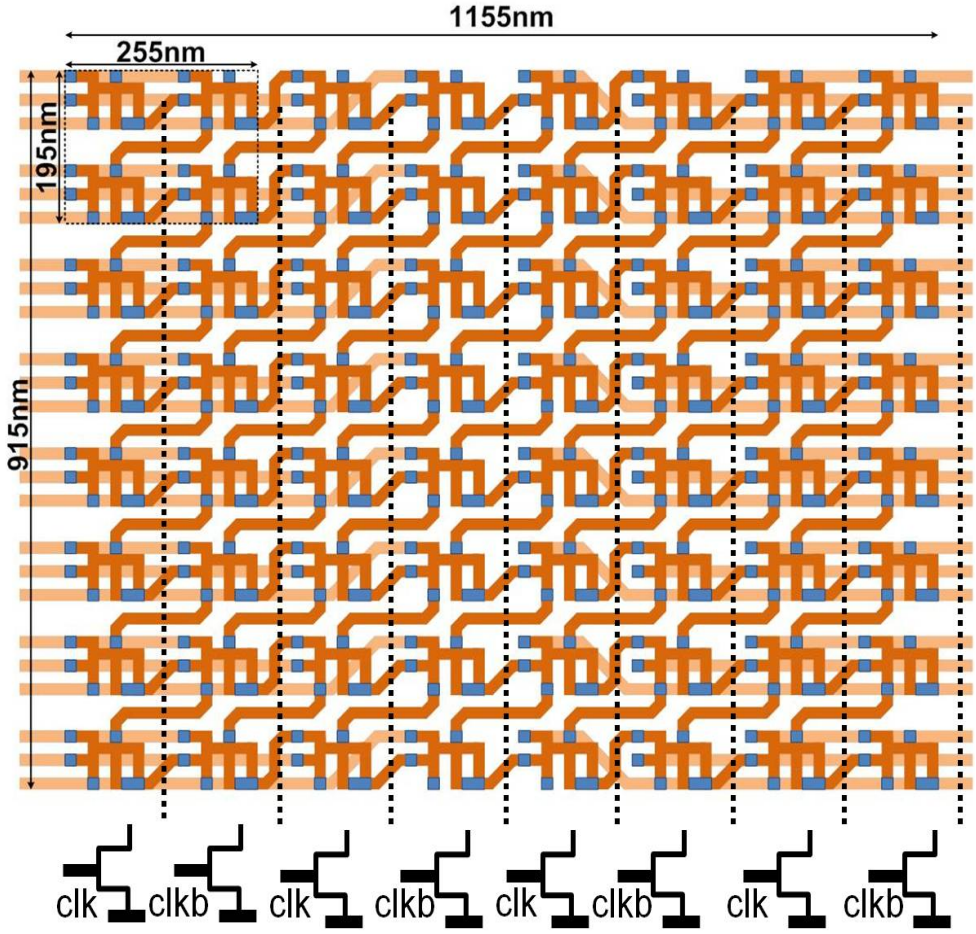


Figure 2.10ASL layout for 8-bit carry save multiplier

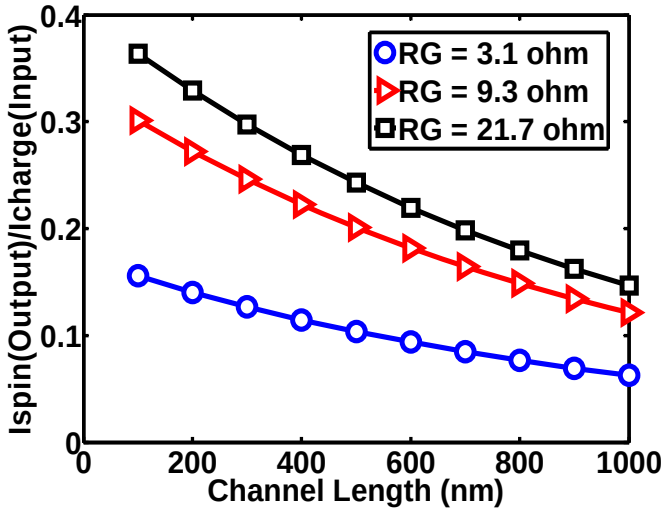


Figure 2.11Spin injection efficiency vs. channel length

increase in the width of the clocked transistors, the dynamic clocking power increases (fig. 12b). Fig. 12 b shows that for the minimum area case (i.e., transistor area = ASL area), dynamic clocking power is negligibly small as compared to the static power. But, with increasing transistor width (and hence Area(Tx)), and reducing supply voltage, the two components become comparable (for Area(Tx) = 15x Area(spin), the two components were found to be equal in simulation). Fig. 12c shows the plot for area saving obtained by the ASL multiplier over 15nm CMOS multiplier for increasing transistor width. The corresponding trend for power saving is shown in fig. 12 d. The power optimal design point can be identified as the saturation point of the total power (the sum of the dynamic and static power components) as shown in fig. 12b. At this point, the pipelined ASL design obtained ~5x lower area and 3x smaller power as compared to a 15nm CMOS design. Needless to say that, the area corresponds to transistors used for clocking the ASL gates.

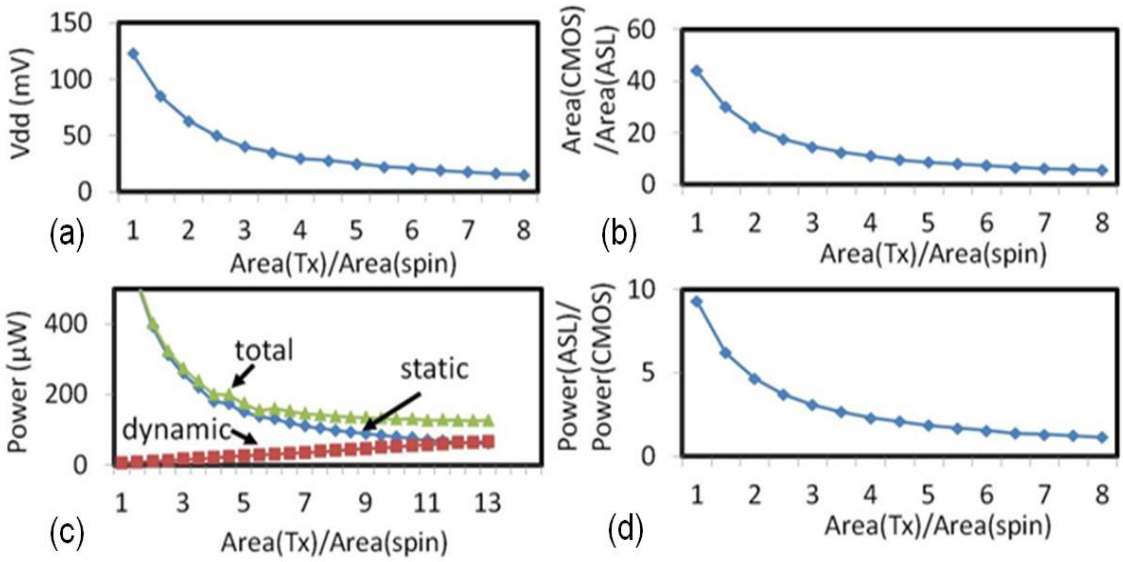


Figure 2.12(a) Supply voltage needed for pipelined ASL (for 300MHz operation) reduces with increasing Area(Tx), (b) Higher transistor width and lower supply voltage leads to reduction in static power but the dynamic clocking power increases. (c) Area benefit of pipelined ASL over 15nm CMOS design reduces with increasing Tx area (as expected). (d) Power consumption of pipelined ASL reduces with increasing Tx area and reaches a minima, but saturates after a certain point due to increase in dynamic power consumption.

From the foregoing discussion, it is apparent that in the pipelined ASL scheme, application of standard CMOS transistors for clocking leads to stringent tradeoff between power and area-efficiency, thereby eschewing the overall benefits of 2-phase clocking. However, despite achieving poor energy-efficiency, the minimum area ASL using standard CMOS transistors provides advantage in terms of robustness. For the non-pipelined design, the operating speed is determined by the critical delay path in the multiplier block (~16 full-adders delays, 8 for the multiplier-block and 8 for the merging part in the carry-save multiplier architecture). Hence, for achieving high frequency operation, current per magnet must be increased. Simulation results show that, (after including the overhead due to buffering magnets in the pipelined case) a non-pipelined

8x8 multiplier requires more than ~12x higher current injection to achieve the same performance as the pipelined design. However, since there are no transistors, the voltage required is also significantly low (~20mV, for optimized input-lead and ground resistance of ~35Ω). As a result, non-pipelined case consumes only ~1.8X higher power than the minimum area pipelined design. But, the main bottleneck of the non-pipelined design is the high current requirement, leading to unreliably high current density (~10^8 A/cm^2) in the metal leads. This amounts to ~0.5 mA of input current for each 15x30x1nm^3 magnet (with parameters given in fig. 10c) for ~60ps switching time, to achieve ~500MHz throughput. The pipelined design however could use ~40μA of current for each magnet for 1ns switching time for the same throughput.

2.4.3 3-D ASL for ultra-high density and low power computation blocks

As described in an earlier section, use of clocked transistors in pipelined ASL necessitates the use of higher voltage levels. This increases the power consumption as compared to the ideal case (with zero on resistance transistors). 3-D ASL design depicted in fig. 13 however, can overcome this disadvantage. The proposed 3-D ASL design constitutes of multiple ASL layers stacked vertically. In such a design, each horizontal, 2-D layer performs computation independently. *Nano-magnets* in multiple 2-D layers sharing the same 2-D coordinates can be supplied charge current through a common via. A spin-scattering layer can be deposited on the top of each magnet in order to prevent spin current interaction along the vertical vias.

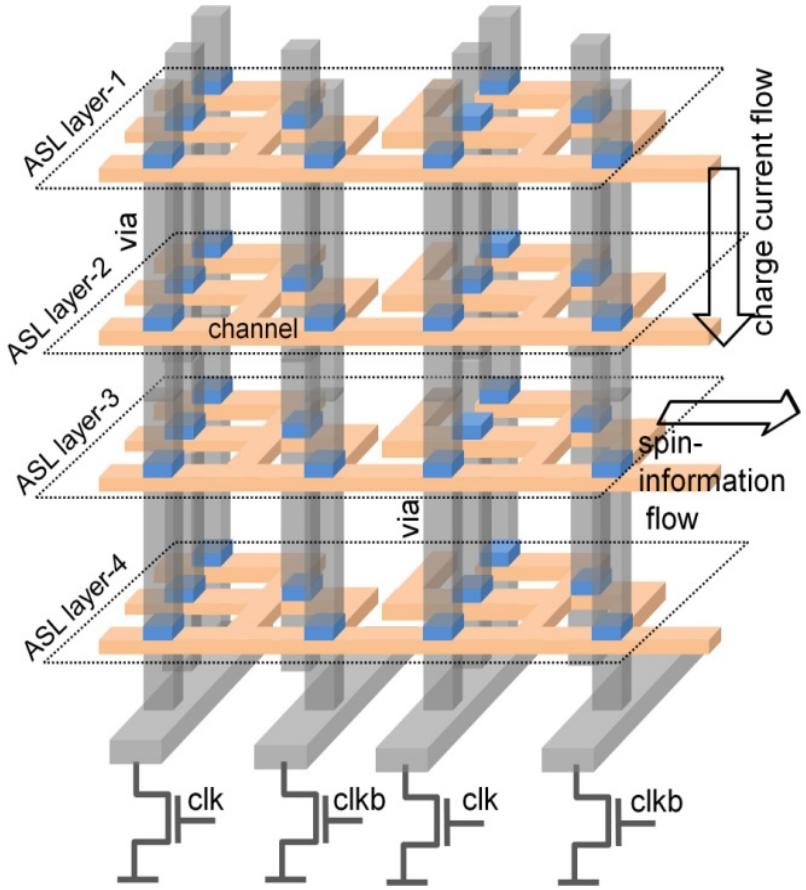


Figure 2.13 3-D ASL can be constructed by stacking 2-D ASL layer along the vertical direction. All the ASL layers in the vertical direction are supplied current using the same CMOS transistors.

In the original design, a group of *nano-magnets,* belonging to a particular logic stage in a pipeline, were clocked using a single, large transistor. For the 3-D design, the same transistor can supply current to that particular group of magnets in all the vertical stacks. Since, the overall resistance of the metallic vias is negligibly small as compared to the transistor, there is no significant increase in the supply voltage as compared to the one layer, pipelined ASL case (fig. 13). This implies that, the total power consumption remains the same as that for one layer design. Thus the effective power saving as well as area benefit over CMOS is enhanced by a factor of $N(V_I/V_N)$, where V_I and V_N are the

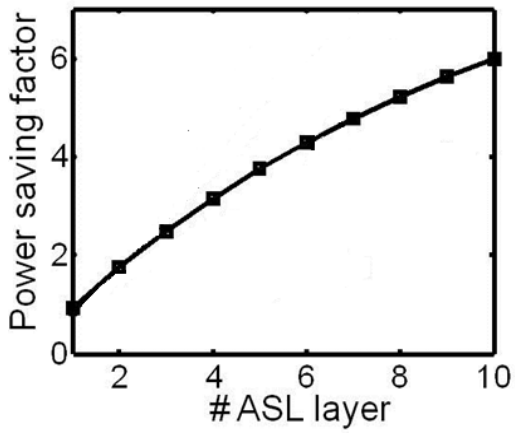


Figure 2.14 Power saving for stacked ASL vs. number of stacked layers for minimum area case.

voltages required to supply the same amount of current per-magnet for single and *N*-layer stacks respectively. Based on layouts, for minimum area case (when clocking transistor area equals ASL area), ~1μm wide 15nm transistor could supply current to 6 full-adders (FA) in parallel for 500MHz operation (total ~1mA of current). 10 stacked ASL layers offer ~ 350Ω resistance per supply lead (i.e. per magnet). This corresponds to ~11Ω load-resistance (350Ω x 5x 6 for 6-FAs) per-μm of transistor width. Due to the finite increase in the load-resistance, $V_N$increases with *N*, thereby limiting the overall power saving for larger *N*. This trend is shown in fig. 14. For N= 10, V_N was found to be ~210mV (as opposed to ~160mV for single, non-stacked pipe-lined ASL layer), thereby providing a factor of ~7.6 reduction in power as compared to the non-stacked case. As compared to the non-stacked, non-pipelined case, the overall power-saving was therefore ~15 (note that pipe-lining alone provided ~2x improvement in power saving with minimum area transistor).

It would be expected that for the stacked ASL, increasing the CMOS area should improve the power saving further, due to reduction in the required supply voltage. Notably, for the minimum area case, the dynamic power dissipated in switching the transistor was found to be ~100x smaller than the static power in the ASL. Hence there is a significant amount of room for trading of power saving with CMOS area, by increasing the transistor widths. The maximum power saving can be obtained for the case when the supply lead resistance become dominant as compared to the CMOS resistance. For the minimum area case, the ration between the transistor resistance and the supply-lead resistance was found to be ~ 3.5:1. Hence, ideally using much larger transistors (say ~10x larger), the required supply voltage can be further reduced by a factor of ~3 or more, resulting in ~98% (~50x) overall power saving as compared to the standard, non-clocked ASL. Note that, this would increase the dynamic power by a factor of ~10, however it is still small enough to be ignored as compared to the static power. Apart from loss in area efficiency, the use of larger transistors may complicate the routing of current from the transistors to the ASL devices. It is therefore advantageous to use methods that trade of dynamic power involved in clocking with the static power in ASL, without significantly increasing the CMOS area, as discussed in the next section.

2.4.4 Choice of transistor characteristics and operating point

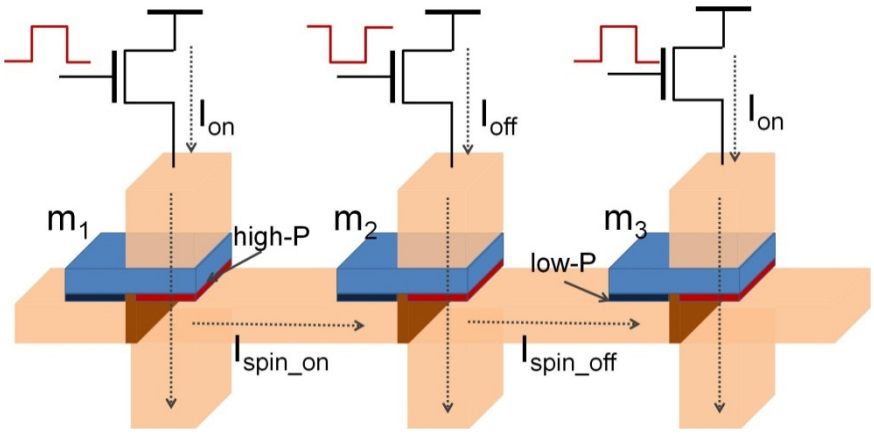


Figure 2.15 Three pipeline stages of ASL, depicting *ON* and *OFF* currents of the clocking transistors.

As discussed earlier, based on simulations we estimated that CMOS transistor of 1µm width was required for 5 ASL-FAs (10-stackes) at 500MHz, 0.7V clock, with the total terminal voltage of ~220mV. Considering ~5fF/µm switched capacitance offered by 15nm CMOS, this evaluates to ~0.5fJ capacitive switching energy per FA-stack. The static current of ~35µA flowing through each supply lead implies ~40fJ of switching energy per FA-stack (usingT_{sw}= 1ns, V ~230mV). The supply voltage for ASL can be further reduced by increasing the gate drive voltage of the clocked transistor. Another effective method that can be used in parallel is to drastically reduce the threshold voltage of the driving transistors. Assuming linear region operation of the transistors,

$$I_d \sim (V_{gs}\text{-}V_t)\, V_{ds} \tag{2.11}$$

Let's compare the results for nominal value V_{gs}andV_t (say 0.7V and 0.3V) respectively, with the case of 2xV_{gs}(1.4V), and an ultra-scaled V_t(say ~0.1V). The ratio for the two cases obtained using eq. 11 is ~3.25. This would allow an effective reduction of ~3x in $V_{ds.}$ leading to ~12fJ of energy-dissipation due to static-current per FA-stack (as opposed to ~40fJ with V_{gs} ~0.7V and V_t~0.3V). The 4-fold increase in dynamic switching power

still leads to ~2fJ of energy per FA-stack. Hence the static power, leading to ~3x power saving over the stacked ASL with minimum size-transistors with original biasing conditions. This implies ~50x power saving over standard ASL, with minimum area clocking transistors.

Assuming that 1.4V gate-to source voltage is within reliability limits of the ultra-low V_t clocking transistors, an important concern can the increase in leakage current due to the aggressive scaling of V_t. The standard expression for source to drain subthreshold leakage is given by eq. 12

$$I_{leak} = I_o \exp\left(\frac{-qV_t}{mkT}\right) \tag{2.12}$$

Where, m is a dimensionless ideality factor typically ~1.2 [146]. The off-state leakage current would further increase by about ten times for every 0.1-V reduction of V_t. For a standard CMOS IC, this would imply exponential increase in leakage power. However, for ASL, static current is inherently used in logic computation during the active clock cycle. Hence, even a drastic increase in off-state leakage current of the transistors does not significantly alter the overall power numbers as long as the ON/OFF ratio for the transistor current is significantly lessthan ~10. This is evident from the simple expression for energy-dissipation due to static current in clocked ASL, given by eq. 13

$$E_{static} = V\, T_{clk}\, (0.5\, I_{ON} + \ 0.5\, I_{OFF}\,) \tag{2.13}$$

Where, V is the supply voltage and T_{clk} is the clock period.

An important concern with the use of transistors with poor I_{on}/I_{off}ratio can be the robustness of the logic operations. Fig. 15 depicts an exemplary ASL pipeline with one magnet per-stage. The off-state leakage-current (I_{off}) injected into the channel

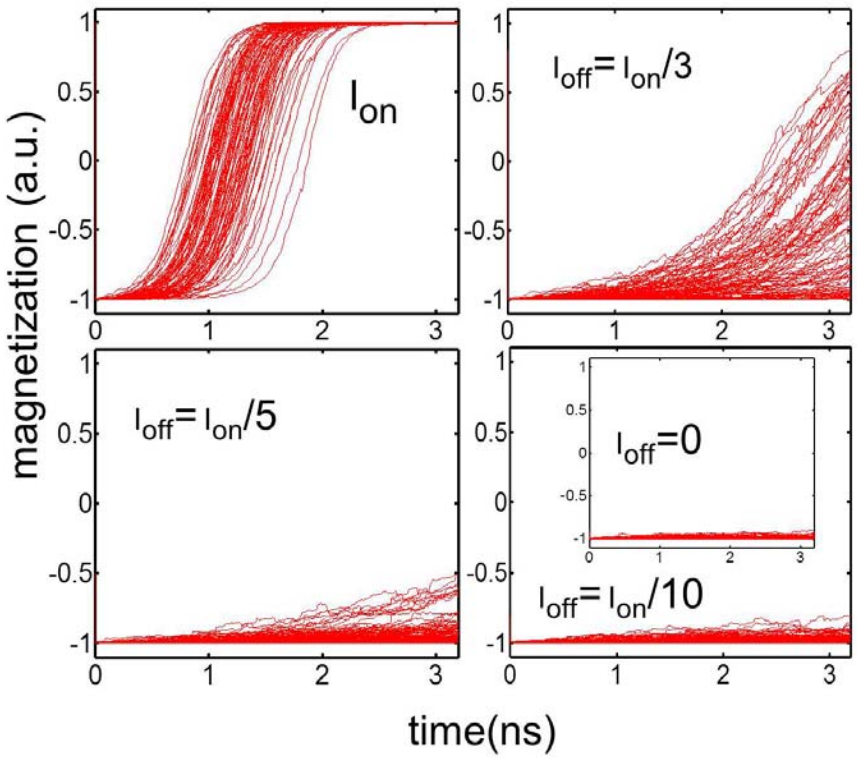


Figure 2.16Stochastic LLG simulation plots for magnet dynamics under the application of *ON*-current and different levels of-*OFF* currents.

through the high-*P* side of a receiving magnet (m_2) can disturb the state of a transmitting magnet (m_3) in the next stage. Hence a safe limit for I_{on}/I_{off}ratio must be maintained.

The switching speed of a nano-magnet increases linearly with switching current [10]. Hence, a factor of ~10X reduction in current can be expected to increase the switching delay by the same amount, thereby providing a sufficient disturb-margin. Fig. 16 shows the simulation of nano-magnet dynamics under the application of *ON*-current (corresponding to 3ns switching time) and three different levels of *OFF*-current. Stochastic Landau-Lifshitz –Gilbert (LLG) equations (eq. have been used to capture the

effect of thermal noise [9]. The plots show that an I_{on}/I_{off}ratio of ~10 may provide sufficient robustness for error-free logic flow in pipelined ASL. Note that a much lower read-disturb margin (read to write current ratio of 2 to 3) is commonly used in magnetic random access memory (MRAM). However, in ASL heating effects can be more prominent due to continuous injection of bias currents.

2.5 Performance Summary

Table-2 depicts the benefits of the design techniques presented in this work. As mentioned before, the key feature of ASL is its compactness, whereas, the energy-inefficiency resulting from relatively larger magnet-switching delay can be identified as the down-side of it. In a non-pipelined ASL design, achieving 500MHz operation for an 8x8 multiplier would mandate ~120ps switching-speed for individual magnets, requiring untenably high current-levels for magnets. This would result in large static-power in the ASL device (which is the only power component for a non-clocked, non-pipelined ASL) as shown in the figure. Introduction of minimum area clocking-transistors for 2-phase pipelining can help reduce the magnet-switching time (and hence the current-levels), leading to large reduction in power-dissipation in the device.

As mentioned before, the switching speed of a nano-magnet is proportional to the switching-current. Hence, the I^2R power-dissipation(I = current, R = device resistance)in the ASL device bears a quadratic dependence upon the switching-speed. The use of relatively high-resistance transistors however requires significantly higher voltage (~150mV as compared ~20mV), and leads to significant static-power dissipation

across the transistors. Such a minimum area design barely offers any energy benefits over non-

Table 2.1 Performance comparison of proposed pipelined, stacked ASL with standard ASL

Logic style (T_{del} ~2ns)	Vclk (V)	Vdd (mV)	I_{static} (per FA) (mA)	T_{sw} (FA) ns	C_{TX} (/FA) (fF)	E_{static} (/FA) V_{dd}. I_{static}. T_{del} (fJ)	E_{dyn} (/FA) C_{TX} V_{clk}^2 (fJ)	# of magnets (8x8 mult.)	E_{total} (pJ)	Area w.r.t CMOS
Standard ASL	-	25	3.0	0.12	-	150	0	424	12.3	1/48
Pipelined ASL, min area	0.7	165	0.16	2	1.16	53	0.56	580	5.3	1/48
Pipelined ASL (optimized Power)	0.7	20	0.16	2	12	6.4	5.8	580	1.2	1/5
Pipelined Stacked ASL N=10	0.7	220 /10 (eff.)	0.16	2	1.16 /10 (eff.)	7.0	0.056	580	0.7	1/480
Pipelined Stacked ASL (low V_t Tx) N=10	1.5	70 /10 (eff.)	0.16	2	4.6 /10 (eff.)	2.2	0.2	580	0.24	1/480

pipelined ASL, but does improve the reliability of the design by reducing the current injection for a given circuit performance by more than an order of magnitude. Area can be traded-off with power by increasing the size of the clocking transistors. This reduces the associated I^2R dissipation in the transistors. However, this results in increase in

dynamic power consumption due to increased switched capacitance. We proposed the design of 3-D ASL that can enhance both energy-efficiency and area-density, by sharing a common CMOS substrate with multiple ASL layers. The energy efficiency is improved

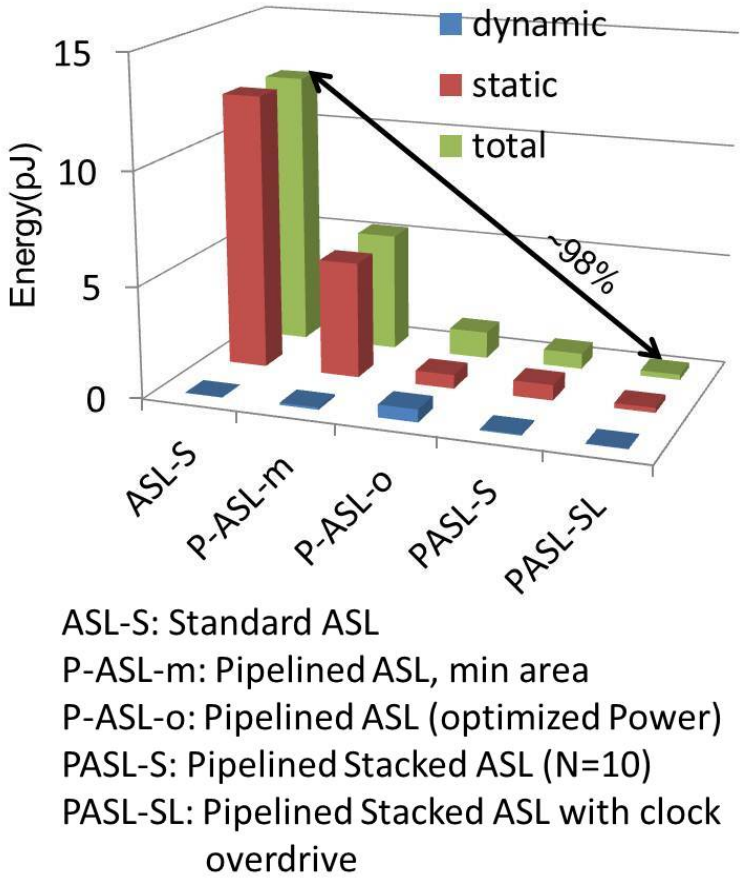


Figure 2.17 Comparing energy-efficiency of proposed pipelined, stacked ASL scheme with standard ASL

by a factor proportional to the number of layers, as discussed before. Finally, we explored an alternate method for trading of the dynamic power dissipation with its static counterpart in the stacked ASL design. We employed large overdrive voltage for the clock transistor to achieve minimal ON-resistance for the transistors. We also proposed to

use drastically scaled threshold voltage for the transistors for the same purpose, exploiting the leakage tolerance of ASL. These methods led to ~98% improvement in energy efficiency over standard non-clocked ASL (fig. 17) and ~99.8% improvement in effective area density. Close to two orders of magnitude improvement in switching-energy for ASL may render it comparable (or possibly better, with further improvement in ASL device characteristics like spin injection efficiency, efficient magnet-channel interface, low contact-resistances and impurity-scattering and high-spin diffusion length) to low voltage 15nm CMOS, as discussed in section 3.1.The most attractive feature of 3D ASL is evidently the ultra-high area-density. The prospects of achieving ~1000x higher logic density as compared to CMOS may be a motivating factor for the on-going research and experiments in this field.

3. NON BOOLEAN COMPUTING WITH SPIN TORQUE

3.1 Motivation

Most of the spin-torque based computation schemes proposed so far have been focused on digital logic, and, their benefits over robust and high performance CMOS remains debatable. Ultra-low voltage, current-mode operation of magneto-metallic spin-torque devices can potentially be more suitable for non-Boolean computation schemes like, neural networks, which involve analog processing.

The fundamental processing element (PE) of such non-Boolean hardware can be regarded as 'neuron' owing to its functional resemlance to the processing elements in the biological brain. Such a computing unit or neuron essentially performs analog summation of a number ofinputs recevied and mulitplied by input-weights called 'synapses', followed by comparison with a threshold (fig. 1). A large number of such neurons can be interconnected in different ways to realize different classes of non-Boolean computing architectures that can be functionally much superior to the conventional Von-Neumann designs for a certain class of applications. Such prospects have encouraged numerous design attempts in past that aimed to achieve the neural terms of power consumption as well as area. For instance, a digital CMOS implementation of a

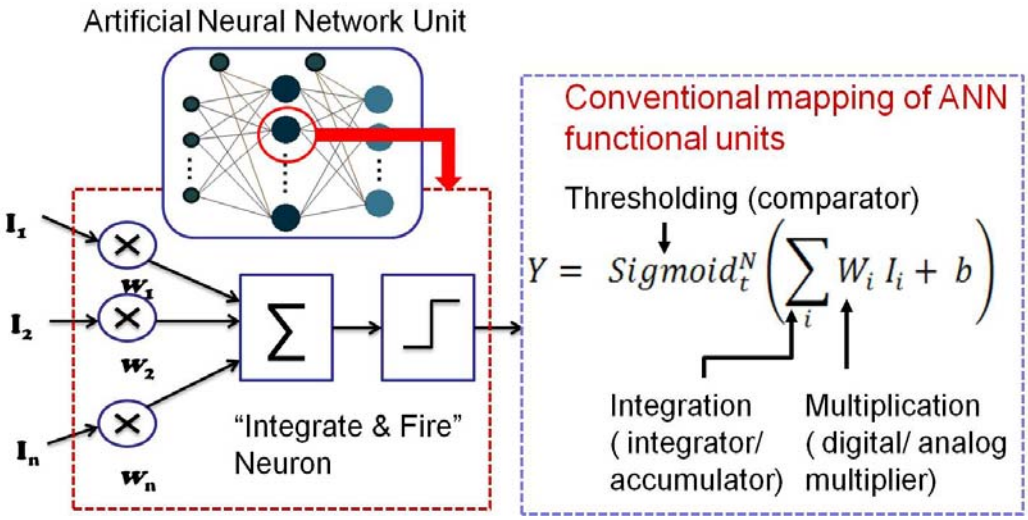


Figure 3.1Thresholding neuron as a neural computation unit with input weights W_i, called synapses.

neural computing unit would employ multipliers to emulate weights, and adders and comparators tofunctionality using CMOS transistors [19-23]. However, realization of such a functionality using conventional CMOS design incurs prohibitively high cost interms of power consumption as well as area. For instance, a digital CMOS implementation of a neural computing unit would employ multipliers to emulate weights, and adders and comparators to perform the summation and the thresholding operations respectively. Such a bulky design would fail to leaverage the algorithmic and structural benefits of the non-Boolean computing architectures altogether. It has been argued that analog CMOS circuits, owing to their compactness, can be better suited for such tasks [24]. However, large static power consumption in such circuits eschews the energy benefits of non-Boolean designs realized using analog and mixed-signal CMOS. Despite continual algorithmic developments achieved in the field of non-Boolean computing in

recent years, the aforementioned design bottelcecks have illuded the parallel effors for its efficient hardware mapping.

The emerging spin-torque devices may hold the key to this bottleneck. As mentioned earlier, low voltage current-mode switching of spin-torque devices can be exploited to obtain the critical analog functionality of summation and thresholding, needed for non-Boolean computing. However, a comprehensive solution to this research problem would involve, not only evolving spin-device models for ‘neurons’, but to explore efficient ways of integrating such devices into suitable circuits and architectures to assess the overall implementation feasibility and system-level benefits. In this work we plan to explore different device structures for "spin-neurons" and compare them with respect to computation efficiency, fabrication ease, reliability and compatibility to hetero-integration.

In the following sections, we first present the analysis of neural networks using analog CMOS neurons and estimate its performance. Following this we introduce the proposed spin device models for neurons that can out-perform conventional CMOS circuit models for neurons by orders of magnitude.

3.2 Neural Networks with resistive memory synapses and CMOS neurons.

3.2.1 Resistive memory as synaptic network

The input synapses in an analog neural-network hardware can be visualized as resistive conduits with different conductance’s (preferably programmable), connecting the analog input levels to the neuron circuit. In recent years several device solutions have been proposed for fabricating CMOS compatible, nano-scale programmable resistive elements,

generally categorized under the term ‘memristor’[39-45]. Continuous range of resistance values obtainable in these devices can facilitate the design of multi-level, non-volatile

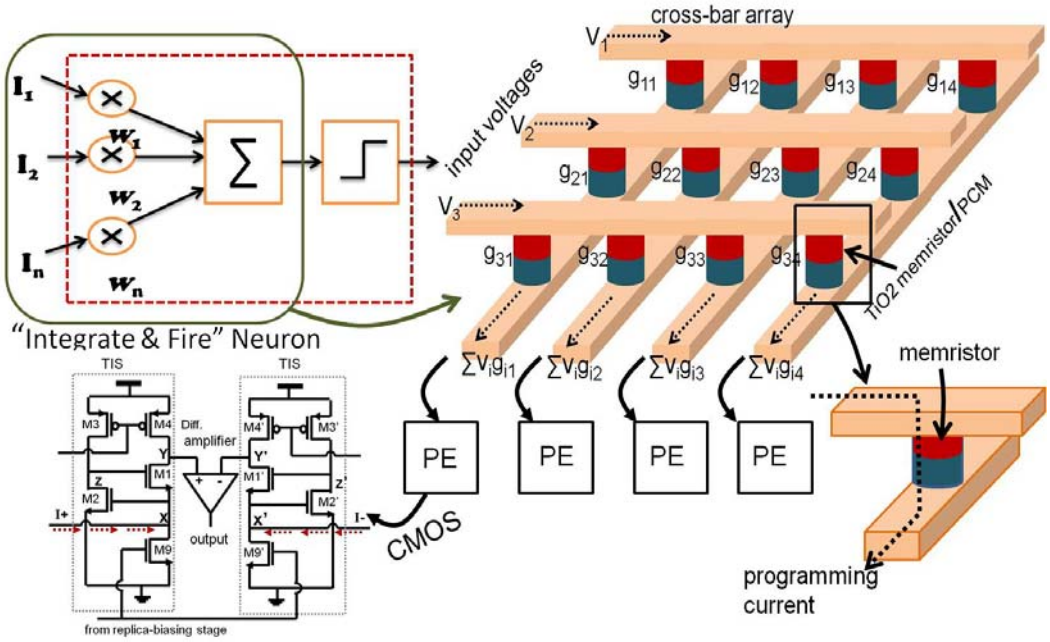


Figure 3.2Neural Computing with resistive cross-bar array : a memristors connecting a set of horizontal and in- metal lines can be programmed by applying writing pulses across the two lines[45]. For computing, inputs are applied to the horizontal lines and the current mode summationsare obtained along the in-plane lines.

memory [38-40]. Such devices can be integrated into metallic crossbars to obtain high density resistive crossbar networks (RCN) [39-45] (fig. 2). The Resistive-Crossbar Network **(**RCN) technology can be exploited for implementing non-Boolean computing architectures like neural networks, where the memory elements can be used as compact weights or synapses.[5, 11].

Fig. 1 depicts a resistive crossbar network and its analogy with the non-Boolean processing module discussed earlier. The RCN constitutes of memristors (Ag-Si) with conductivity g_{ij}, interconnecting two sets of metal bars (i_{th} horizontal bar and j_{th} in-plane

bar). The horizontal bars shown in the figure receive input currents/voltages. Assuming the outward ends of the in-plane bars grounded, the current coming out of the j_{th}in-plane bar can be visualized as the dot product of the inputs $V_i$and the cross-bar conductance values g_{ij}(fig. 1). An RCN can therefore, directly evaluate the weighted summation of analog inputs and hence provides an efficient model for synapse or weighted input connections for a neural network. Each of the in-plane bars in fig. 8 therefore can be input to an analog unit that can provide the essential neural functionality of thresholding.

Several design schemes for non-Boolean/neuromorphic hardware based on RCN have been proposed in literature [69,70] that would employ analog CMOS circuits to perform the thresholding task (fig. 1). In the following sub-section we analyze the design and performance of such an analog CMOS neuron circuit. Fig. 2b depicts an ideal circuit-model for a neuron with step transfer-function;

$$Y = \text{sign}\,(\Sigma W_i I_i + b_i) \tag{3.1}$$

where, $I_i$denote thei_{th} input to the neuron, W_i the corresponding synapse-weight and b_i the neuron-bias. The input-weights (that can be positive or negative) can be realized using compact programmable memristors. The synapse-weights are implemented using programmable conductance elements G_i(which can potentially have negative values). Input voltages V_i applied to the synapses result in a current $\Sigma G_i V_{i,}$, which can be either

positive or negative, depending upon the set of inputs and the weights. The

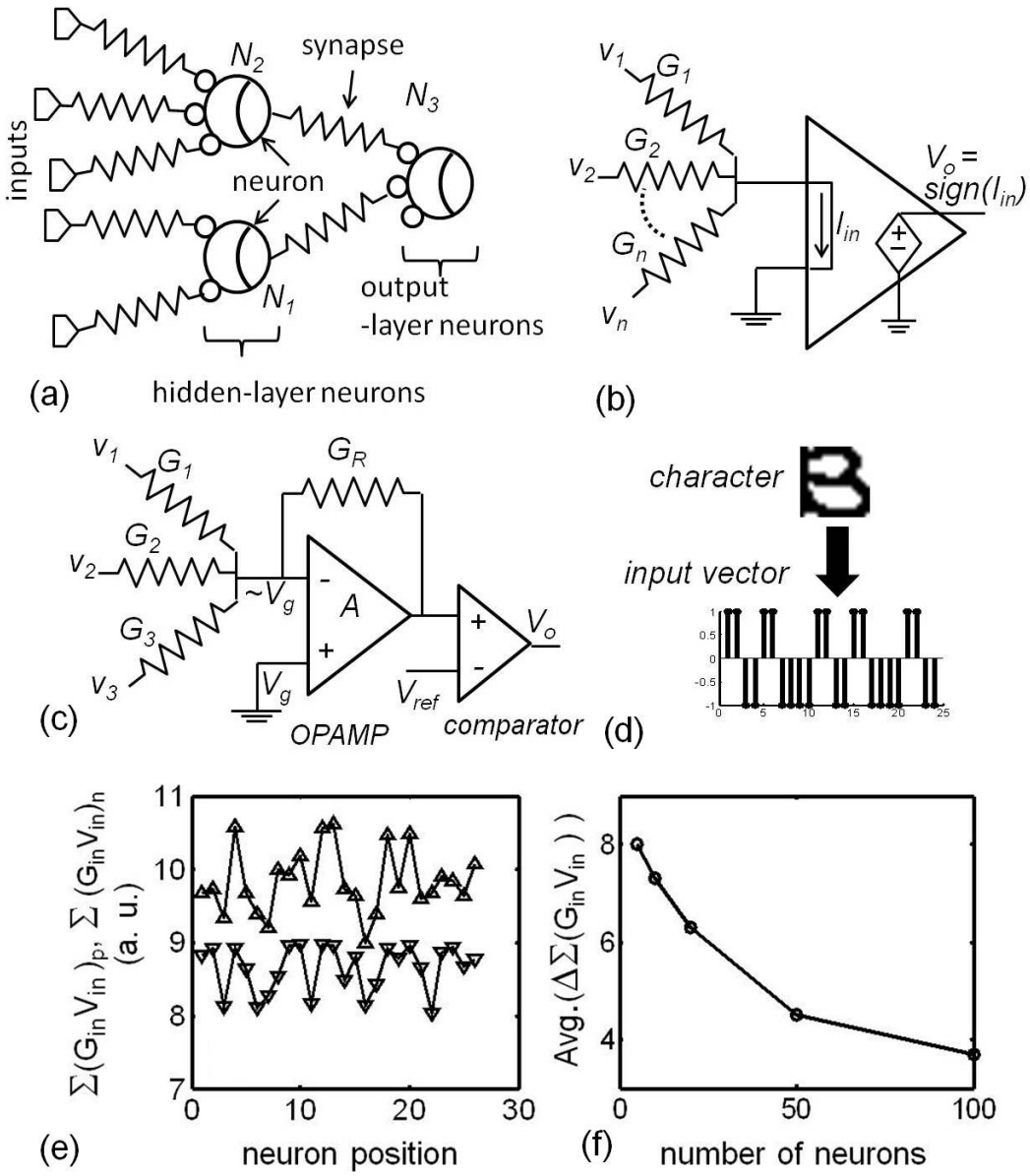


Figure 3.3(a) A feed-forward Neural Network constituting of multiple neurons, (b) an ideal circuit model for step-transfer function neuron, (c) an analog CMOS realization neuron., (d) input vector generation from character images using method described in [9], (e) $|\Sigma(G_i V_i)_p|$ and $|\Sigma(G_i V_i)_p|$ values for 26 output neurons for character-recognition operation, (f) ΔVG_i vs. number of neurons.

neuron-output, acting as a current-dependent binary voltage-source, assumes a high (+1) or a low (-1) value, depending upon the sign of the total current. It is important to note the essential input characteristics provided by the idea neuron model. The input port provides a fixed potential (in this case, ground potential) and offers small input impedance (ideally zero). This essentially implies that there is negligible change in the voltage potential at the input port. Note that any significant deviation in the input potential from a desired value would result in a net current of $\Sigma G_i (V_i - V_{in})$, where V_{in} is the non-zero input potential. This would cause erroneous network outputs when V_{in} varies randomly for different neurons.

A practical CMOS circuit design to implement the ideal neuron model presented in fig. 2b is given in fig. 2c. An operation amplifier (OPAMP) is used at the first stage of the circuit, which, for a sufficient amplification-gain, forces its two inputs to remain close to each other. Thus, by applying a fixed voltage on one of the two inputs (ground-potential V_g), the other input, (which is used as the neuron-input terminal) is also clamped to the same potential. Assuming V_g=0, the output voltage of the OPAMP can be visualized as $Vo = (1/G_R)\Sigma G_i (V_i)$, which can be positive or negative. The result is compared with zero using a comparator. For an appropriate choice of $G_R$the output voltage swing can be made sufficiently large so that a simple inverter can be used as a comparator in the second stage.

This example shows that the conventional circuit model of neuron employs an OPAMP for providing a low-impedance (fixed-voltage) input-node for linear summation of input-currents, and for transimpedance conversion of the current-mode summation, to

yield the neuron output. Thus the energy-efficiency and the performance of such a neuron model would be limited by the characteristics of the OPAMP, which is a power and area consuming circuit.

The summation term in eq. 1 can be divided into its positive ($\Sigma(G_i V_i)_p$) and negative ($\Sigma(G_i V_i)_n$)constituents. The result of the sign operation is determined by the difference between these two terms ($|\Sigma(G_i V_i)_p|$ -$|\Sigma(G_i V_i)_n|$), which is essentially $\Sigma(G_i V_i)$.

In this work, we used the network parameters for a 2-layer feed-forward neural network for character recognition. The output layer of the network has 26 neurons, each corresponding to one of the 26-alphabetic characters. In order to obtain the network weights and hence the corresponding physical values of conductance elements G_i's ,Matlab neural-network tool box was used. The inputs for training were obtained by extracting edge features from 16x16 pixel hand-written alphabetic characters. The edge-map were obtained by performing pixel-wise addition along four different direction in an input image (horizontal, vertical, and +/- 45^o) , and concatenating the resulting four vectors to form a single analog vectors. The network was trained using steepest gradient descent algorithm. For the training purpose, the step transfer-function of the spin-neuron was approximated by a 6^{th} order sigmoid function.

Fig. 2e shows the plot for $|\Sigma(G_i V_i)_p|$ and $|\Sigma(G_i V_i)_n|$ for the 26-neurons for the case when the input character belongs to the particular nodes. Note that in the basic

neuron operation defined by eq. 1, the output for a given computing step is independent of the previous state of the neuron and hence does not require any rest operation. Ideally, eq.1 depends upon only upon the sign of the net input, determined by $\Sigma(G_i V_i)$, and hence an infinitesimal positive or negative value of this sum triggers the change in the neuron's state. However, for a practical hardware $|\Sigma(G_i V_i)|$ must be larger than a

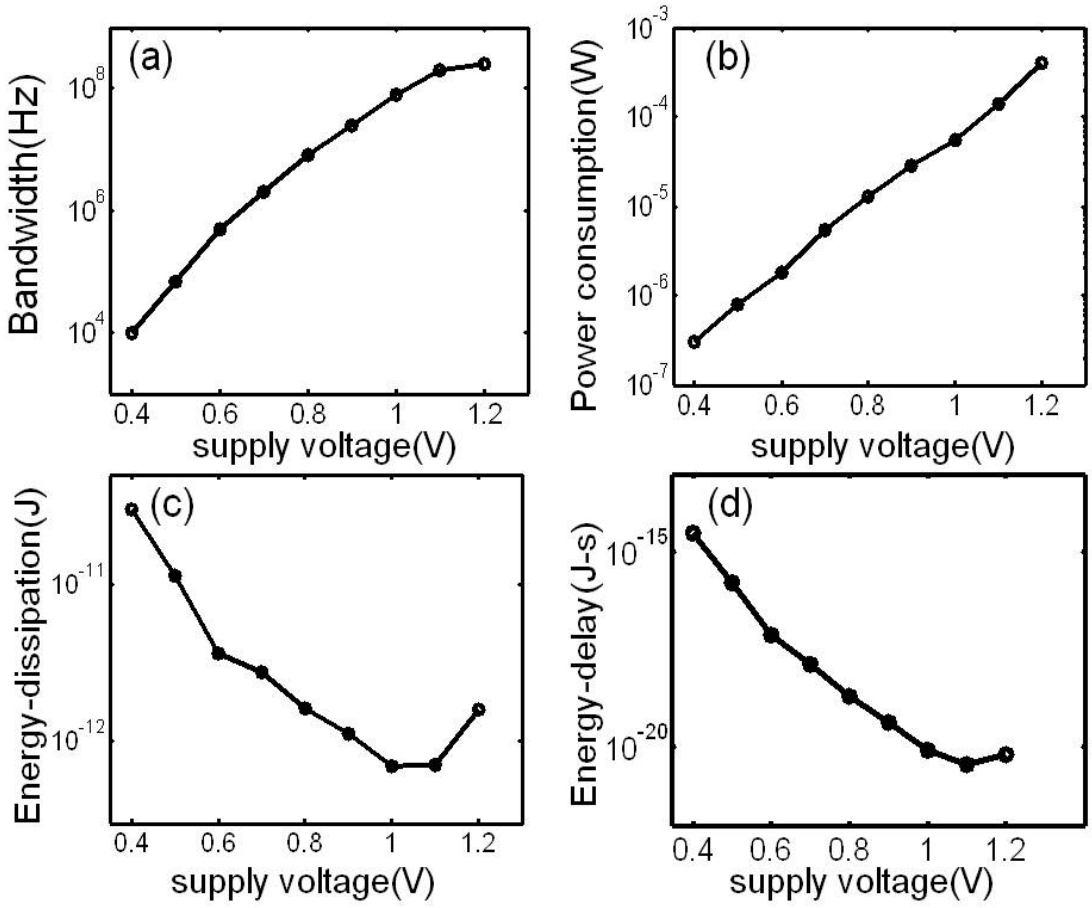


Figure 3.4(a) Bandwidth of CMOS neuron circuit vs. supply voltage, (b) power consumption vs. supply voltage, (c) energy-dissipation per computing operation of CMOS neuron vs. supply voltage, (d) energy-delay product vs. supply voltage.

minimum value to flip the state of the neuron from one of the two binary levels to the other, depending upon the sign of the summation. This enforces $|\Sigma(G_i V_i)|$ tobe greater than a non-zero constant for a practical computing circuit or a device. This minimum non-zero value, $|\Sigma(G_i V_i)_{min}|$ defines the threshold of the neuron. Results show that $\Sigma(G_i V_i)$ can be less than 10% of the total positive and negative current (denoted by $|\Sigma(G_i V_i)_p| + |\Sigma(G_i V_i)_n|$) flowing through the synapses. Thus, the resolution required for the neuron for correct operation can be defined as the ratio given in eq. 2

$$\Delta VG_i = (|\Sigma(G_i V_i)_p| - |\Sigma(G_i V_i)_n|)/ |\Sigma(G_i V_i)_p| + |\Sigma(G_i V_i)_n| \text{ x } 100 \quad (3.2)$$

Fig. 1f shows that $\Delta VG_i$for a neuron reduces with increasing number of inputs. For neurons with larger than ~25 inputs, this value can be lower than ~5%. This translates to stringent constraints upon the variations in the input voltage of the neuron. As mentioned above, any random variation in the bias voltage of the input port would result in deviation from the ideal neuron equation, resulting in computing errors.

Results show that after considering 10% σ variations in the input weights, we are left with less than 3% tolerance for the variation in the input node-voltage. For OPAMP supply as well as the binary-input level of $\pm$ 0.5 V in 45nm technology, this would translate to ~30mV of tolerance. Notably, the random offsets in an OPAMP can be few tens of millivolts. The sizing and gain of the OPAMP must be large enough to meet the offset requirements. With the aforementioned constraints, we obtained the power-consumption, delay, energy (per-operation) and energy-delay product for a 25-input

CMOS neuron-circuit shown in fig. 2c, for different supply voltages (rail to rail). The results are given in fig. 4 a-d. At the optimal point, power-consumption and the bandwidth (delay-1) were found to be around ~70μW and ~100MHz respectively. This provided an optimal energy-dissipation of ~0.7pJ per-neuron per-cycle. The energy-delay-product can be obtained as ~$3.5e^{-21}$ J-s. The maximum current per-synapse used for this case was ~3μA. Notably variability-related design constraints may become increasingly more stringent at lower technology nodes for conventional analog circuits, leading to heightened design challenges.

Thus, due to significant static-power dissipation and scalability limitations pertaining to analog CMOS neuron models, the energy benefits of RCN as a non-Boolean computing tool are significantly undermined. We next present the design and analysis of spin-torque based neuron and discuss its energy-benefits over CMOS model discussed above.

3.3 Spin based neuron models

We noted that ultra-low voltage, current-mode operation of magneto-metallic devices like lateral spin valves (LSV) and domain wall magnets (DWM) can be used to realize analog summation/integration and thresholding operations with the help of appropriate circuits, and, can be used to model energy efficient "neurons" [27]-[29]. Such device-circuit co-design can lead to ultra-low power neuromorphic computation architectures, suitable for different data processing applications. In the following sections spin neuron models proposed as a part of this work are presented.

3.3.1 Spin neuron using LSV

Ultra-low voltage, current-mode switching phenomena in magneto-metallic LSV's can be used to model energy efficient 'neurons' as described below.

3.3.1.1 Bipolar Spin Neuron

Fig 5a shows our proposed device structure for bipolar spin neuron [26, 29]. It constitutes of an output magnet $m_4$with MTJ based read-port (using a reference magnet m_5), and two anti-parallel input magnets m_1 andm_2,with their'easy-axis' parallel to that of m_4. A

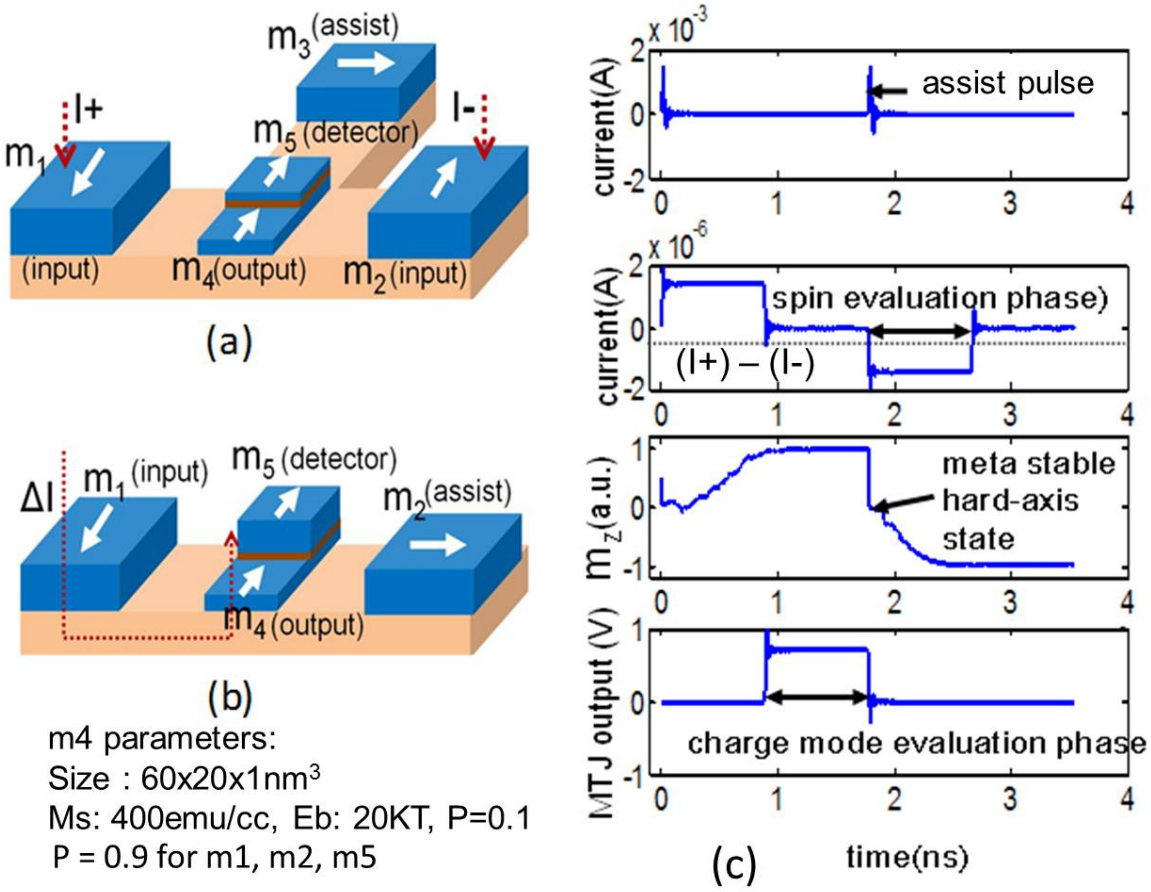


Figure 3.5(a) Device structure for bipolar spin neuron using LSV, (b) device model for unipolar spin-neuron using LSV, (c) simulation waveforms for bipolar spin neuron.

preset-magnetm_3, with an orthogonal easy-axis, is used to implement current-mode Bennett-clocking (BC). A current pulse input through m_3, presets the output magnet, m_1, along its hard-axis. The preset pulse (fig. 5c) is overlapped with the synchronous input current pulses (I_1 and I_2)received through the magnets m_1 and m_2. After removal of the preset pulse, $m_4$switches back to its easy-axis. The final spin-polarity of $m_4$depends upon the sign of the difference ΔI, between the current inputs through $m_1$and m_2.

As mentioned earlier, a neuron receives external stimulus and inputs from other neurons through synapses. The conductivity of a synapse (or weight) can be either positive (excitatory) or negative (inhibitory). Transfer-function of an artificial neuron can be expressed as the *sign-function* of sum of inputs received through multiple synapses. In the proposed device, the neuron functionality is realized by connecting the positive and the negative synapses to its two complementary inputs. The output magnet, in effect, evaluates the sign function with the help of current-mode Bennett-clocking described above [9].

3.3.1.2 Unipolar Spin Neuron (USN)

Fig. 5b shows the device structure for unipolar spin neuron based on LSV [26, 29]. In this case, a single input magnet, m_1, receives the difference of current from positive and negative synapses. This implies that the subtraction between the two current components is carried out in charge-mode, outside the neuron device. As this device receives only the difference ΔI between the two current components.

The minimum value of ΔI that can be correctly detected by the LSV-neurons, determines the resolution of the device. As explained earlier, a smaller ΔI allows smaller

input current levels, leading to smaller power dissipation. The lower limit on the magnitude of the resolvable current input ΔI for the LSV-based neurons (hence, on current per-input for the neuron),for deterministic switching, is imposed by the thermal-noise in the output magnet, and, imprecision in Bennett-Clocking (like misalignment). We estimated the resolution of the spin neurons by including such non-ideal effects. Stochastic LLG was used to assess the effect of thermal noise and ~10^{o} of misalignment for the hard-axis magnet was included (fig. 6a). As for spin-torque

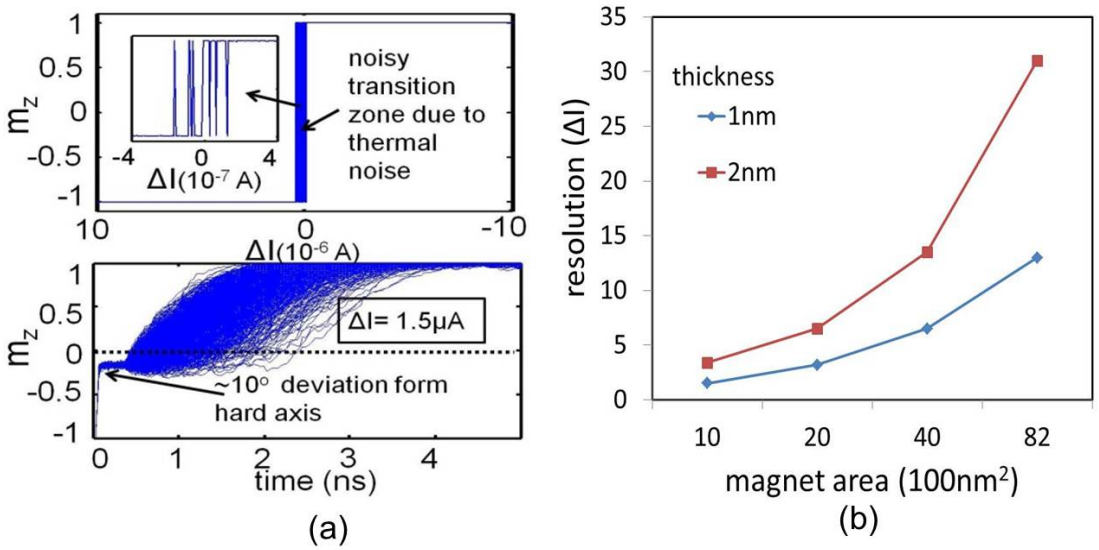


Figure 3.6 Due to noise in the neuron-magnet and imprecise BC (leading to $m_z \neq 0$ during preset), larger ΔI (hence, current for inter-neuron signaling) is required for correct switching, than the ideal case. Minimum input current level can be determined on the basis of bit error rate (BER) resulting from these effects. (transients show correct switching for 10000 runs with ΔI =1.5μA for 60x20x1 nm^3 magnet, i.e., BER<0.01%), (b) resolution of spin-neuron vs. magnet size estimated using stochastic LLG simulations

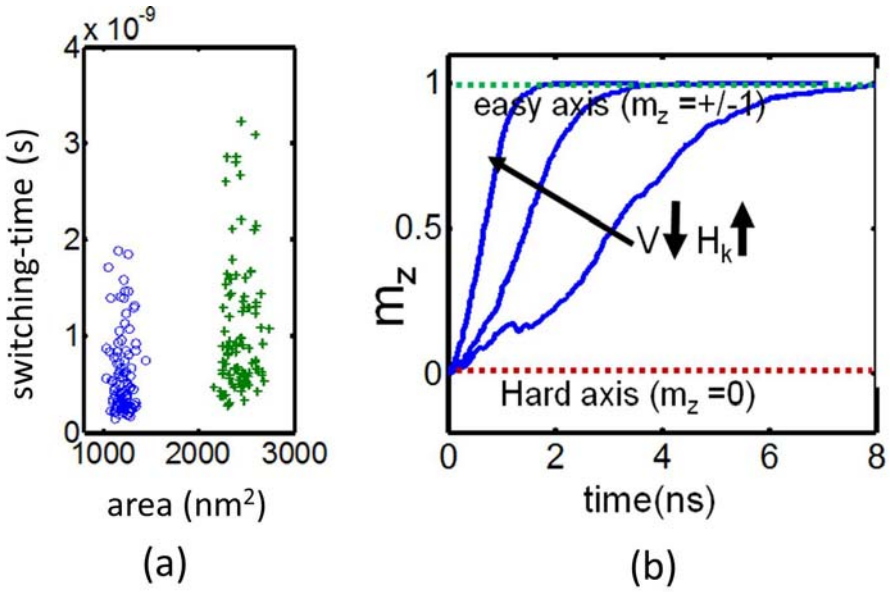


Figure 3.7(a) spread in switching time (hard-axis to easy axis relaxation) for two different magnet areas (with 20% variation), showing larger time spread for larger area, under same ΔI., (b) effect of magnet scaling on easy-axis relaxation time.

switching between easy axis (without hard-axis assist), the input current ΔI required for deterministic switching with current-mode Bennett clocking also reduces with the size of the output magnet (fig. 6b). Transient results for a highly scaled output magnet ($20x60x1nm^3$) is depicted in fig. 6a , showing the possibility of ~1μA resolution.

The switching time for the proposed spin neuron is mainly determined by the time taken for hard-axis to easy-axis transition, after the removal of the assist pulse. Fig. 7a shows the effect of area variation upon the switching time, depicting that scaling down the output-magnet in fact reduces the spread in switching time for the same current (same is true for spread in other critical magnet parameters like Ms and α). This is because, a smaller magnet with higher critical field H_c shows a faster response to the spin current ΔI.

The 4-component spin-circuit model described in chapter-2 has been used for the LSV neurons. Several other non-idealities like the effect of spin-scattering at an imperfect magnet-channel interface and in the metal channel have not been considered in this work. Non-local spin-injection can also be used for the LSV-based neuron model described above. In that case, the MTJ can cover the entire free-layer and the input current flows into a ground lead located below the output magnet [28]. However, the spin-injection efficiency for the non-local case is expected to be significantly lower, as opposed to the local case, as discussed in chapter-2.

Next, we present spin neuron based on domain wall magnet (DWM).

3.3.2 Spin neuron using Domain Wall Magnet (DWM)

3.3.2.1 Domain wall magnet: simulation-model

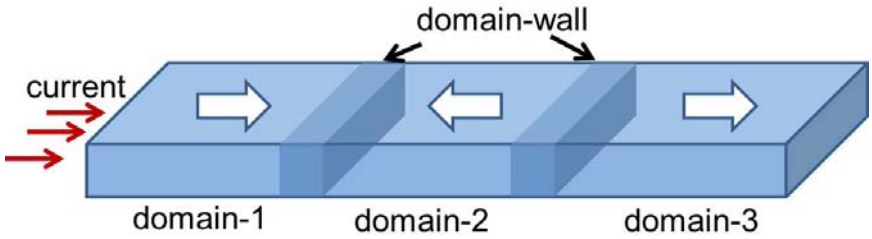


Figure 3.8A domain wall magnet strip with three spin domains

A domain wall magnet (DWM) constitutes of multiple nano-magnetic domains separated by non-magnetic regions called domain wall (DW) (fig 8). The DW can be moved along a magnetic nano-strip using current injection [31]. For instance, in fig. 8, electrons passing from the right-spin domain-1 to the left spin domain-2 are right-spin polarized. These spin-polarized electrons exert spin- torque on the regions of domain-2 close to the

domain wall and subsequently result in reversal of spin polarity of those regions in domain-2. This effectively moves the DW between domain-1 and domain-2 towards right in fig..8. This phenomenon can be employed to switch the spin-polarity of a particular region in the magnetic nano-strip. Some of the spin-domains in such a nano-magnetic strip can be selectively stabilized while others can be designed to be switchable through current induced domain wall motion.

In the following subsections we describe the modeling of spin-torque induced domain wall motion which has been employed to model spin-neuron in this work. Modeling of domain wall dynamics consists of two parts, first, the simulation of spin-polarized current transport along the magnetic nano-strip and second, the solution of magnetization dynamics in the nano-strip.

3.3.2.1.1 Spin-polarized current-transport along the magnetic nano-strip

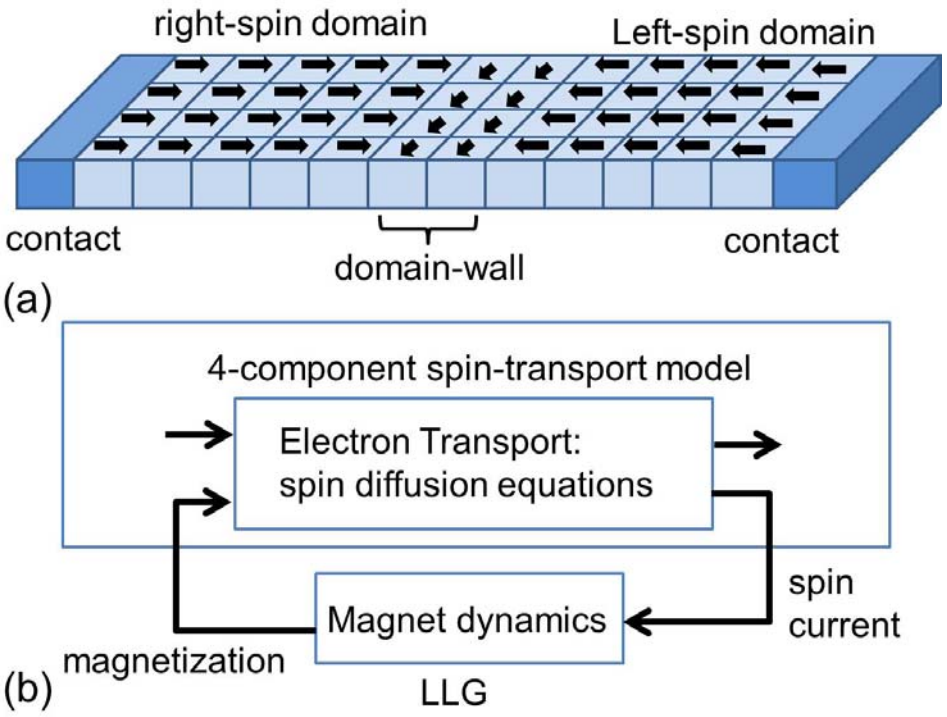


Figure 3.9(a) nicro-magnetic, multi-domain simulation model for domain wall magnet, with a magnetic-nano-strip divided into small nano-magnetic grids, (b) self-consistent simulation of spin-transport and magnetization dynamics, as proposed in [31].

Figure 9 illustrates the magnetic nano-strip we have considered for transport calculation. The structure consists of a matrix of nano-magnets (mi,j) obtained by dividing the nano-strip into a two dimensional (x-z) square grids. For example, in a nano-magnet strip with dimensions200nmx60nmx10nm (length:200nm and width:60nm) a grid size of $10x10x10nm^3$ results is total 120 square grids. Each grids has separate magnetization vectors as shown in fig. 9. Each nano-magnet is modeled as a Π conductance-network [10, 31] with shunt and series components, G_{0F} and G_F (Four Component Spin Transport model), respectively using Valet-Fert diffusion model [66] and interface model by Brataas [65]. Note that both G_F and G_{0F} are matrices with 4x4 elements as shown:

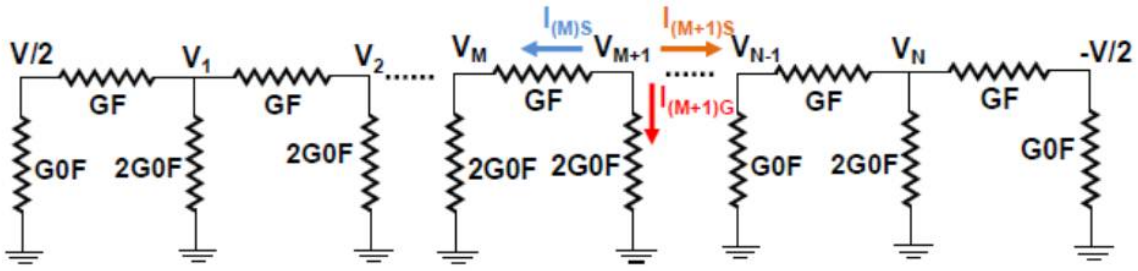
$$GF = \begin{bmatrix} \frac{A'}{\rho l} & gP & 0 & 0 \\ gP & g_{scz} & 0 & 0 \\ 0 & 0 & g_{sex} & 0 \\ 0 & 0 & 0 & g_{sey} \end{bmatrix}$$

$$G0\Gamma - \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & g_{shz} & 0 & 0 \\ 0 & 0 & \Gamma+\Gamma^* & i(\Gamma-\Gamma^*) \\ 0 & 0 & -i(\Gamma-\Gamma^*) & \Gamma+\Gamma^* \end{bmatrix} \quad (3.3)$$

where A' is the magnet area, ρ is the resistivity, l is the length. Note that g_{shz} is the shunt component in z-direction, $g_{se[x,y,z]}$ are series conductances in x,y and z directions[31]. Γ is given as

$$\Gamma = 1-r_l r_r^* \quad (3.4)$$

where r_l and r_r reflection coefficient of left and right spin respectively. Figure 10 shows the equivalent network of resistors corresponding to the magnetic nano-strip. I-V relations are solved for each elemental magnet using $I=GV$. Both I and V contain 1x4 elements, where one component corresponds to charge (V_C, I_C) and other three correspond to three spin components, one for each direction: (V_X, I_X) for x, (V_Y, I_Y) for $y$and (V_Z, I_Z) for z. By solving KCL at all the nodes, one can estimate spin-voltages and spin-currents at nodes 1 through N for a given voltage (V) across the nano-magnet strip.



$$[I]_{1x4(M)S} = [GF]_{4x4}\ ([V]_{1x4(M+1)} - [V]_{1x4(M)})$$
$$[I]_{1x4(M+1)G} = 2[G0F]_{4x4}\ ([V]_{1x4(M+1)})$$
$$[I]_{1x4(M+1)S} = [GF]_{4x4}\ ([V]_{1x4(M+1)} - [V]_{1x4(M+2)})$$

Figure 3.10Basic equations and models used in the simulation framework, illustrating the circuit representation of the Four Component Spin Transport Model for spin diffusion and LLG for magnetic dynamics. The current, voltage and conductance metrics for each elemental nano-strip has been shown.

Spin-currents exert torque on the elemental nano-magnets and are given as inputs to the magnetic dynamics calculation, which is presented next.

3.3.2.1.2 Domain Wall Dynamics Using 2D LLG with Spin-Torque

The magnetic dynamics of each nano-magnet can be described by the LLG equation (Eq. 5), where *m* is the magnetization vector.

$$\frac{dm}{dt} = \pi - |\gamma| m \times H + \alpha\, m \times \frac{dm}{dt} - \frac{1}{qN_s} m \times (m \times I_s) \tag{3.5}$$

Where H is given as:

$$H = H_{DIP} + H_{EX} + H_{DMAG} + H_{Ku} \tag{3.6}$$

Where, H_{DIP} (eq. 7) is the dipolar field, H_{EX} is the exchange field (eq. 8) ,H_{DMAG} is the demagnetization field (eq. 9) and H_{Ku} is the anisotropy field (eq. 10).

$$H_{DIP} = M_S V \frac{(3(m.r)r - mr^2)}{r^5} \tag{3.7}$$

$$H_{EX} = \frac{2A}{3M_S}\sum_{nn=1}^{8} m_{nn} \tag{3.8}$$

$$H_{DMAG} = -4\pi M_S m_x \overline{x} \tag{3.9}$$

$$H_{ku} = \frac{2K_u}{M_S} m_z \overline{z} \tag{3.10}$$

Where, M_S is the saturation magnetization, A is the exchange energy coefficient, K_u is the anisotropy energy constant. $I_S$is the spin-current which can be obtained from spin transport calculation described earlier. LLG contains three terms on the right side, where the first one is related to the dynamics under magnetic field (H), the next one is related to the damping torque term and the third is related to spin-current (I_S). Since each square grid is modeled as a magnet, we need to solve LLG for each magnet with a given total magnetic field H and spin-current I_S. Moreover, every square grid is coupled to its neighbors and as a result LLG of each magnet need to be solved self-consistently with LLGs of neighboring magnets. In the given example, we have considered 120 square grids and hence 120 LLGs need to be solved self-consistently.

Once the magnetization of each grid is calculated, the magnetization information is given as input to Four Component Spin Transport model for calculating spin-currents and spin-torque at the next time-instant. The steps (LLG and spin-transport) are repeated for every time-step (~psec) until the simulation time is completed. We have benchmarked the proposed simulation framework with experimental demonstration of DW movement. Figure 11 illustrates the average DW velocity with current density and shows good match with experimental data in [5].

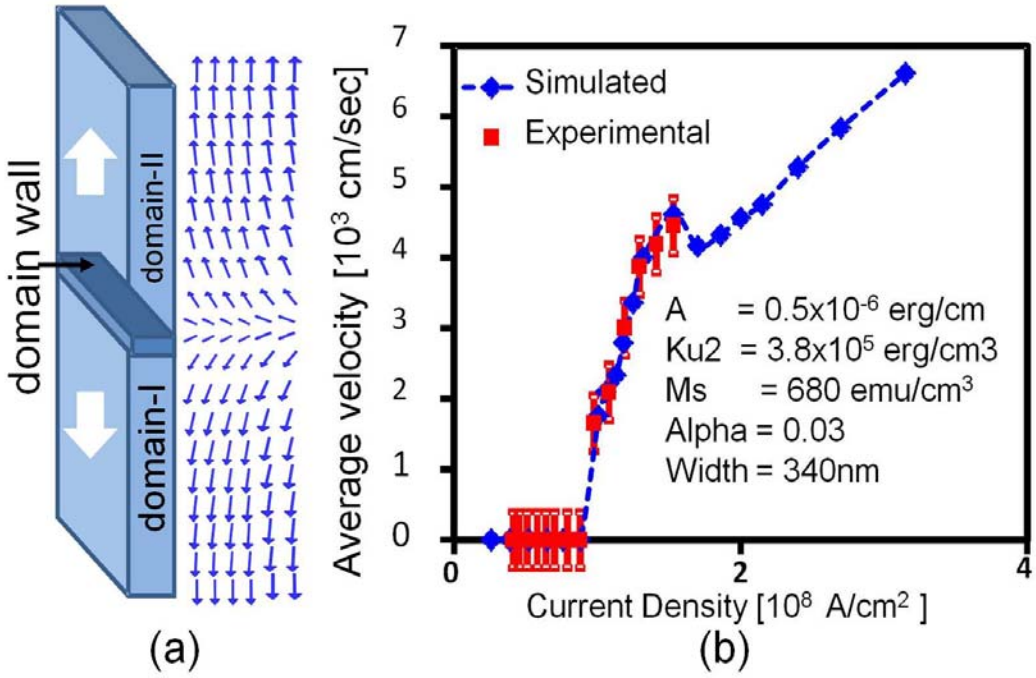


Figure 3.11(a) Fig. 2.Domain wall magnet (b) DW velocity as a function of current density with experimental data in [5]

3.3.2.2 Spin Neuron based on Domain Wall Magnet

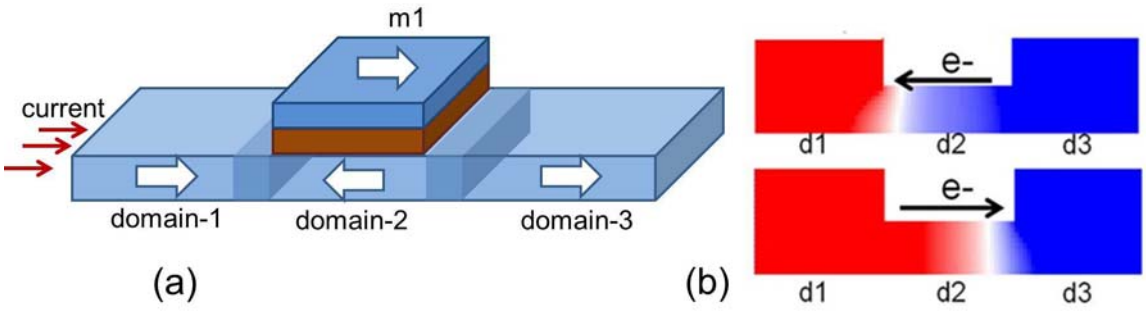


Figure 3.12(a) Spin neuron based on domain wall magnet (b) micro-magnetic simulation for neuron switching.

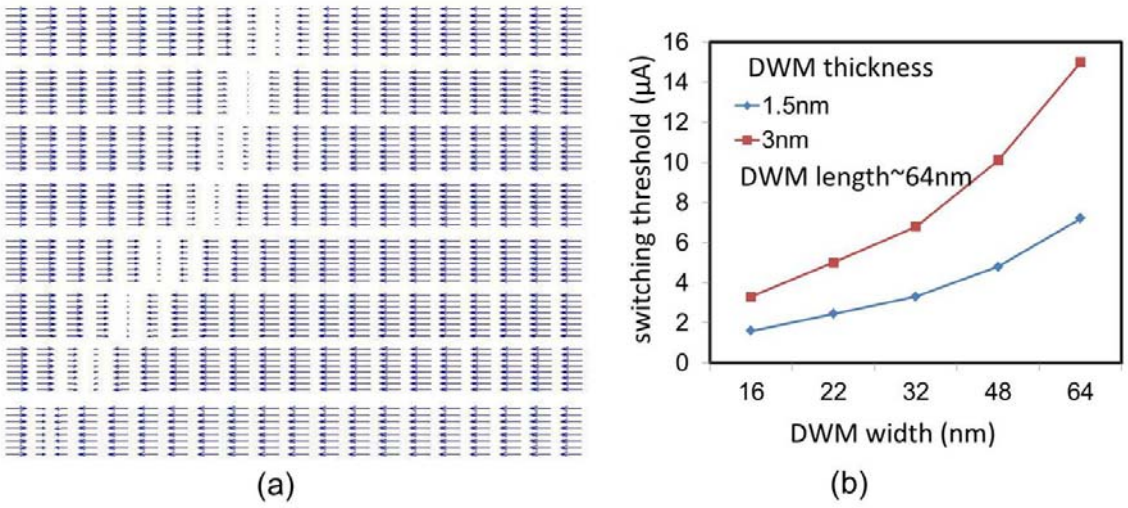


Figure 3.13 (a) micro-magnetic simulation plot for DWM neuron with free layer size ~48x16x1.5nm^3 with an input current of ~2μA and total simulation time of ~2ns (snapshots at equal time steps for the 1.5ns simulation time have been presented) (b) scaling of switching current threshold with free-layer size.

Recent experiments have achieved switching current-density of less than ~10^7A/cm^2 for nano-scale DWM strips, and, a switching time of less than 1ns [13, 14]. The current threshold as well as the switching time of DWM may scale down with device-dimensions [15]. Thus, it might be possible to design a DWM switch with highly scaled dimensions (say 20x60x2nm^3) that achieve a switching current threshold of the order of 1μA [16, 18]. Such low-resistance, low-current magneto-metallic switches can operate with small terminal voltages and can be suitable for analog-mode, non-Boolean computing applications. Such a DWM based spin-neuron structure is shown in fig. 12. It constitutes of a thin and short (16x48x1.5 nm^3) *nano-magnet* domain, d_2(domain-2) connecting two anti-parallel *nano-magnet* domains of fixed polarity, d_1(domain-1) and d_3(domain-3). Domain-1 forms the input port, whereas, domain-3 is grounded. Spin-polarity of the DWM free-layer (d_2) can be written parallel to $d_1$or $d_3$by

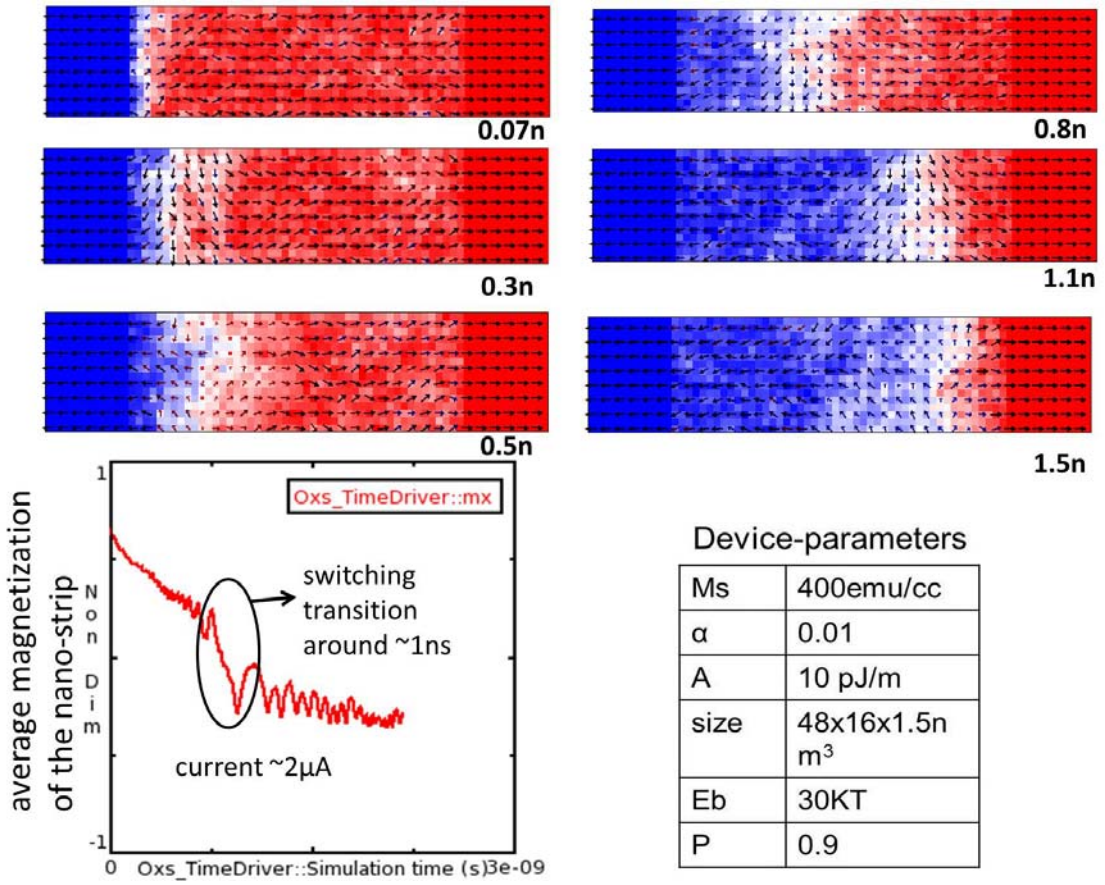


Figure 3.14 OOMMF simulation results for neuron switching

injecting a small current along it from $d_1$to d_3 and vice-versa. A magnetic tunnel junction (MTJ) formed between a fixed polarity magnet m_1 and d_2 is used to read the state of d_2. The effective resistance of the MTJ is smaller when m_1 and d_2 have the same spin-polarity and vice-versa ($R_{parallel}$~5kΩ and $R_{anti\text{-}parallel}$~15kΩ).Thus, the neuron can detect the polarity of the current flow at its input node. Hence it acts as an ultra-low voltage and compact current-comparator that can be employed in energy efficient current-mode data processing [26]. A non-zero current threshold for DW motion would result in a small hysteresis in the neuron switching characteristics. It is desirable to reduce the threshold to get closer to the step transfer function of an ideal comparator. Fig. 13 shows that for

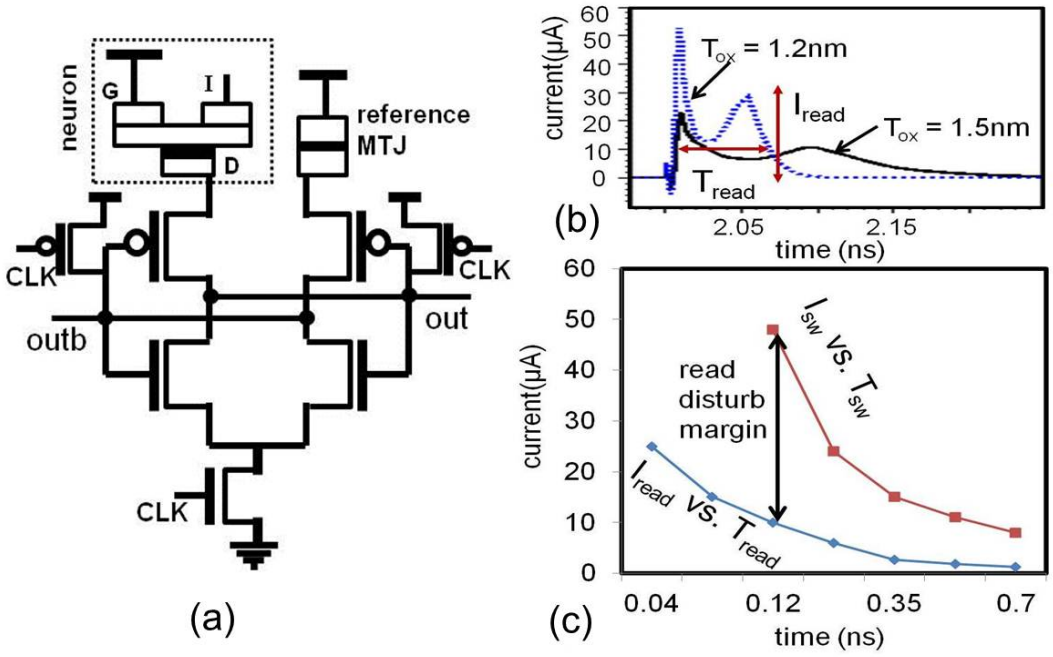


Figure 3.15(a) Dynamic CMOS latch for sensing the neuron MTJ, (b) For thicker tunnel oxide (T_{ox}), the peak transient read current (I_{read}) reduces and read time (T_{read}) increases. (c) Plot comparing DWM switching threshold (I_{sw}) for different switching time (T_{sw}), with that of $I_{read\ Vs.}\ T_{read,}$ show that sufficiently large read disturb margin is available for a wide range of I_{read}.

highly scaled free-layer (domain-2) size used in this work, a switching current threshold of the order of ~1μA may be achievable. Fig. 14 shows the micro-magnetic simulation plot obtained from OOMMF for such a scaled device. The effect of thermal noise has been incorporated by employing stochastic LLG [9].

Note that a DWN-based neuron employs a homogenous magnetic write-path and hence does not suffer from non-idealities like interface and channel-spin scattering that limit spin-injection efficiencies in devices based on lateral spin valves [29]. We employ dynamic CMOS latch for reading the MTJ (fig. 15), which results in only a small transient current drawn from the output port (domain-3) of the DWM neuron, which can be kept below its switching threshold. Additionally, the time domain threshold for

domain wall motion also helps in preventing read disturb from the small transient current [10].

Application of similar DWM-based device structure for digital logic design has been proposed earlier in [16]. However in such a scheme non-volatility of the DWM switch is critical, as the free-layer is required to store the logic information for half a clock cycle. Moreover, the domain-wall between the free-domain and the two fixed domains must be stabilized in absence of input current to preserve the logic state. This is generally achieved by incorporating notches at the boundary of the fixed and the free domain [31]. The use of notch can increase the switching current threshold significantly. In the neuron operation however, the non-volatility of the free-layer is inconsequential, as the CMOS latch is used to transfer the spin-mode information into full-swing voltage levels, while the input current is present. Thus the need of stabilizing the DW in absence of input current can be mitigated. This can facilitate lower switching current as compared to that used in [16].

Further improvement in switching current may be possible by the application spin-orbital assisted domain-wall switching [103], which has been predicted to achieve an order of magnitude reduction in critical switching current for current-induced DW-motion in nano-scale magnetic nano-strip [103]. In this work we employ a current threshold of ~1.5μA for 1.5 ns switching-speed based on the simulation results for the DWM neuron free-domain size of 48x16x1.5nm^3, which corresponds to the current-density of 5MA/cm^2. This dimension of the free-domain would offer an effective resistance of ~200Ω.

3.3.3 Neural Network with Spin Neurons

Fig. 16 shows a spin-neuron with inputs synapses connected to domain-1. Domain-3 is connected to the ground potential. Due to low-resistance of the magneto-metallic write-path of the neuron, in absence of any input signal, the input terminal of the neuron is also clamped to the ground potential. This naturally fulfills the requirement of low impedance input-node along with a fixed input potential for the neuron device. Assuming a neuron with ~25 inputs and a maximum current of ~3μA per input, $|\Sigma(G_i V_i)_p| + |\Sigma(G_i V_i)_n|$ and $\Sigma(G_i V_i)$ come close to ~40μA and ~3μA. This implies an overall current-flow of ~3μA in and out of the DWM neuron (with resistance ~200Ω), which would result in a fluctuation of ~0.6mV at the input node. Thus even for input voltages as small as 30mV, the percentage fluctuation in the input-node-voltage can be less than ~2%. Moreover, it should be noted that, this fluctuation (positive or negative) is caused by the input itself. The net input-current injected into the neuron changes the input voltage in the direction determined by the larger of positive and negative current components (ie., according to the direction of the current flow at the neuron input). Hence it may not affect the final outcome unless it is large enough to reduce the current (difference between the positive ($\Sigma(G_i V_i)_p$) and the negative ($\Sigma(G_i V_i)_n$) components) injected into the neuron below its switching threshold (in this case designed to be ~2μA).

The state of the neuron's free layer (domain-2) can be detected using a high-resistance voltage-divider formed between a reference MTJ and the neuron MTJ, with the help of the CMOS latch in fig. 15. Thus the spin-neuron simultaneously provides

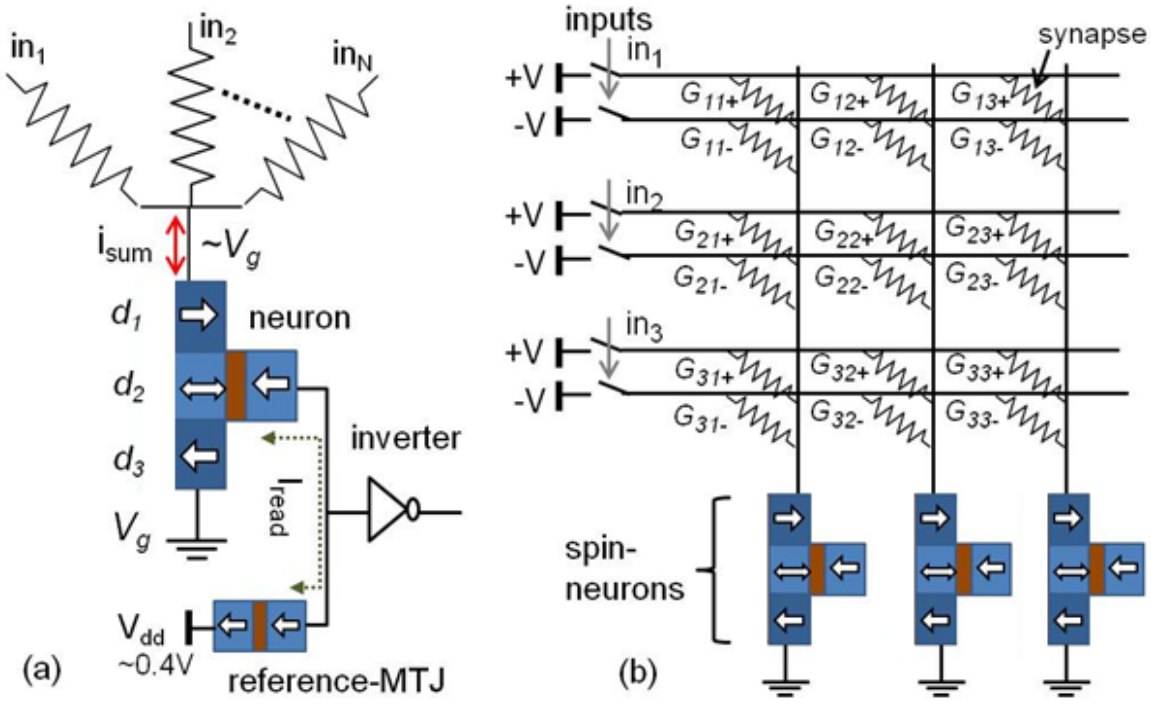


Figure 3.16(a) Spin-neuron connected with input synapses, (b) neural-network circuit using spin-neurons

transimpedance conversion for the input-current, thereby realizing the complete neuron equation in a single device.

The energy dissipation for the spin neuron has two components. First, the switching energy due to the static current flow between the input voltages and the neuron. These components equal to the product of the total input-current flowing across the synapses, the input-voltage levels and the neuron switching time. For an average of ~40μA of current flow across input voltage levels of ±30mV for 1ns switching time, this component evaluates to ~1.2fJ. 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 $\pm$ 100mV of input levels, which might be more easily achievable, the first energy component is limited to ~4fJ, which is more than two orders of magnitude less than that obtained for the CMOS neuron. The second component of energy-dissipation in the spin-neuron can be ascribed to the MTJ-based read operation which was found to be ~0.5fJ of energy dissipation comes from the latch's operation. Thus the total energy-dissipation in a spin-neuron for ~1ns switching speed can be close to 1fJ. This leads to the possibility of three to four order of magnitude improvement in energy-delay product as compared to a conventional CMOS implementation. Apart from ultra-high energy efficiency, another attractive feature of the spin-neurons is their compactness. In the CMOS layer a compact CMOS inverter replaces an area consuming OPAMP. Hence, spin neurons can facilitate higher integration density for neural-network circuits.

A 3x3 neural-network circuit using spin neurons is shown in fig. 16b. The network has two conductances (that can be implemented using multi-level spintronic memristors)G_{i+} and G_{i-} for each input in_i. When an input is high (logic '1'), a voltage signal $+\Delta V$ and $-\Delta V$ are applied to the conductances G_{i+} and G_{i-} respectively, resulting in proportional current flow into the input terminal of the neuron, as shown in fig. 4b. The net current due to the i_{th} input in_i, injected into the j_{th} neuron, therefore, can be written as $\Delta V(G_{ii+}-G_{ii-})$. Thus, the input weights needed for the neurons can be obtained by programming G_{i+} and G_{i-} to appropriate states.

The write path of the neuron is connected to ground. Using Kirchhoff's law it can be visualized that the net current flowing into the input node of the neurons is given by the following equation:

$$I_{sum} = \Sigma\Delta V((in_i(G_{ij+} - G_{ij-})) \quad (3.11)$$

This expression is essentially same as the term within the braces in eq.1. The sign function over the current-mode summation is carried out by the spin-neurons, thus realizing the energy-efficient neural-network functionality. At the level of network-design, another noticeable advantage of spin-neurons is ultra-low energy-dissipation in cross-bar interconnects in the synapse network shown in fig. 16b. This results from the ultra low-voltage operation of entire network, facilitated by the spin neurons. Notably, the LSV bases spin neurons with ~1.5μA switching current threshold and ~2ns switching time would also result in similar energy figures.

3.4 Summary

In conclusion, memristor based resistive cross-bar memory can be a very attractive option for non-boolean/neuromorphic as well programmable Boolean computing hardware. However, application of conventional analog circuits with resistive-memory may lead to energy inefficient and complex design. However, magneto-metallic spin neurons can be ideally suited for memristor based computing, as they act as fast, compact, low voltage, and ultra low energy current-mode thresholding elements. Spin neurons can also be combined with other deivces operating as synpases, to realize neural network hardware. For instance, non-programmable synapses can be realized with CMOS transistors of different dimentions. Similarly multi-level MTJs can also be used as compact programmble weights in a corss bar network. A fully spin-mode neuron-synapse unit can also be formed by using DWM as prograamble input synapse and LSV as the neuron unit.

In the subsequent chapters we present different design examples for neural hardware with spin neurons with different circuit-integration schemes.

4. HYBRID NEURAL NETOWRK DESIGN USING SPIN NEURON AND MAGNETIC DOMAIN WALL SYNAPSE

In this chapter we present the spin based neuron-synapse model. First we discuss the application of domain wall magnet as a synapse. Following this, the neuron model is described which is based on the lateral spin valve structure discussed in chapter 3. Application of the proposed device in spin-CMOS hybrid neural network design is described and the overall performance is discussed.

4.1 Domain Wall Magnet as Synapse

As mentioned earlier, domain wall can be moved along a magnetic nano-strip by application of magnetic field or by injection of charge current along the nano-strip. Application of DWM in the design of non-volatile memory and logic design has been explored by several authors. In the present work, we propose the use of DWM as synapse, where its programmable spin injection strength is used for implementing spin-mode weighting operation. Fig. 1a shows a domain wall magnet interfaced with the non-magnetic channel of a neuron.

In order to write the weight into the DWM, current is injected along the length of the domain wall as shown in fig. 1a. Under this condition the channel is kept

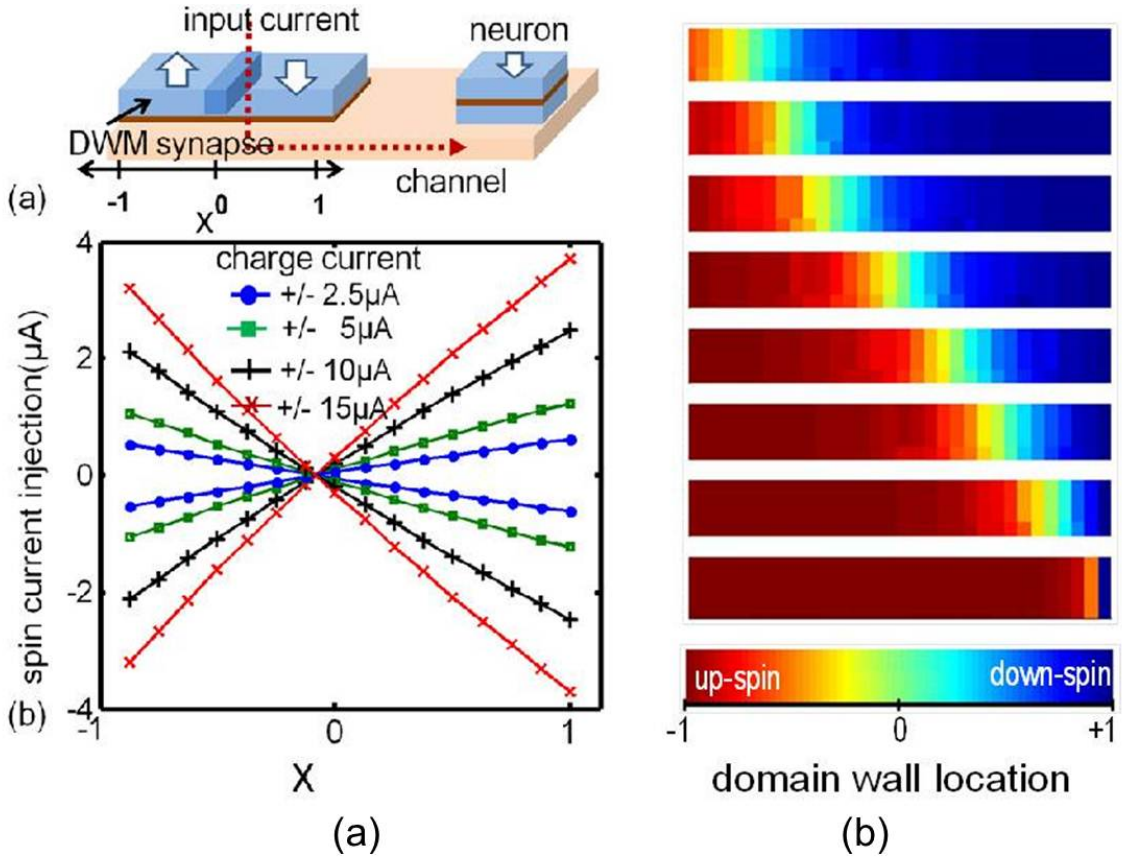


Figure 4.1(a) Spin polarization strength current injected through DWM as a function of DW location, (b) Fig. 2 Magnetization state of the DWM at equal time intervals after starting of DWM motion.

in a floating state. A thin MgO layer incorporated at the top and bottom surface of the DWM reduces the fringe current passing through the parallel path provided by the floating channel and the input lead, during the write operation. The interface oxide also imparts an effective resistance to the input lead of the DWM that makes it dominate the parasitic resistance of the signal-routing metal- lines. During computation, the input current is injected into the channel through the domain wall in the vertical direction. Fig. 1b shows the plot for spin polarization of current passing into the channel through the DWM vs. domain wall location for different charge current values. It can be observed

that, spin polarization strength of the charge current reaching the channel is proportional to the offset of the domain wall location from the centre. For the extreme left location of the domain wall, the charge current reaching the metal channel is maximally up-spin polarized and vice-versa. The net polarization is reduced to zero for the central location of the domain wall, as equal amount of up and down spin electrons are injected into the channel in this case.

In the simplest case, the two extreme locations of the domain wall can be employed for implementing programmable binary weights. Neural networks with binary weights can be applied for logic synthesis and pattern recognition applications However, network with binary weight may require larger number of neurons for a given operation, as compared to a network with higher number of weight levels depending upon the size of the exhaustive training set. Larger number of weight levels can be obtained by employing longer DWM stripes that can facilitate better quantization of domain wall location. It has been shown that incorporation of nano-scale notches in the DWM strips can enhance the stability of DW at the notch sites [31]. The incorporation of notches along the length of the DWM synapse can help in achieving higher writing accuracy. In this work we incorporate DWM synapses with a cross section area of 350x80nm^2. Notches etched at 22nm interval along the 350nm long DWM strip can provide 16 levels of weight. Fig. 1b shows the magnetization state of the DWM at equal time intervals after the application of 250psec voltage pulse train.

4.2 Spin based Neuron model

Transfer function of an ‘integrate’ and ‘fire’ neuron is given by eq. 1.

$$Y = f\left(\sum w_i I_i + b\right) \tag{4.1}$$

Here, w_i and I_i are the weights and corresponding inputs and b is the neuron bias. The bias can be chosen to be zero. It however aids in training convergence and can be easily implemented by an additional synapse magnet which is driven by a clock. The function f*(x)* is given by eq.2 and approximates a step transfer function for a sufficiently large N.

$$f(x) = \left(1 + e^{-N(x-t_o)}\right)^{-1} \tag{4.2}$$

Here t denotes the threshold of the neuron. It can be inferred that a higher $|t|$ would require a larger value of $|x|$ to switch the neuron. For a given set of normalized weights $W_{i,}$ this translates to larger levels of the input signals $I_{i.}$ For the spin based neuron model, this implies larger input current per synapse and hence higher power consumption. Therefore, switching threshold of the output *nano-magnet* needs to be reduced. We incorporate current-mode Bennett-clocking to achieve this. The device structure for the neuron with three inputs is shown in fig. 2. The 'firing-magnet' forms the free layer of an MTJ. The two anti-parallel, stable polarization states of a magnet lie along its easy axis (fig. 2). The direction orthogonal to the easy axis is an unstable polarization state for the magnet and is referred as its hard axis [10, 13]. The preset-magnet shown in fig. 2 has its easy axis orthogonal to that of the neuron magnet (MTJ free-layer which is in contact with the channel). In the beginning of a clock-period, current-pulse injected through the preset-magnet forces the neuron-magnet to the hard-axis configuration (fig. 3). As soon as the hard-axis biasing-pulse goes low, the free-layer makes transition to the easy-axis polarity governed by the polarity of net spin-polarization of the channel-current flowing

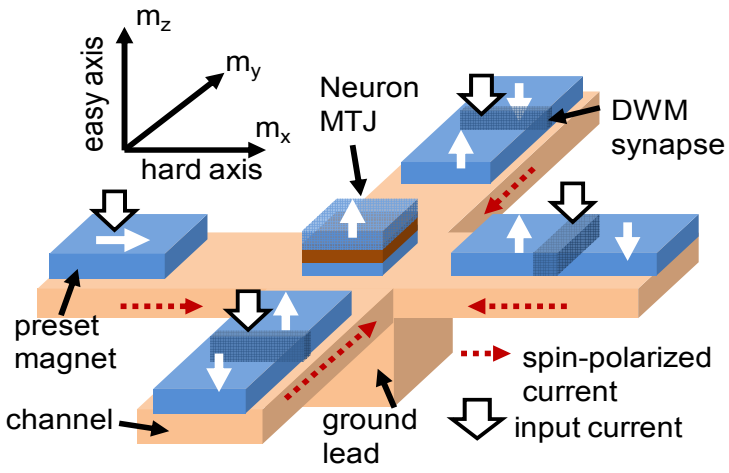


Figure 4.2 LSV-based neuron-model (with non-local spin injection) with three inputs (DWM synapses). The free layer of the neuron-MTJ is in contact with the channel and its polarity, after preset, is determined by spin polarity of combined input-current in the channel region just below it.

under it. As a result, the firing-magnet, i.e., the free layer of the MTJ acquires either parallel or anti-parallel polarization with respect to the fixed-layer. Note that, summation of the 'spin-weighted' input currents (eq. 1),received through multiple DWM synapses, takes place in the metal-channel. Whereas, the symmetric step-transfer function upon the summed spin-current (eq. 2), is realized with the help of Bennett-clocking of the neuron-magnet [9]. Note that, in this work non-local spin injection has been used for the LSV neuron. However, local spin injection can also be used, as discussed in chapter-3, for higher spin-injection efficiency.

When the clock is low, a CMOS-based detection unit (discussed later) reads the state of the neuron MTJ. For a parallel configuration, it generates a high output whereas for the anti-parallel configuration, it settles to a low value. Hence, the detection unit converts the spin-mode information of the neuron magnet's state into a charge-mode

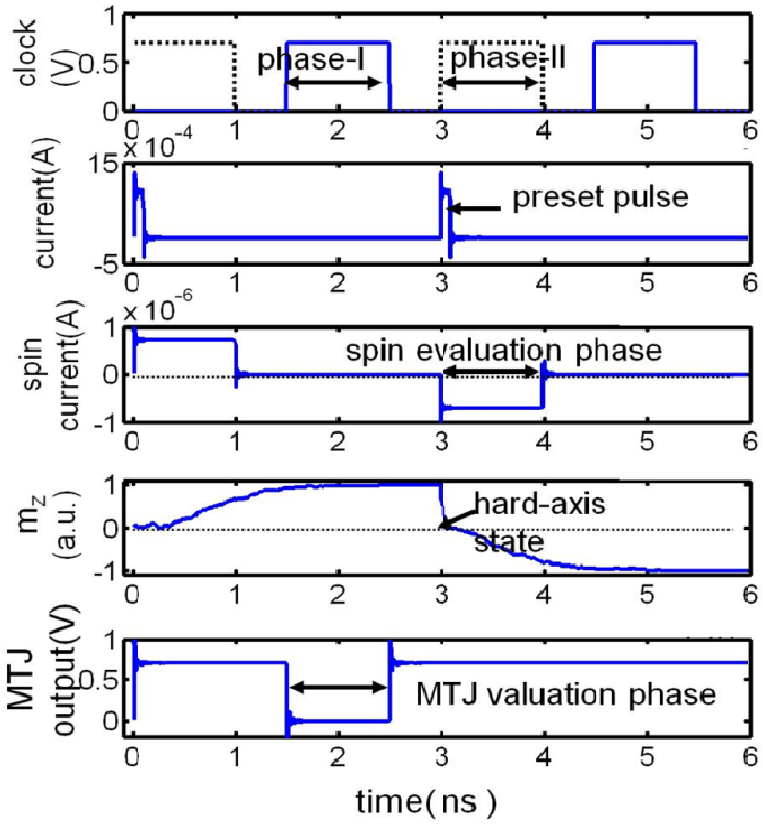


Figure 4.3 3 Timing waveform for the proposed neuron model

signal. For a particular stage of network, spin and charge mode evaluations occur in alternate clock phases (fig. 3). For a multistage, feed-forward neural network, neurons in alternate stages are driven by complementary clock phases. This results in a fully parallel and pipelined network.

In the proposed neuron model, the use of non-local STT switching allows a low resistance path for static charge current flow that includes the DWM synapse and the non-magnetic channel. This allows application of very small voltages, which in turn results in ultra low energy operation for the magneto-metallic neuron-synapse unit.

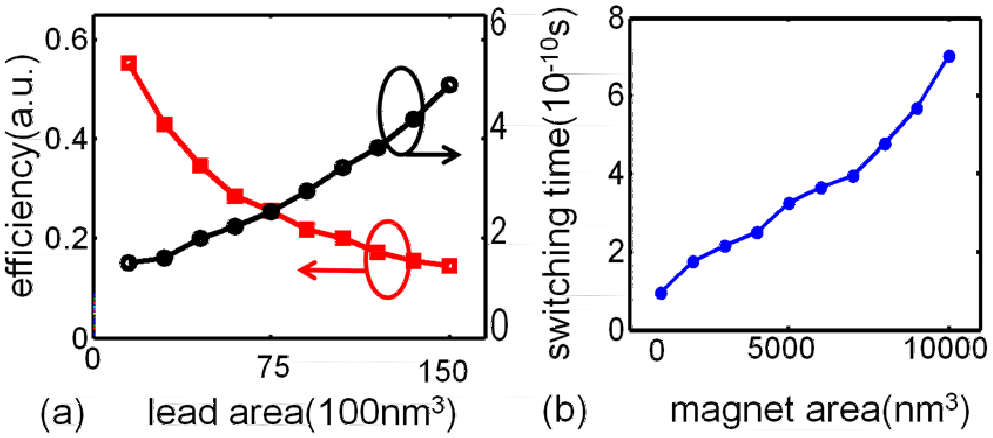


Figure 4.4(a) Increase in spin injection efficiency and switching speed through scaling of ground lead for a fixed current input (b) Reduction in switching time with combined scaling of neuron magnet for a fixed current input.

The detection scheme, employs dynamic CMOS latch discussed later, involves negligibly small transient current flow through the high resistance MTJ stack.

Performance metrics of the neuron-device, like, spin injection efficiency (for the non-local case), switching energy and switching-speed can be improved by the appropriate choice of magnet parameters, device geometry and operating conditions. Non-local spin injection efficiency in the device can be defined as the ratio of spin current $I_{s,}$ injected into the output magnet and the net spin polarized charge current in the channel under the neuron MTJ. As discussed in chapter 2, for non-local spin injection, the spin components of the combined synapse current gets divided between the output magnet and the ground lead (fig. 2). Thus the spin injection efficiency for a given charge current input is enhanced by increasing the resistance of the ground lead (fig. 4a).

Smaller volume for the output magnet, along higher coercive field H_k leads to higher switching speed for a given spin current (fig. 4b) [9]. It also leads to faster easy

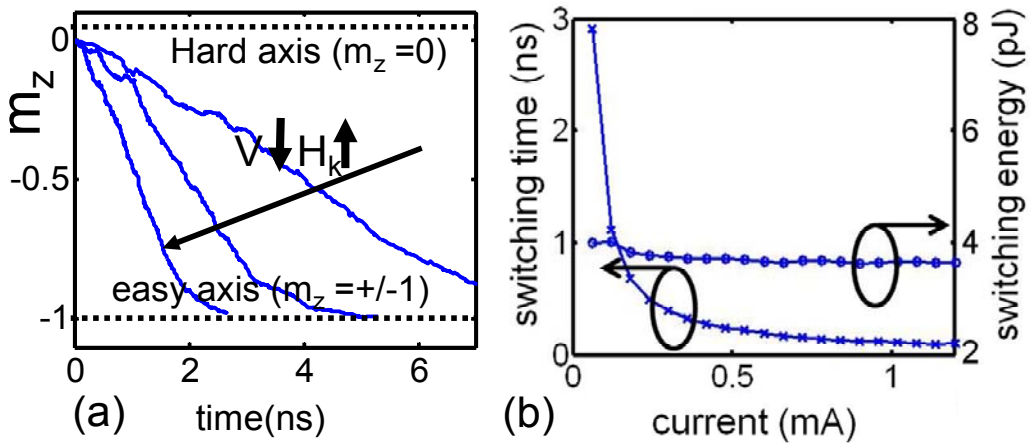


Figure 4.5(a) Increase in easy-axis restoration speed with H_k and reducing magnet volume (for spin current of 0.5 μA) (b) Hard-axis switching time and switching energy vs. switching current.

axis restoration (fig. 5a). In order to maintain the spin injection efficiency, resistance of the ground lead needs to be scaled up proportionately.

Hard-axis switching-energy is a significant portion of the energy dissipation per-neuron, per-cycle. Fig. 5b shows that, the hard-axis switching current increases with switching speed (~direct proportionality [11]). Hence for a given terminal-voltage, the switching-energy remains almost constant. In the present work, the hard axis biasing current is supplied through a transistor operating between a small terminal voltage. In order to allow a small transistor width and hence, lower clocking power, it is favorable to choose the smallest possible value for switching current and hence maximum possible preset pulse width for a given operating frequency. In this work we employed preset-current pulse of amplitude 300μA and pulse width 0.5ns.

A center-surround layout for a neuron with 12 input synapses is shown in fig. 6. Spin-polarized charge current inputs from DWM synapses combine in the channel and

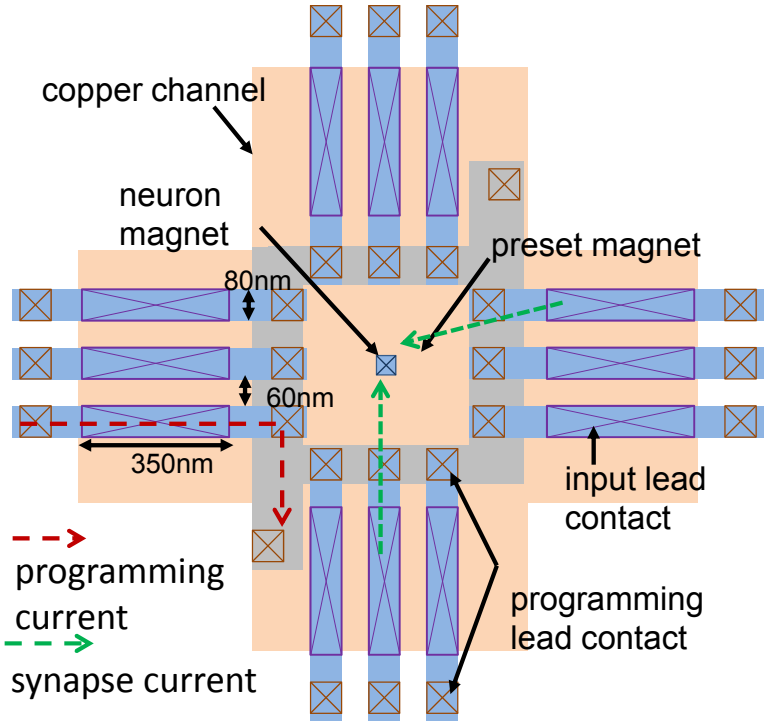


Figure 4.6 Centre-surround layout of the proposed neuron-synapse unit. Spin-weighted current inputs from DWM synapses combine in the central region of the 2-D metal channel, where the neuron is located

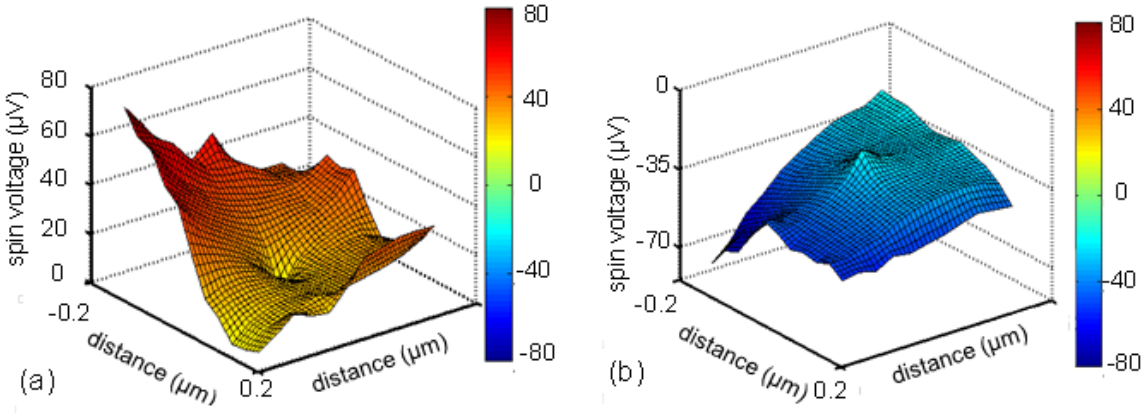


Figure 4.7(a) Channel spin potential of a 16 input neuron under firing condition (b) Channel spin potential under non-firing conidtion

flow into the ground lead located near the neuron MTJ. Spin polarization strength of charge current decays exponentially with the distance travelled along the non-magnetic

channel. Thus, the channel-length between the synapses and the neuron must be within 1-2 times spin flip length (λ) [10]. This imposes a limit on the number of input synapses for the structure shown in fig. 6. For copper channel (λ~ 1μm) up to ~32 synapses can be combined directly. For graphene channel (λ~ 6μm) this number can be increased.

Figure. 7 depicts the plot for spin potential in the central-region of the channel, surrounding the output magnet of a 16 input neuron, under firing and non-firing conditions. It shows that, in case of a firing event, the entire channel is dominantly at a positive spin-potential and vice-versa.

4.3 Spin-CMOS hybrid Neural Network

Due to small spin diffusion length of metal channels, spin-mode signaling cannot be used for network connectivity. Hence, in this work the spin-based neuron-synapse modules are interconnected through charge-mode signaling using CMOS. The spin-mode 'firing' information is converted into charge-mode signal using the dynamic CMOS latch, shown in fig. 8a. It compares the effective resistance of the MTJ units in its two load branches. The firing MTJ of the neuron unit connects to one of the loads, whereas, a reference MTJ is connected to the other.

The latch drives a distributed set of current source transistors which in turn supply charge current to all receiving neurons through the respective input magnets (DWM) (fig. 8b). The source terminal of the current source transistors and the ground lead of the spin based neuron modules are biased at $V+\Delta V$ and V volts respectively. Hence, the synapse current flows across a small terminal voltage of ΔV. In the present work, values of V and ΔV are chosen to be 800mV and 30mV respectively. The CMOS units operate between

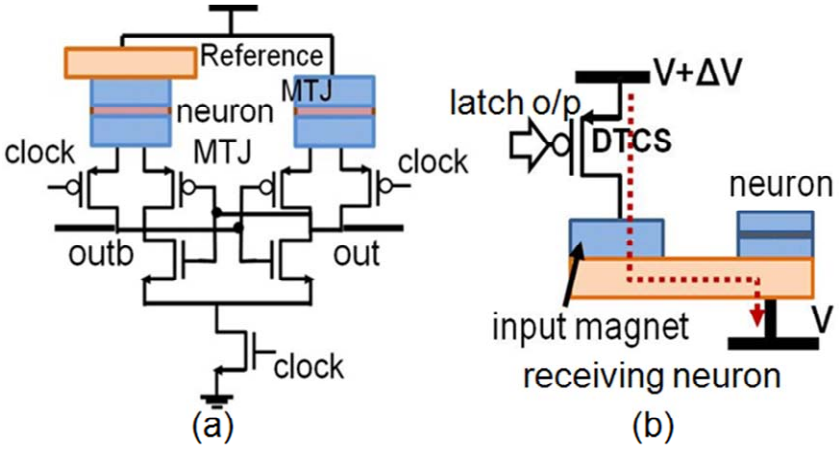


Figure 4.8(a) Differential MTJ latch (b) Inter-neuron current-mode signaling using deep triode current source (DTCS) transistor.

800mV and 0V. Biasing of the spin modules between two relatively high DC levels proves advantageous as compared to direct application of a small supply voltage of magnitude ΔV. This is because, application of differential DC supply can mitigate the impact of I-R voltage drop along the supply lines. It can also be exploited to reject the common-mode noise in the dual supply lines. Moreover, generation of clean DC levels below 100mV is challenging in the state of art CMOS technology, whereas a regulated voltage supply of higher magnitude can be distributed with less than 0.1% fluctuation .

For supplying a current of 5µA per synapse (across a drain to source voltage of 30mV) for 16 receiving neurons, the required source transistor width in 45nm technology is around 2.5µm. In order to minimize the impact of synapse current mismatch, distributed source transistors are used.

Fig. 9 depicts the correspondence between the proposed spin-CMOS hybrid ANN and the biological neural network. The spin potential of the 2-D metal-channel (which is analogous to neuron cell body) depicted in fig. 9, can be related to the

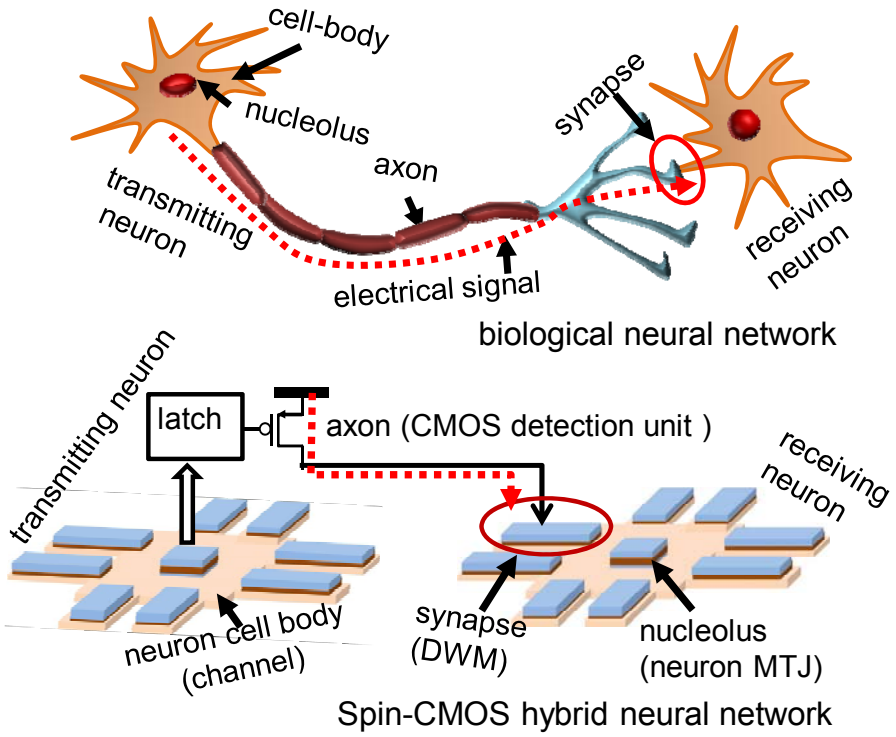


Figure 4.9Correspondence of the spin-CMOS Hybrid ANN to biological neural network

electrochemical potential in biological-neuron's cell-body. Inter-neuron communication in the present design is realized using ultra-low voltage current transmission, which is somewhat similar to the natural mechanism. However, the aim of the proposed model is not to mimic the biological neural network in terms of functionality, but to evolve a model for artificial-neural-network suitable for computational hardware.

4.4 Network Simulation

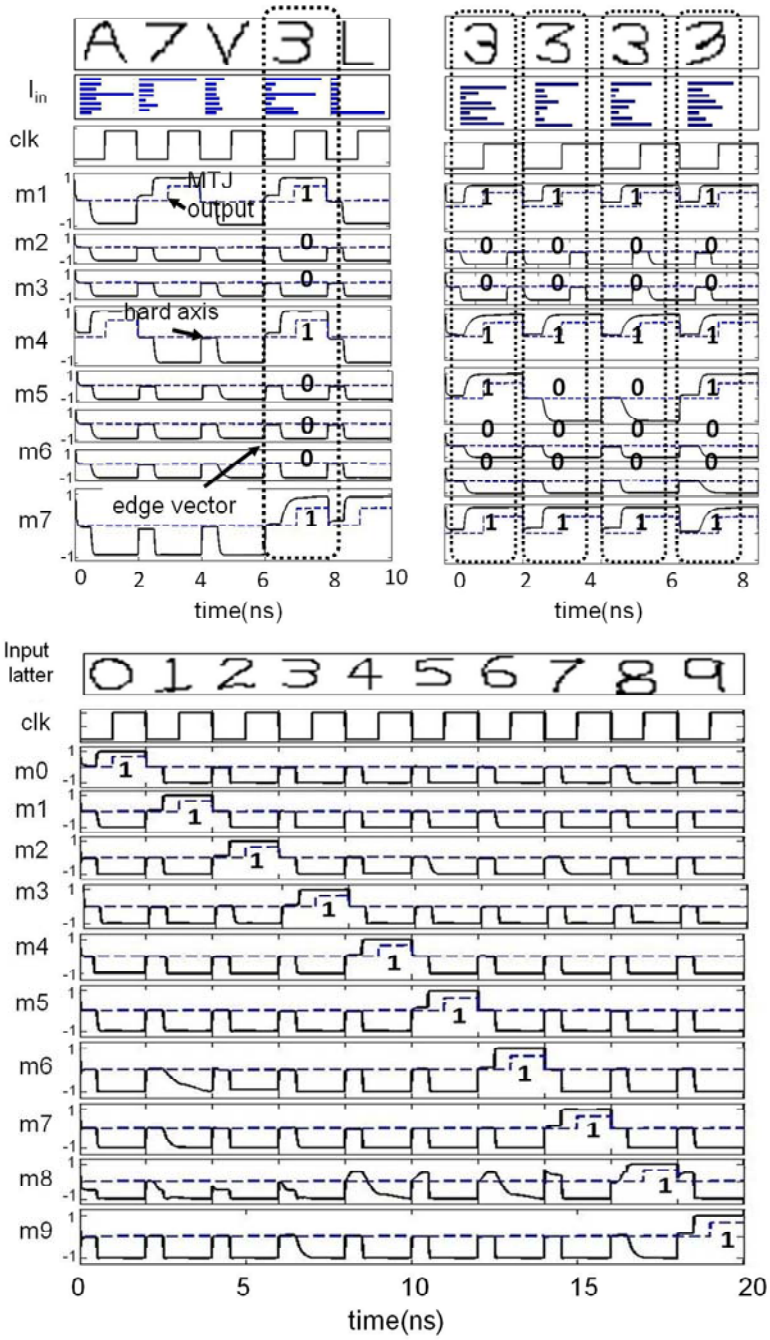


Figure 4.10(a) Barcode generation for horizontal edges in alphanumeric characters, (b) Effect of style variation on horizontal bar code, (c) Output waveforms for numeric character recognition character is depicted in fig. 12a. The solid lines denote the magnetization state of the neuron magnets whereas the dashed lines indicate the corresponding MTJ evaluation.

In this section we describe the network simulation for character recognition as a benchmark application. Impact of process variation upon network performance is assessed. We also compare the performance of the proposed spin-CMOS hybrid ANN with that of a state of art CMOS ANN design.

We simulated character recognition as a benchmark application for the proposed spin-CMOS hybrid design. The overall process for character recognition can be divided into two steps, namely, edge extraction and pattern matching. For edge extraction, column wise pixels form the binary image along four directions - horizontal, vertical and + 45^o are fed to the first stage neurons. These neurons generate a high output if the number of non-zero pixels along a particular column (or equivalently the spin current input $I_{in}$to the neuron) is higher than the neuron threshold. Note that, a desirable threshold for a neuron is set by applying a bias input to it. The horizontal edge extraction process for different input is shown in fig. 10a. Fig. 10b shows the effect of variation in the handwriting style for the numeral '3' on the horizontal bar code. It shows that, significant variations in writing style translate to slight variations in the barcode pattern which can be tolerated by an ANN. Variation tolerance can be enhanced by training with different styles of input characters. The resultant four binary patterns form a 1-D representation of the input character. This pattern is fed to the output stage of the network for classification. The output neurons correspond to the 36 alpha numeric characters. The output evaluation for numeric characters is shown in fig. 10c.

4.5 Variation Analysis

As described earlier, variation aware circuit design techniques, like, the use of distributed and matched current source transistors, can reduce the effect of CMOS process variation upon network performance significantly. The impact of nano-magnet parameter variation upon system performance however, needs to be assessed while modeling an ANN with nano-scale devices.

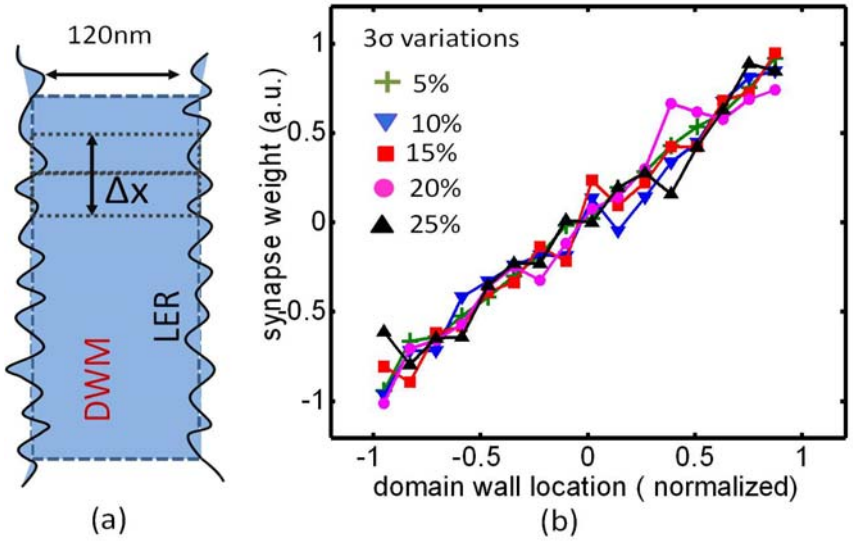


Figure 4.11(a) DWM cross section area showing LER(b) Combined effect of LER, and programming inaccuracy upon DWM weight.

The critical DWM parameters, having impact on computation accuracy, can be identified as, interface oxide thickness, cross section area and domain wall locations. Variation in oxide thickness can lead to mismatch in the effective resistance of the DWM input leads. This leads to difference in charge current injection for different synapses, which in turn introduces errors in weights. However, since the interface oxides are generally grown through atomic layer deposition (ALD), their thickness can be precisely

controlled. Cross section area variation in the DWM synapse leads to variation in spin polarization of the input charge current. Inaccuracy in domain wall programming directly translates to imprecision in synapse weights.

The effect of writing inaccuracy in the domain wall synapse is captured in the

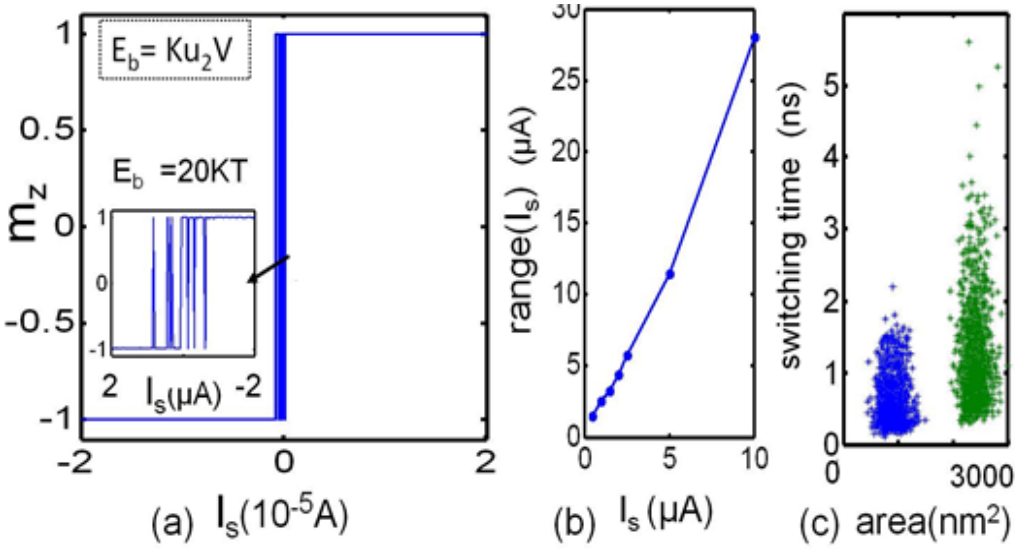


Figure 4.12(a) Near threshold noise reduction for higher anisotropy barrier, (b) Range of spin current injection into neuron magnet vs. average synapse current for a neuron with 16 synapse (c) Scatter plot for easy axis relaxation time under parameter variation and varying input current.

simulation framework by imposing random shifts in domain wall location (fig. 11a). Impact of process variation like line-edge roughness (LER) is incorporated in terms of random variations in the DWM cross section area (fig. 11a). Fig. 11b shows the superimposed effects of inaccurate writing and geometrical imperfection upon DWM weight. The neuron magnet is highly scaled in order to achieve fast easy axis restoration and lower switching current. It is therefore expected to be prone to thermal noise and magnet parameter variations. Fig. 12a depicts the effect of thermal noise on neuron transfer characteristics. Under very small input spin current, the easy axis restoration can

be non-deterministic due to thermal noise. Impact of the noisy transition zone on overall network performance can be ignored as long as it correspond to a small fraction (1-5%) of the range of spin current injection I_s. The range of I_s in turn depends linearly on average synapse current (fig. 12b). Hence, noise

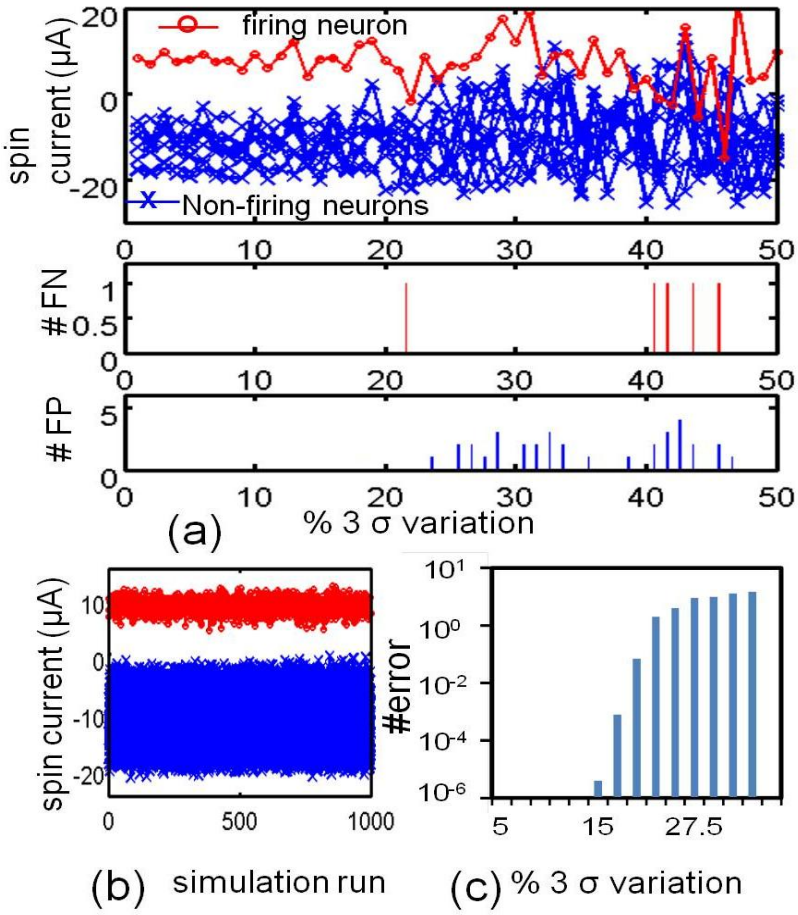


Figure 4.13(a) Impact of process variation on spin current input to neuron magnets, (b) 1000 point simulation for 15% 3σ variations, (c) Monte Carlo results for a neuron under combined process variations.

determines the limit to which the average synapse current can be lowered to reduce the overall power consumption.

Since, Bennett clocking places the neuron switching threshold at origin, irrespective of the magnet parameters, the impact of output-magnet parameter variations upon the device transfer characteristics is significantly mitigated. Parameter variation however, affects the dynamic switching characteristic of the neuron. Easy axis relaxation time for neuron magnet spreads with increased parameter variations, which limits the maximum operating frequency for reliable operation. Fig. 12c shows the scatter plot for neuron switching time for two different sizes of the output magnet. The input current has been varied over two orders of magnitude (20μA to 0.1μA) corresponding to the variation in synapse currents for different input combinations. 25% 3σ variation has been applied for critical magnet parameters. It is evident that lower volume and higher H_k (for a constant switching energy barrier) results in lower spread and hence, facilitates higher operating frequency.

Fig. 13a shows the effect of increasing process variation upon the spin current delivered to the output neurons corresponding to the numeric characters. A negative value of spin current for firing neuron and a positive value of spin current for a non-firing neuron denotes an error. The resulting false negatives (FN) and false positives (FP) are shown in the figure.

Network simulations show that, among different device parameters considered, domain wall location has the maximum impact upon network performance. This is because it bears a direct relation to the synapse weight. As mentioned earlier, incorporation of nano-scale notches along the DWM length can achieve improved programming accuracy. Fig. 21b shows the plot for 1000 simulation points for the network under combined 15% 3σ variations for DWM and neuron magnets. Monte Carlo

TABLE 1
Design Performance for Character Recognition

N_n =86 F_s =500MHz	CMOS 45nm (digital)	CMOS 45nm(analog) (200MHz)	Spin-CMOS Hybrid ANN
Power	65mW	22mW	0.75mW
Network Area	2.1mm^2	0.010mm^2	0.018mm^2

TABLE 2.a
Spin ANN Specs

Spin based ANN	
Number of Neurons*	
Input layer	24
Hidden Layer	24
O/P layer	36
Supply Voltage	
Vdd_H (mV)	825
Vdd_L (mV)	800
Current per firing operation per synapse	20μA
# Weight Levels	16

TABLE 2.b
CMOS ANN Specs

Digital ANN	
#Full Adders	4400
Programmable Latches	6240
weight bit	3 + 1 (sign)
Analog ANN	
Synapse resistance	200K Ω 10 KΩ
Max. synapse Current	4μA
Avg.Neuron power	0.25mW

TABLE 4
Device Parameters

Ku2(biaxial anisotropy)		2x10^6 erg/cm^3	DWM polarization constant	0.9
Magnet size (nm^3)	Neuron	40x30x1	Damping coefficient	0.007
	DWM	350x80x10	Channel material	Cu
H_k(coercivity)		5KOe	Spin Flip Length	1μm (300K)
M_s(saturation magnetization)		400emu/cm^3	resistivity	7Ω-nm

Figure 4.14 Tables depicting the performance of the proposed spin-CMOS hybrid neural network, as compared to CMOS

simulation results for a neuron given in fig. 21 c depicts that it retains accuracy up to more than 18% 3σ variations. Note that 18% variation in a 16 level synapse weight implies a programming error of 3 levels.

4.6 Design Performance

In order to establish a comparison with state of art CMOS technology we implemented the same network architecture in CMOS 45nm technology in two different ways, digital and analog. For the digital design, programmable latches were used to store synapse weights and full adders were employed to implement neuron . For the analog design, memristive synapses were employed. Resistance values in the range of 10kΩ to 200kΩ were used to emulate memristors. In this design analog current-summers were employed for modeling the neuron. The area was estimated based on the cross bar architecture for memristive neural network.

Table-I in fig. 14 compares the two designs with the proposed spin based neural network. The digital implementation consumes large area as well as power due to bulky neuron and synapse units. For a digital neuron with 16 4-bit inputs (as used in this work), the critical path for neuron unit consists of ~16 full adders. With 0.8V of supply voltage a latency of ~0.4ns was obtained for the digital neuron, for which power dissipation of each neuron was found to be ~0.7mW, leading to ~65mW total power for 500MHz throughput. Note that, a fully parallel implementation for the digital ANN was chosen for the purpose of comparison. Area for the digital design can be reduced through sequential processing using smaller number of neuron units, but power consumption is expected to remain almost constant for a given throughput. For the analog design analog CMOS OPAMP were employed as neuron circuits (as described in chapter 3). For a ~200MHz

bandwidth, the power consumption for the analog neuron circuit was ~0.2mW, leading to ~22mW of total power for the entire network. Thus the analog design achieved smaller area and power as compared to the fully parallel digital design, however, at the cost of smaller throughput (200MHz as compared to ~500MHz for digital design).

Power consumption for the spin neuron was found to be dominated by the static component, resulting from current-mode computing in the neurons. For 16-input neurons, ~20μA current per input (with effectively normalized weights) led to around 25μA effective spin-current (positive or negative depending upon input combination), in the neuron channel. Assuming ~20% non-local spin-injection efficiency from the channel to the output magnet, this would provide ~5μA of spin current to the output magnet. This was found to suffice for deterministic current-mode Bennett-clocking of an output magnet of size ~30x40x1nm^2. With ~25mV of DC supply bias for the neuron circuits, this led to ~7μW static power per neuron. After including the overhead due to CMOS latch and clocking, the overall power for the entire network was estimated to be only ~0.75mW for 500MHz throughput. The spin-CMOS hybrid implementation thus achieves both, low power as well as small area, comparable to that of the analog ANN. The power and area benefits of the proposed design can be ascribed to simple and compact spin devices that operate at ultra-low supply-voltages and mimic the neuron operation. Both, low energy consumption, as well as compactness is conducive to integration of large number of neurons for programmable computational-networks for cognitive and Boolean computation. Note that, the use of neuron-model with local spin-injection would further improve the power savings achieved by the proposed spin based neural network.

Table-2 provides some relevant design details. Finally table 3 enlists some of the critical device parameters used in the simulation.

4.7 Summary

In conclusion, spin device phenomena like, majority evaluation, hard-axis switching, and adjustable spin polarization strength of domain wall magnets, clubbed with appropriate clocking scheme can lead to an energy efficient model for neuron-synapse unit. The localized, ultra-low voltage operation of neuron-synapse units, assisted with efficient circuit and architecture level design strategies for inter-neuron signaling and power gating can facilitate high degree of integration. The proposed spin-CMOS hybrid ANN design can be suitable for low power, programmable computation architecture for cognitive as well as Boolean applications. Note that, in this chapter we have employed LSV neuron with non-local spin injection. Local spin injection can also be used for higher spin-injection efficiency and hence power savings. The application presented in this chapter used a simple binary input in the form of alpha-numeric characters, however, analog spin based neural network can also compute with analog inputs. The application of spin-CMOS hybrid neural network for analog image processing application is presented in the next chapter.

5. SPIN-CMOS HYBRID CELLULAR NEURAL NETWORK FOR ANALOG IMAGE PROCESSING

In chapter we present the application of 'spin-neuron', proposed earlier, in an 'on-sensor' image processing architecture. We show that, the spin neurons can be integrated with CMOS transistors to arrive at spin-CMOS hybrid processors (PE). In such a PE, the analog-mode computation can be carried out with the help of the neurons, at ultra-low energy cost. Apart from ultra-low voltage operation, the fast switching of the neuron-magnets also help in reducing the computation energy

5.1 Cellular Neural Network (CNN) : Mathematical model

Cellular neural network (CNN) can be regarded as a fusion of artificial neural network (ANN) and cellular automata [48-49]. It borrows the basic information processing functionality, i.e., the 'integrate and fire' operation upon weighted inputs, from neural networks. The concept of computation based on neighborhood influence, on the other hand, is akin to cellular automata. This class of computation has been found to be highly suitable for several image processing applications, which essentially involves processing of pixel neighborhoods in a parallel fashion.

Fig. 1 shows a cellular neural network array with 3x3 rectangular neighborhoods. Each cell is connected to its eight surrounding neighbors through a 3x3

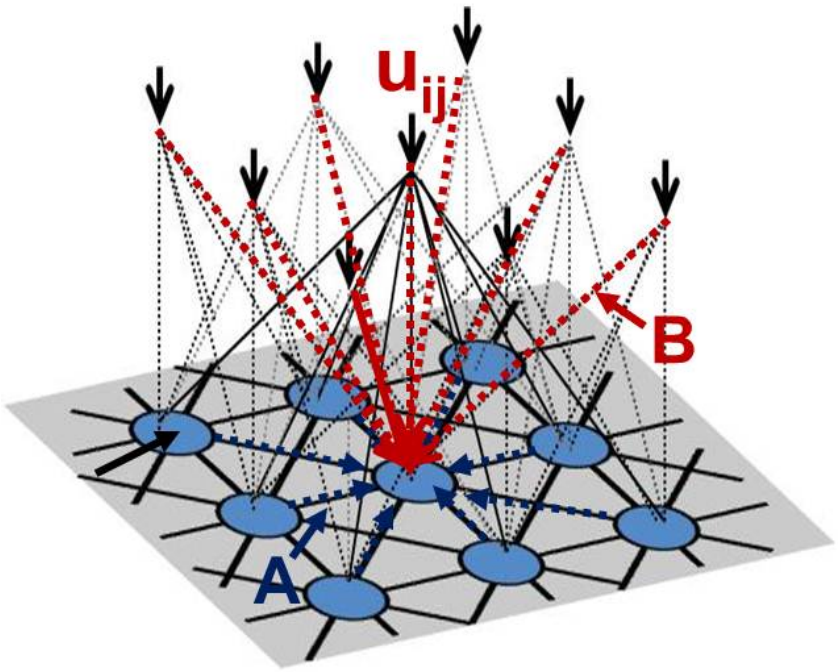


Figure 5.1CNN architecture with 3x3 neighbourhood connectvity

feedback-weight template *A*. *A*(0,0) denotes the self feedback term. The feed-forward template of a cell, *B* (or the input-weight template), determines the connectivity to the neighborhood inputs. In a CNN, each neuron performs integrate and fire operation upon the weighted combination of its neighborhood inputs and outputs in a recursive manner.

The standard expression for a CNN cell state is given by eq. 1 [48].

$$C\frac{dx_{ij}(t)}{dt} = -x_{ij}(t) + \sum_{(k,l)\in N(i,j)} A(i,j;k,l).y_{kl}(t) + \sum_{(k,l)\in N(i,j)} B(i,j;k,l).u_{kl}(t) + z(i,j) \tag{5.1}$$

Where, *x(t)* is the cell state at time *t*, *A* and *B* are the feedback and feedforward template defined above, *u(t)* is the input to cell from its 3x3 neighborhood *N* and *z* is the cell-bias.

The cell output is denoted by *y(t)* which is related to the cell state *x(t)* with a non-linear transferfunction. Time domain dicretization of the CNN state equation leads to eq. 2.

$$x_{ij}(k) = \sum_{(k,l)\in N(i,j)} A(i,j;k,l).y_{kl}(k) + \sum_{(k,l)\in N(i,j)} B(i,j;k,l).u_{kl}(k) + z(i,j) \tag{5.2}$$

Discrete time CNN (DTCNN) employs a step transfer function given by eq. 3.

$$y_{ij}(k) = f'(x_{ij}(k-1)) = \begin{cases} 1 & if \quad x_{ij}(k-1)>0 \\ 0 & if \quad x_{ij}(k-1)<0 \end{cases} \tag{5.3}$$

Application of a step transfer function limits the value of a cell output *y(i,j)* to binary levels of *f'(x)*. The input *u(i,j),* however, can assume continuous values corresponding to the range of pixel intensity.

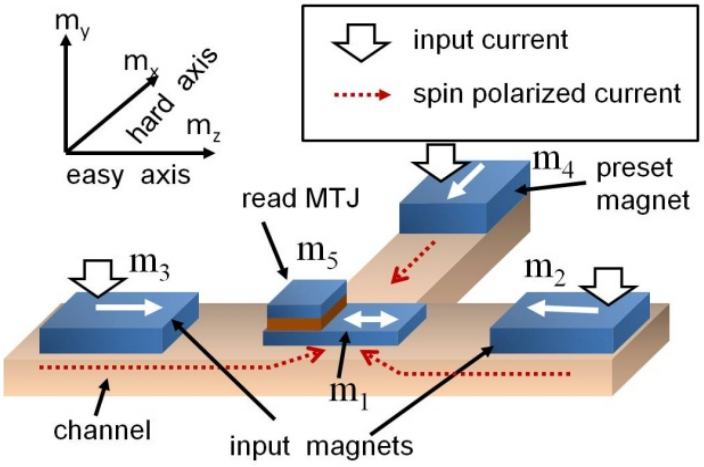


Figure 5.2Bipolar Spin neuron based on LSV (with local spin-injection)

In the spin-CMOS hybrid PE proposed in this work, the two input magnets (m_2 and m_3) of the neuron device shown in fig. 2 are used to realize the inter-neuron connectivity through *A* and *B* templates respectively. All the neighbouring outputs

y(i,j)(/inputs *u(i,j)*) linked to a neuron with positive *A(i,j)'s* (/*B(i,j)'s*) connect to one of the inputs, say m_2, whereas, those, associated with negative terms in the template matrices, connect to the other input m_3. The circuit techniques employed to realize a DTCNN processor (PE) with the spintronic neuron is described in the next.

5.2 DTCNN architecture with Spin Neurons

In this section we describe the design of spin-CMOS hybrid PE that implements the DTCNN functionality for on-sensor image processing. The inputs signal *u(i,j)* for a cell, is the associated photo-sensor input. Transistors of weighted dimensions are used as deep-triode region current sources (DTCS), to implement *A* and *B* templates. The neuron in a PE, receives sensor input signals and outputs of its neighbouring PE's through the DTCS's in the form of charge current. The current mode signals combine in the metal channel of the neuron, where the Bennett clocking of the output magnet realized, eq. 3. A dynamic-CMOS detection unit however, converts the bipolar spin information pertaining to the state of the neuron-magnet, into unipolar voltage-level. Hence, the final PE output is given by eq. 3. The circuit operation corresponding to these step are described in the following paragraphs.

Fig. 3 shows a photodiode that converts the illumination intensity received at a pixel into a voltage signal. The transistor M_1first presets the photodiode capacitance to Vdd-V_t, where Vdd is the supply voltage and $V_t$is the threshold voltage of the transistor. The capacitance is then discharged by the photodiode current, rate of discharge being proportional to the incident illumination intensity [54]. At the end of discharge period of a fixed duration, the transistor M_2 samples the photodiode voltage. The sampled voltage

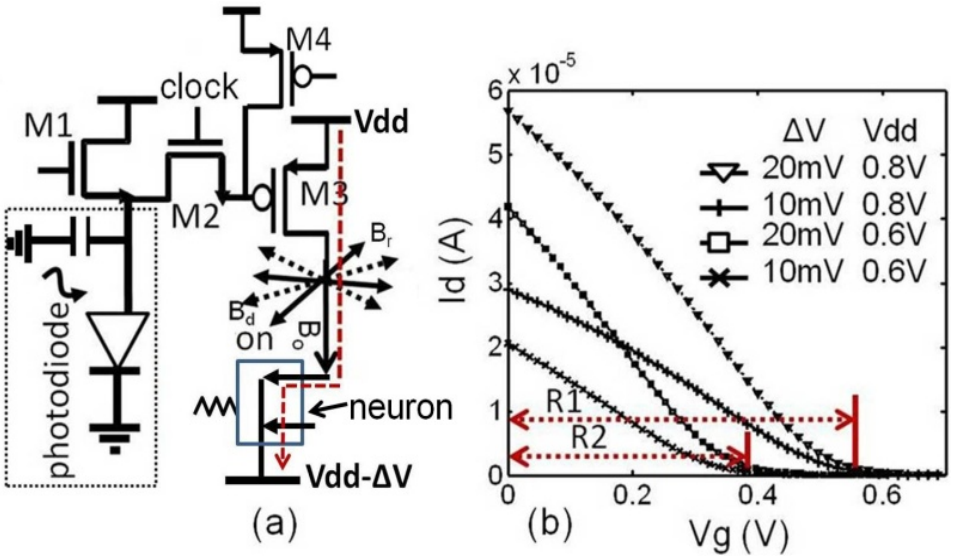


Figure 5.3(a) Circuit for *B*-template realization (b) deep-triode region characteristics of the DTCS transistor M_3driven by the sampled photo-sensor voltage

at the gate of M_3 ranges from $Vdd\text{-}V_t$ to 0V, corresponding to the illumination intensity at the pixel. M_3supplies input current to the neurons located in the 3x3 neighborhood of the pixel through separate and weighted fingers, with dimensions corresponding to the elements of the *B* template. A second DC level $Vdd\text{-}\Delta V$ is used in the design, in order to exploit the low-voltage operation of the spintronic neurons. It connects to the lead terminal of the neurons as shown in fig. 3a. The current supplied by M_3 therefore, flows through a small terminal voltage ΔV, which can be of the order of ~10mV. Note that, since the resistance of M_3 is significantly higher than that of the magneto-metallic neurons, it accounts for most of the ΔV-voltage drop. Fig. 3b shows that the output current of *M3* is a fairly linear function of the sampled gate voltage for the deep-triode region operation.

Fig. 4 shows the circuit scheme used to realize the *A*-template. The corresponding simulation waveforms are shown in fig. 5. When the clock is low, output

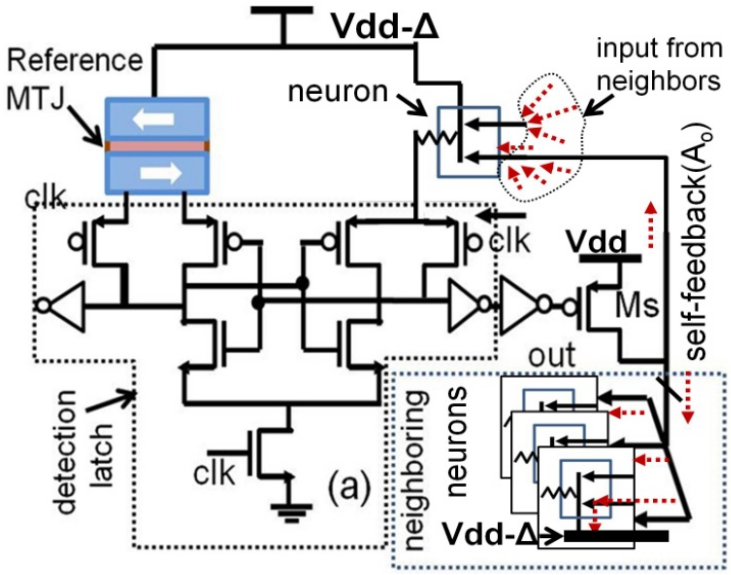


Figure 5.4CMOS detection unit senes the state of the neuron magnet and transmits current mode signal to the neighboring neurons through a deep triode current source transistor.

of the dynamic-CMOS latch is precharged to *Vdd*. The latch is activated at the positive edge of the clock signal. The two load branches of the latch are connected to he detection terminal, *D,* of the neuron and a reference MTJ respectively. The latch compares the difference between the effective resistances in its two load branches through a transient discharge current. It drives negligible static current into the high resistance neuron-MTJ stack. For the anti-parallel state of the neuron-MTJ (which can be regarded as the 'firing state'), the latch drives the DTCS transistor M_s shown in the figure. $M_{s,}$ in turn, supplies current to the neighbouring neurons through separate weighted fingers corresponding to the *A* template. After a time delay that is sufficient for the latch to evaluate and settle to its final value, the neuron device receives the preset current through a clock driven

DTSCS. Note that, a delayed preset pulse with respect to the clock edge ensures that the latch evaluates correctly according to the neuron-MTJ state stored in the previous

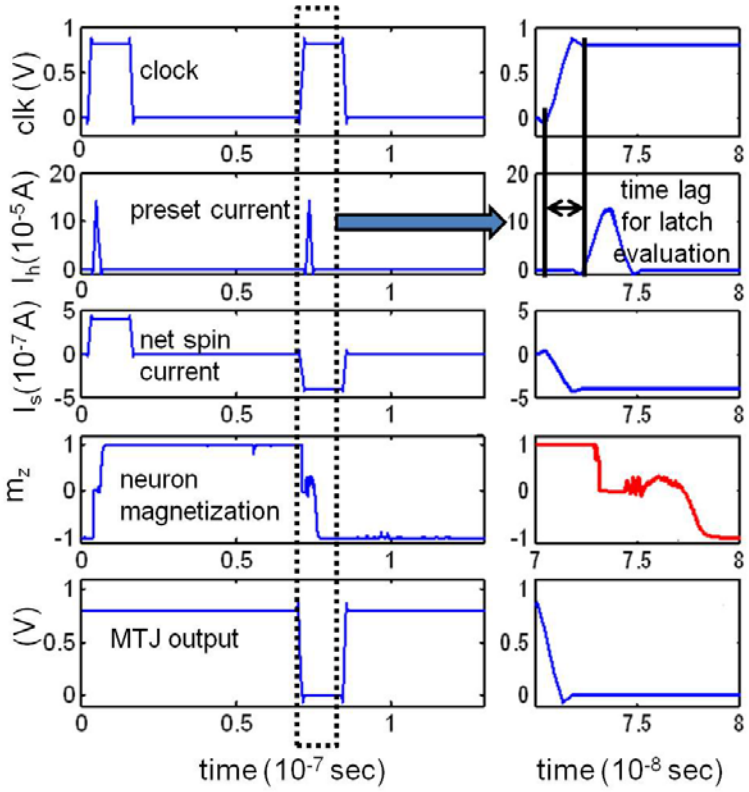


Figure 5.5Simulation waveform for DTCNN operation of the spin-CMOS hybrid PE

evaluation cycle. Once evaluated, the latch can not change its state until it is precharged again, despite the flipping of the neuron MTJ. At the positive edge of the clock, the latches in all the PE's evaluate simeltaneously and conditionally drive their respective DTCS outputs. Hence, a neuron recieves input currents from its neighbors, during the period when the clock is high.As soon as the preset signal goes low, the neuron magnet settles to one of its stable states, depending upon the overal spin current received through its inputs. Thus, the recursive operation of DTCNN PE, given by eq. 2 is realized by the application of an appropriate clocking scheme. Note that, the current supplied by the

DTCS outputs of the latches also flow across the two supply levels, *Vdd* and *Vdd-ΔV*, as shown in fig. 4.

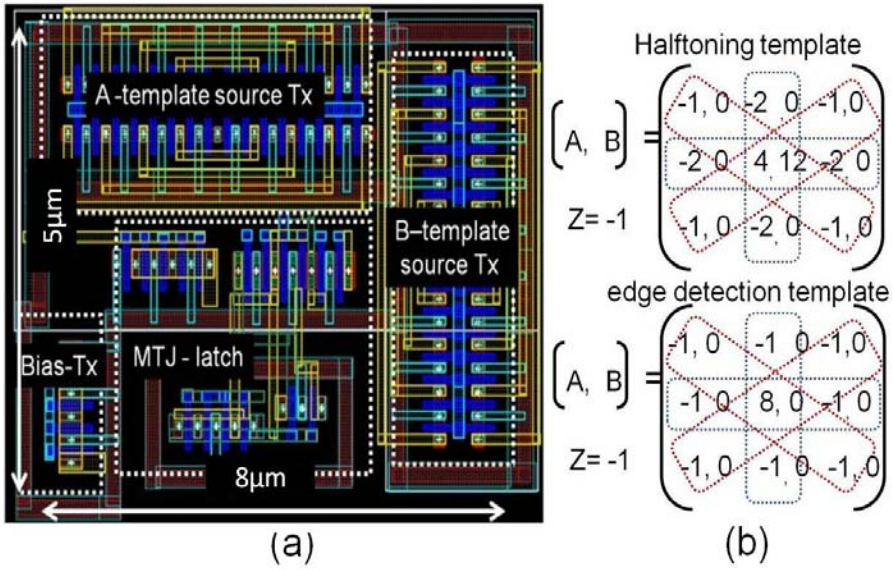


Figure 5.6(a) Layout of the CMOS circuit (90 nm tehnology) in the PE showing that the source transistors occupy larger portion of the PE area. (b) DTCNN templates for edge detection and halftoning

Fig. 6 shows the layout for the CMOS circuitry employed in the spin-CMOS hybrid PE. It shows that a major portion of the PE area is occupied by the triode-region source-transistors (M_3 in fig.3a and M_s in fig. 4). As mentioned earlier, in order to realize non-overlapping inter-neuron connectivity, we employed separate fingers in the source transistors. Moreover, a matched layout of the fingers was considered. Fig. 5 shows the values of *A* and *B* templates for two common applications, halftoning and edge detection. As mentioned before, for an application specific design, the fingers of DTCS's are weighted according to the templates. In the simplest case, for a given connectivity, number of fingers equal to the weight (matrix element) magnitude can be chosen. The

sign of the weight, determines the connectivity, to one of the complementary input of the corresponding neuron.

As discussed before, application of current mode Bennett-clocking reduces the required amount of current injection for a neuron, per- input, to few microamperes. Hence, the multi-finger DTCS transistors can supply the required current even at a small terminal voltage ΔV. Hence, two DC supply levels separated by a difference of ~20mV can be chosen. This achieves reduced static power consumption for current-mode inter-neuron signalling.

As long as input currents of the neurons are large enough to overcome the impact of thermal noise in the neuron-magnet, the precision of computation achievable, with the proposed scheme, is limited, mainly, by the supply noise. As the accuracy of on-chip DC supply regulation, in the state of art technology is limited to ~0.1%, high precicion imaging applications may seem out of scope of the proposed design. However, the use of dual supply rails proposed in this work may significantly compensate this disadvantage. Differential supply lines can significantly mitigate the impact of the noise sources, that lead to common-mode fluctuations. Hence a thorough modelling and analysis of this effect needs to be considered, in order to assess the noise tolerance of the proposed scheme. In the present work, we have included the effect of supply and process variations, and we discuss these in the next section on simulation framework.

5.3 Application Simulation

In the following sub-sections we present simulation results for some common image processing applications like edge detection, halftoning and digitization.

5.3.1 Feature Extraction

Edge detection (fig. 7a) is one of the most common image processing step, applied in most vision applications. As an example, motion detection (fig. 7b) employs comparison

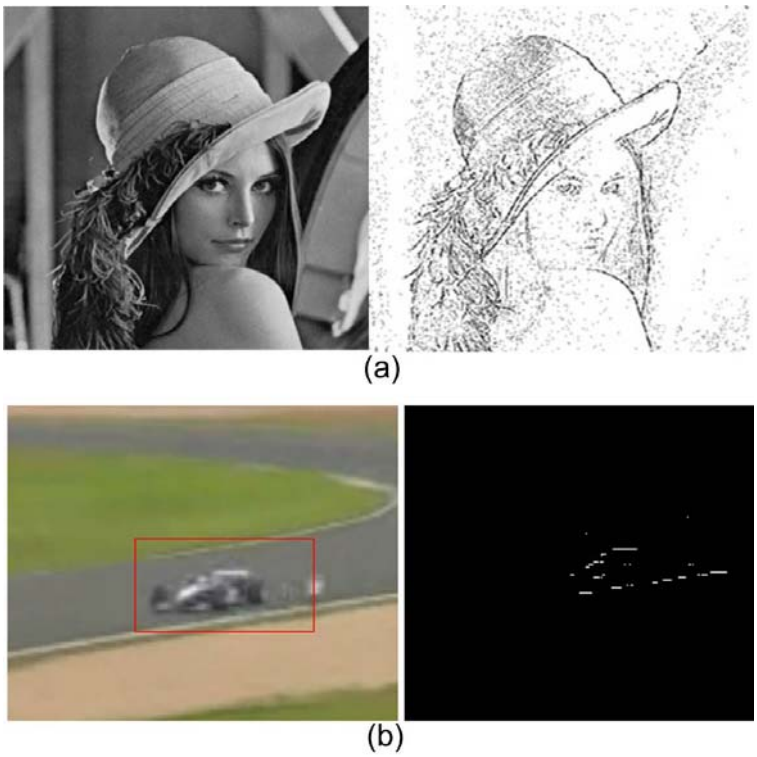


Figure 5.7 (a) Result of edge detection from a grey-scale image, (b) Motion detection on the basis of temporal difference in edge maps.

between the edge maps of a still background, sampled one after the other. This can be achieved by employing extra storage registers per PE to store a sequence of edge maps.

5.3.2 Halftone compression and sensing

Halftoning is a process in which a grey scale image is recorded as (or compressed into) a binary image, with just two levels, in a way such that important details in the image are preserved. Several algorithms for decompressing halftone images have been proposed in literature. This technique can be used for sensing, storing and transmitting images in

bandwidth limited sytems. Simulation result for halftoning of a statellite image is shown in fig. 8. Fig. 9 shows the halftoned image of Lenna along with the effect of reduction in ΔV upon the halftone output. With decreasing ΔV the effect of noise becomes increasingly more prominent.

Figure 5.8Simulation results for halftoned image of a satellite picture

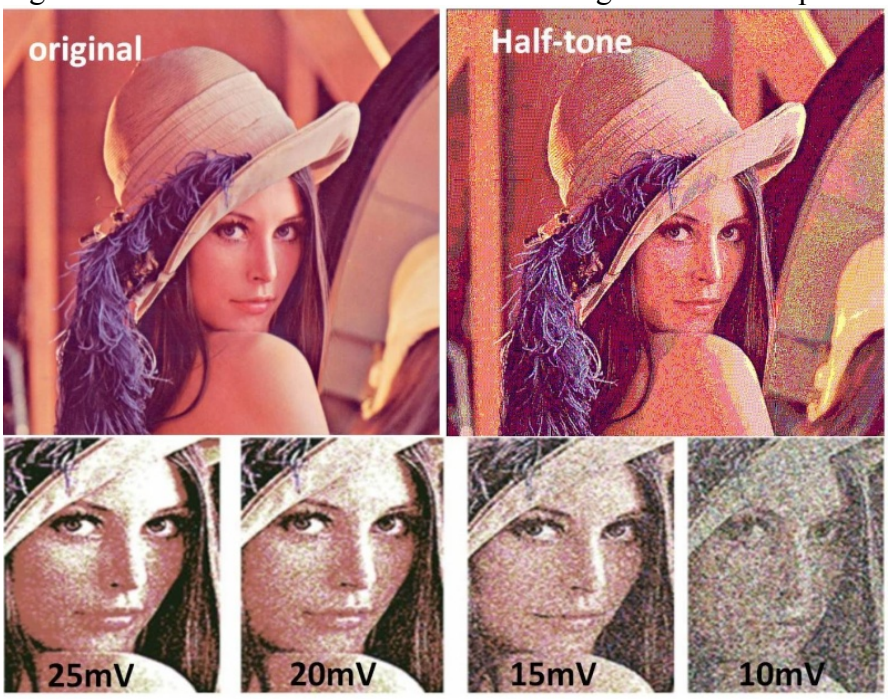


Figure 5.9(a) Halftone of Lenna (b) effect of reduction in ΔV upon the output, with 0.1% supply noise.

5.3.3 Digitization

Successive-approximation-register (SAR) analog-to-digital converter (ADC) is one of the most common data converters employed for on-sensor image quantization (fig 10a) [64]. The data conversion algorithm employed in an SAR-ADC can be explained as follows. To begin the conversion, the approximation register is initialized to the midscale (i.e., all but the most significant bit is set to 0). At every cycle a digital to analog converter (DAC) produces an analog level corresponding to the digital value stored in the register, and, a comparator compares it with the input sample. If the comparator output is high, the current bit (MSB) remains high, else it is turned low and the next bit is turned high. The process is repeated for all the bits. At the end of conversion, the SAR stores the digitized value for the pixel intensity, which can be read out in a column-wise manner from the sensor array. In a cicuit implementation of SAR-ADC, most of the power consumption results form the comparator and the DAC units. The SAR unit consists of a bank of

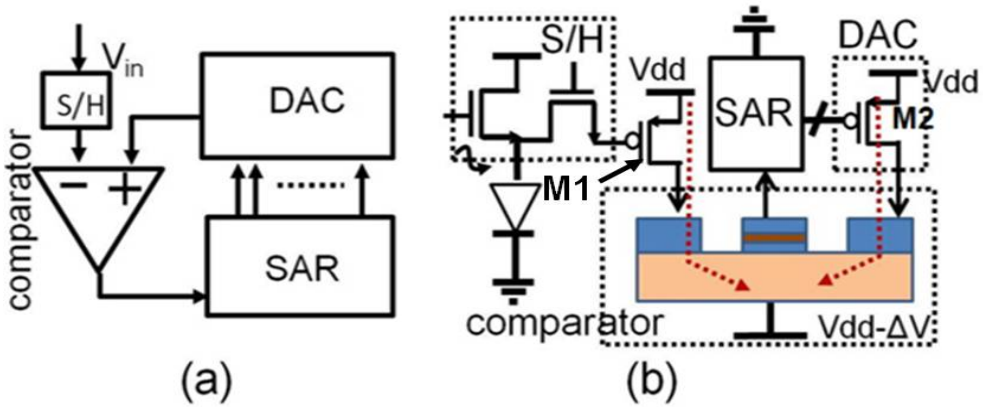


Figure 5.10(a) SAR ADC block diagram (b) compact and low power SAR ADC using spintronic neuron.

CMOS latches and a simple control logic, which consumes negligible power as compared to the analog units.

As the SAR-ADC essentially employs recursive evaluation, akin to the CNN equation, the PE circuit decribed in the previous section can be easily extended to realize

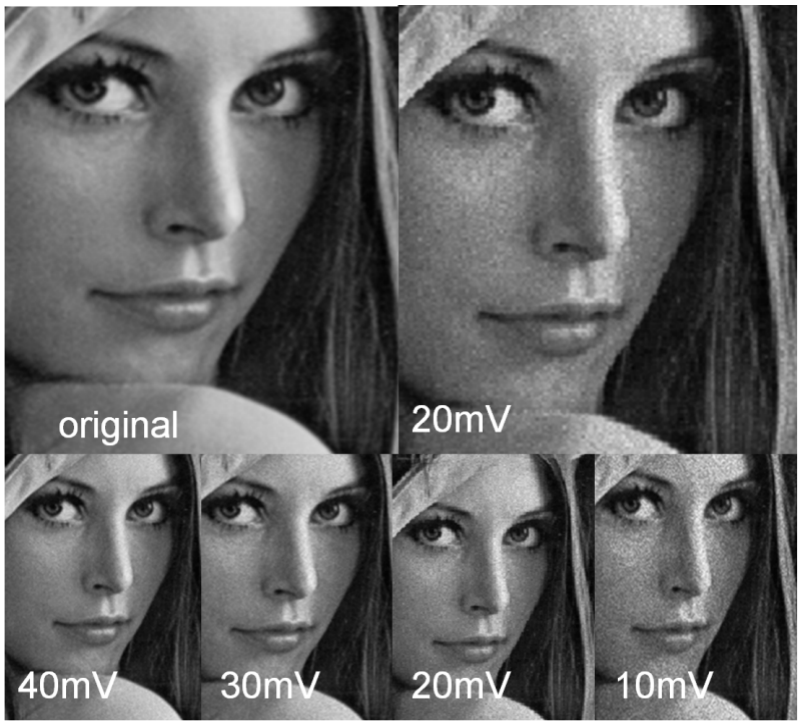


Figure 5.11Simulation result of spin-CMOS hybrid 8 bit-SAR-ADC and the effect of lowering ΔV upon the output, with 0.1% supply noise.

a compact and low power *N*-bit SAR-ADC. In the schematic diagram for the proposed ADC, shown in fig. 10b, the DTCS M_1 converts the sampled output of the photo sensor into a current signal, that is injected into one of the inputs of a three input neuron. The SAR simply consists of a bank of *N* CMOS latches, which in turn drive *N* different fingers of the DTCS *M2*. The multiple fingers of *M2* are binary weighted and hence, it acts as a compact DAC and injects current into the second complementary input of the neuron. Current mode Bennett-clocking of the neuron, using the third input (a preset

magnet, not shown in fig. 15b)), at the beginning of each conversion stage, realizes the comparator operation. Note that, in the proposed ADC design, the analog computation current flows across the two supply levels, i.e., across a small terminal voltage ΔV, thereby resulting in small power consumption. Moreover, in each frame, the current flow is restricted to the small period of conversion just after the data is sampled.

Fig. 11 shows the simulation results for an 8-bit SAR-ADC based on the proposed scheme. Degradation in image quality due to supply noise can be perceived. Note that, in this work we have not considered any coupling between the two supply levels and independent noise sources have been used in simulation. Hence a thorough analysis of the proposed differential supply scheme would be need to assess the computation precision, achievable by the proposed hardware.

5.4 Design Performance

Fig. 12 depicts the architecture for on-sensor image processing [64]. Such a design employs PE's integrated on each of the photo-cell. The output of the photo-detectors are directly processed by the PE's and the result is read out column-wise.

In such an architecture, the total energy dissipation per-input frame can be expressed as the sum of computation energy (E_{comp}), the read-out energy (E_{read}) and the energy that is wasted in the form of leakage current (E_{leak}).

$$E_{tot} = E_{comp} + E_{read} + E_{leakage} \tag{5.4}$$

E_{comp} can be expressed as a sum of neuron-preset-energy, (the energy associated with current mode Bennett-clocking), E_{preset}, the energy associated with current mode inter-neuron signaling, E_{evl}, and the dynamic switching energy in the PEs', $E_{dynamic}$(including the clocking power). A first order expression for these components can be derived using

the design parameters, namely, the two supply levels *Vdd* and *Vdd-ΔV*, the read-out voltage V_{read}, the preset time T_{pre}, the evaluation time I_{evl}, the effective switched capacitance in a PE, C_{PE}, the bit-line capacitance $C_{BL,}$ the word-line capacitance $C_{WL,}$

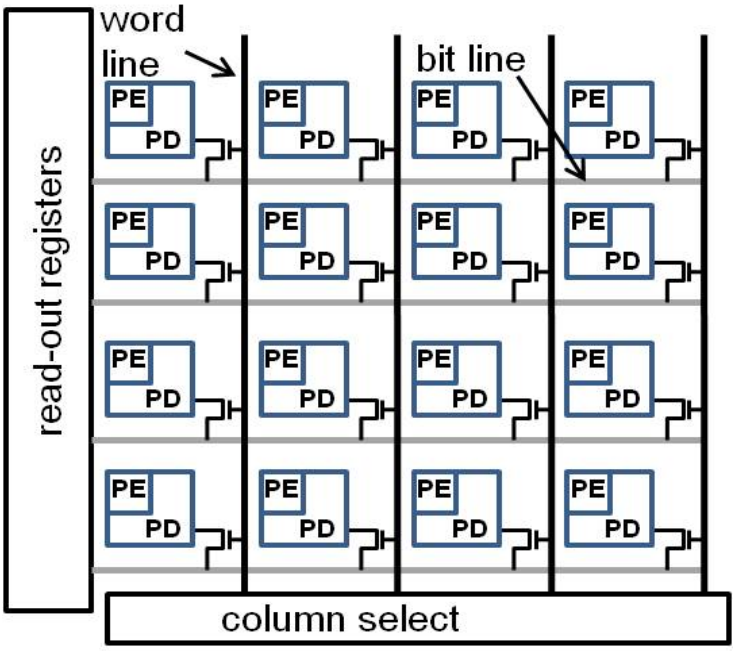


Figure 5.12An on-sensor image processing architecture contains PE's embedded into the pixel locations, and an addressing arrangement for reading out the PE outputs in a column-wise manner

number of cells in the array *N*x*N,* the switching activity factor, α, and the number of iteration required per-frame for a given operation, *M*:

$$\begin{aligned} E_{comp} &= N^2 M(E_{preset} + E_{evaluation} + E_{dynamic}) \\ &= N^2 M(\Delta VT_{pre} I_{pre} + \Delta VT_{evl} I_{evl} + \alpha C_{PE} V_{dd}{}^2) \\ &= N^2 M(\Delta VT_{pre} I_{pre} + \Delta VT_{evl} I_{evl} + \alpha C_{PE} V_{dd}{}^2) \end{aligned} \tag{5.5}$$

The read-out energy, in the case of column-wise read-out can be obtained using the effective bit-line capacitance that is switched to read out *K* bit data per PE from the entire *N* x *N* frame,

$$E_{read} = KN(N(\alpha' C_{BL} V_{dd} V_{read}) + \alpha' C_{WL} V^2) \approx KN^2(C_{BL} V^2) \quad (5.6)$$

E_{leak} can be ignored, as there are well known gating techniques that can make the leakage power for the PE's negligibly small during the read-out period. The results given in table-1, based on the design parameters in table-2 and table-3, in fig. 13, indicate that for the proposed architecture, E_{comp} is of the same order as E_{read}. Hence, the energy component, related to static power consumption due to analog-mode computation, can become comparable to that associated with dynamic power consumption in the peripheral digital-circuits.

As described earlier, the advantage of using the proposed spin-CMOS hybrid scheme for analog computation comes from two main factors. The first, static current flow across a small voltage ΔV, and the second, pulsed operation of the spintronic neurons with a narrow pulse-clock. Although, gating of analog modules in low frame rate image processing architectures have been proposed, gating of analog circuits for high frame rates can be challenging. Moreover, it might not be possible to gate analog cirucits with a pulse-width of a few nano-seconds, which is possble with the spintronic neurons.Comparison with on-sensor image processing designs for feature extraction, given in table-IV, shows more than two orders of magnitude improvement in computation energy. Note that, the effect of technology scaling has been included through a mutiplicative factor of S^2 , where, S is the ratio of the technology scale between the reference design and the presented work (90nm CMOS).Table-4 compares the performance of the proposed SAR-ADC with some recent CMOS designs. Note that ADC is one of the few

Table-I
Design Performance for 256x256 array

Frame rate: 10000 fps	E_{comp}	E_{read}	Power
8- bit quantization	13nJ	8nJ	180μW
Edge detect.	4nJ	1nJ	40μW
Halfton.	6nJ	1nJ	50μW

Table-II
Design Parameters (90nm CMOS)

Vdd	900mV	C_{PE}	6fF
ΔV	20mV	N	256
(I_{evl})	60μA	M, K :	
I_{pre}	120μA	ADC	8 , 8
Tevl	12ns	Edge det.	3 , 1
Tpre	2ns	halfton	4, 1
C_{BL}	200fF	C_{BL}	200fF
V_{read}	100mV	α	0.5

Table-III
Magnet-Parameters

Ku_2 (biaxial anisotropy)		$2x10^6$ erg/cm^3	polarization constant	High: 0.9 Low: 0.1
Magnet Size (nm^3)	neuron	60x20x1	Damping coefficient	0.007
	DWM	350x80x10	Channel material	Cu
H_k (coercively)		5KOe	Channel spin flip length	1μm
Ms(saturation magnetization)		500emu/cm^3	resistivity	7Ω-nm

Table-VI
Comparison with CMOS designs for feature extraction

	CMOS Tech (T)	Fps (frame rate)	N (# PE)	Power	FOM*	FOM(proposed)/ FOM (given)
[45]	0.35μ	2000	32x32	600μW	$3.4x10^3$	253
[4]	0.6μ	100k	1x1	85μW (per PE)	$1.1x10^3$	200
[31]	0.25μ	4000	128x128	20mW	$3.2x10^3$	470
[46]	0.35μ	2000	160x120	25mW	$1.5x10^3$	560
[47]	0.35μ	100	1	0.06μW	$1.66x10^3$	500

Table-V
Comparison of the proposed ADC with state of art CMOS design

Ref 8 bit	CMOS tech.	Fs	Power (W)	Spintronic ADC (W)	FOM** ratio
[35]	0.18μ	370KHz	32 μ	0.04μ	133
[36]	0.18μ	500kh	7.75μ	0.06μ	32
[37]	0.25μ	100KHz	31μ	0.012μ	40
[38]	90nm	10M	70μ	1μ	70
[39]	90nm	20Mhz	290μ	4μ	72

*FOM = (S^2) x(#PE x Fps)/Power

**FOM = (S^2) /Power

S : technology scaling ratio

Figure 5.13 Tables for performance comparison

analog modules for which power consumption reduces with scaling. Results show that the spin-CMOS hybrid ADC can achieves ~30x low power consumption, as compared to some of the latest designs. In this work we have assumed two supply sources *Vdd* and *Vdd- ΔV*. It can be assumed that charge supplied by the higher supply, is restored in the second source, and, can be utilized by other circuit components in a large-scale, heterogenous architecture. Effect of supply noise needs a more thorough analysis. Supply routing techniques, that can exploit the differential supply scheme employed in this work to mitigate the effects of supply noise, need to be explored.

Though, high precision computation on analog images may seem challanging with the technology limits associated with supply noise, the proposed scheme can be highly suitable for several low-level and middle-level analog image processing applications, for which, the conventional mixed signal designs consume large amount of power.

5.5 Summary

In this work we explored the application of the proposed spins, in on-sensor image processing applications. It was shown that a spin-CMOS hybrid PE can handle analog processing functionality in an highly energy-efficient manner. The theoritical analysis presented, showed that, substituting some of the conventional analog processing units in an image acquision and processing hardware, by the spintronic neuron, can achieve ultra low power computation. This can facilitate the design of very high integration density hardware for sensory signal acquisition and processing.

6. ULTRA-LOW ENERGY ASSOCIATIVE COMPUTING ARCHITECTURE WITH SPIN NEURONS

6.1 Introduction

As discussed in chapter 3, resistive crossbar memory (RCM) can be highly suitable for non-Boolean data-processing applications like associative computing. Owing to the direct use of nano-scale memory array for associative computing, it can provide very high degree of parallelism, apart from eliminating the overhead due to memory read. Associative computing of practical complexity with RCM is essentially analog in nature, as it involves evaluating the degree of correlation between inputs and the stored data. As a result, most of the designs for associative hardware using RCM's proposed in recent years, involved analog CMOS circuits for the processing task. Recent experiments on analog-computing with of multi-level Ag-Si memristors employed analog operational amplifiers for current-mode processing. However, as we showed in the last chapter, application of multiple analog blocks for large scale RCM may lead to power hungry designs, due to large static power consumption of such circuits. This can eclipse the potential energy benefits of RCM for non-Boolean computing. Moreover, with technology scaling, the impact of process variations upon analog circuits becomes increasingly more prominent, resulting in lower resolution for signal amplification and processing. Hence, conventional analog circuits may fail to exploit the RCM technology for energy efficient, non-Boolean computing.

The solution to this bottleneck may lie with alternate device technologies that can provide a better fit for the required non-Boolean, analog functionality, as compared to CMOS switches. Recent experiments on spin-torque devices have demonstrated high-speed switching of scaled nano-magnets with relatively small current density [13-14]. Application of emerging spin-torque switching techniques like, those with spin-orbital coupling assist [104], may facilitate the design of low-current nano-scale spintronic switches. Such magneto-metallic devices can operate at ultra-low terminal voltages and can implement current-mode summation and comparison operations, at ultra-low energy cost. Such current-mode spin switches or 'neurons' can be exploited in energy-efficient analog-mode computing. In this chapter we present the design of RCM based analog associative memory using such "spin neurons". The spin neurons form the core of hybrid processing elements (PE) that are employed in RCM based associative modules and achieve more than two orders of magnitude lower computation energy as compared to conventional mixed-signal (MS) CMOS circuits. Application of spin neurons to RCM can therefore greatly enhance its prospect as a non-Boolean computation tool

6.2 Computing with RCM

6.2.1 Multi-level RCM

Although, the basic concept of using spin-neurons for RCM-based computing can be applicable to different resistive memory technologies, in this work, we use the devices for CMOS compatible Ag-Si memristors [104]. Fig. 1 depicts an RCM network. It constitutes of memristors (Ag-Si) with conductivity g_{ij}, interconnecting two sets of metal bars (i_{th} horizontal bar and j_{th} in-plane bar). High precision, multi-level write techniques for isolated memristors have been proposed and demonstrated in literature that can

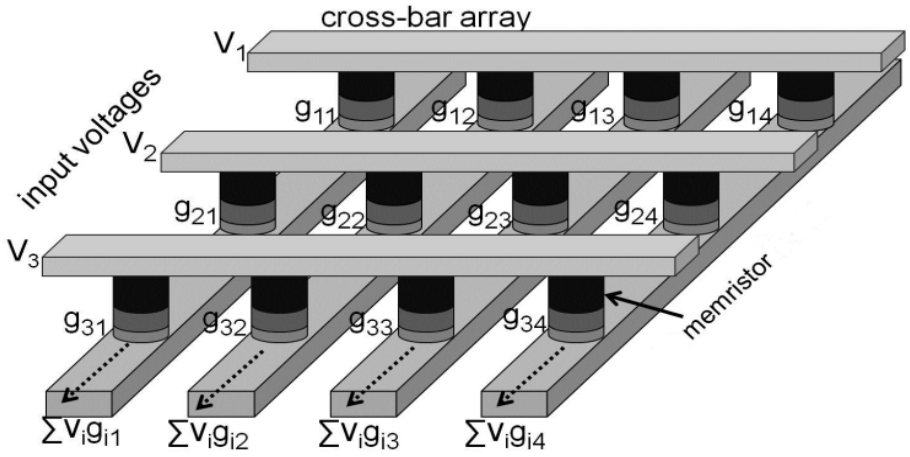


Figure 6.1A Resistive crossbar network used for evaluating correlation between inputs and stored data

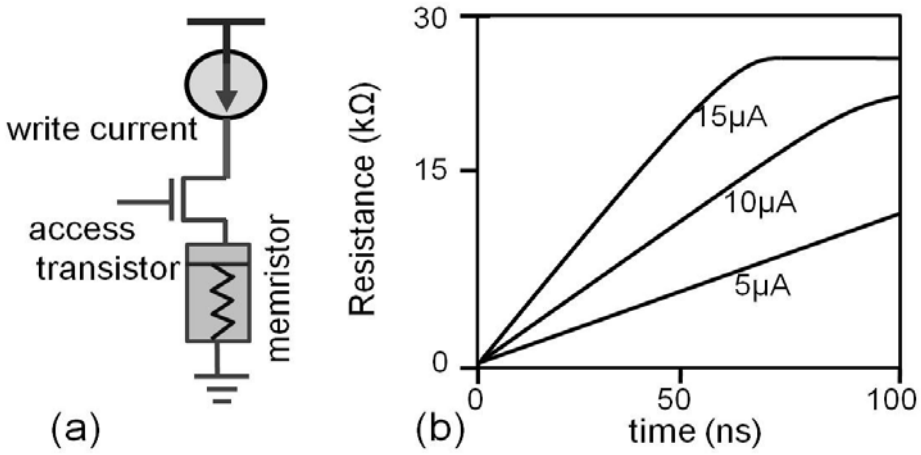


Figure 6.2(a) A resistive memory cell with access transistors, (b) transient change in resistance for different magnitude of programming current.

achieve more than 8-bit write-accuracy [104]. In a crossbar 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 (fig. 2a), and diodes have been proposed in literature that facilitate selective and disturb free write operations [105].

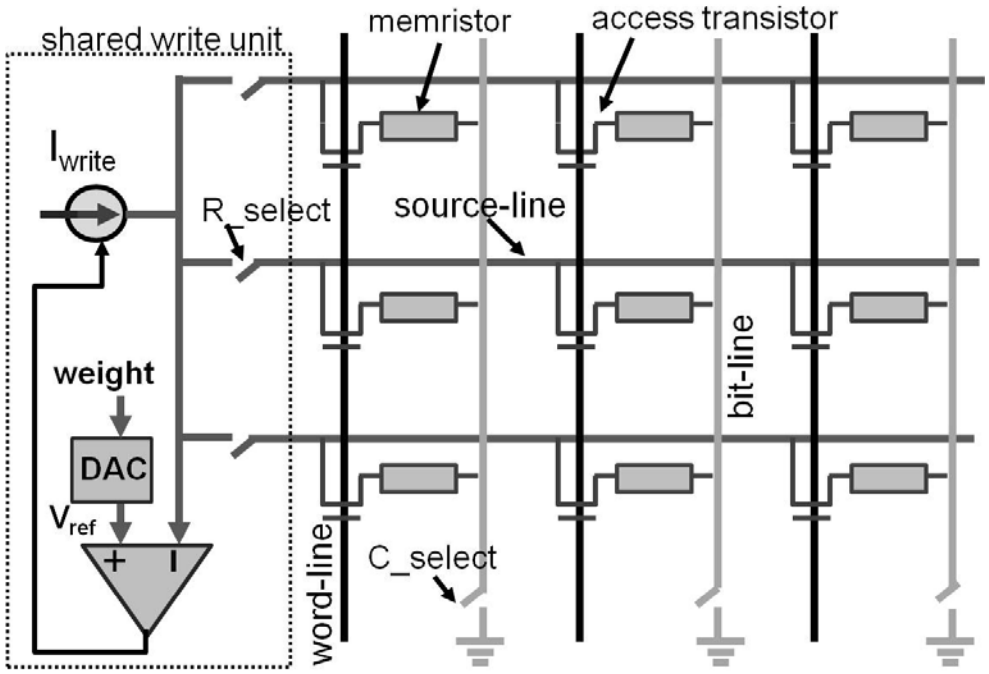


Figure 6.3a A resistive memory array with multi-level programming periphery

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 [106]. Such schemes can be applicable only if programming speed is not a major concern.

Fig.3a depicts a possible array-level schematic of multi-level writing scheme for memristors, using adjustable pulse-width [107]. The memristor-cells to be written are selected by choosing the corresponding set of the word-line, the source-line and the bit line. For infrequent write operations, a single write unit can be shared among large number of rows, as shown in fig. 3a. However, for maximum write-speed, each row can

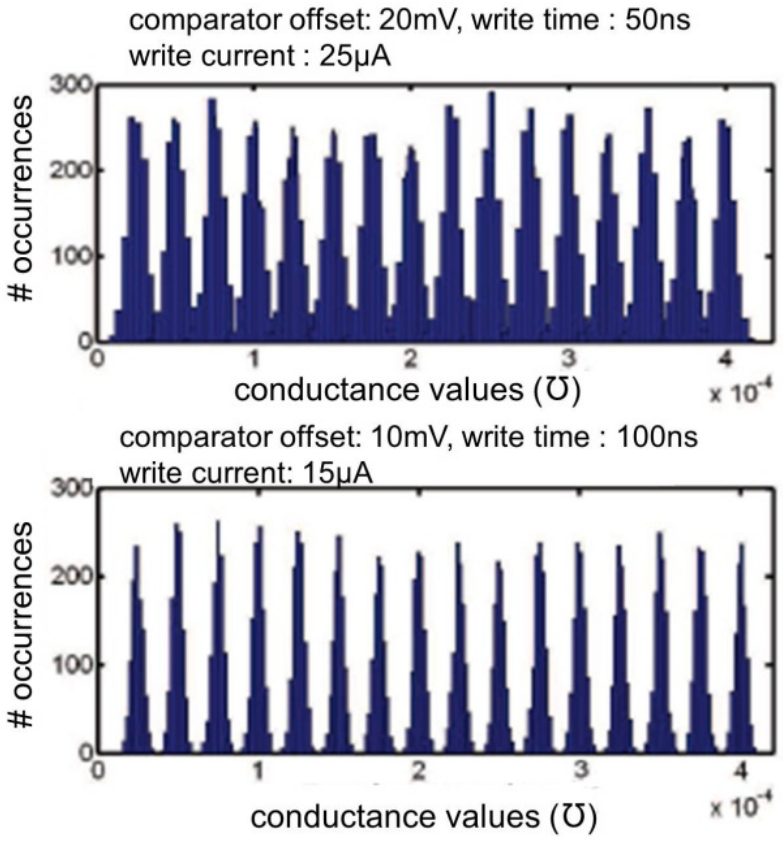


Figure 6.3b Simulation results for feed-back-based write show that higher write precision can be obtained by employing higher resolution comparator and longer write time. These trends have been obtained using analytical model for memristors [1].

have a dedicated programming cell. This would allow writing of one column at a time, by selecting a particular world line.

In order to accomplish the write operation, a constant current can be injected into the selected cell and the voltage developed on the source line is compared with a comparator threshold. The threshold, in turn, is set proportional to the target resistance, by using a compact switched capacitor digital to analog converter (DAC). The current source is disconnected as soon as the accessed memristor acquires the target resistance value. As shown in fig.2a, lower value of write current results in slower ramp in the

resistance value and hence, allows more precise tuning. Analytical model for memristor have been used for simulation in this work [107]. Experimentally it has been observed that memristive devices (including Ag-Si) do exhibit a finite write threshold for an applied current/voltage, below which there is negligible change in the resistance value [108]. As described in the following sections, application of spin-based neurons in RCM facilitates ultra-low voltage (and hence low current) operation of the memristors for computing and hence, can achieve reduced read-disturb for the array.

The write-precision in method described above, is mainly limited by random offset of the comparator, inaccuracy in the current source and the DAC. Larger accuracy would entail higher design-complexity for these blocks and lower write-speed (fig. 3b).

6.2.2 Associative Computing Using Multi-Level RCM

Memory based pattern-matching applications generally apply some form of feature reduction technique to extract and store only the essential 'patterns' or 'features' corresponding to different data samples. The extracted patterns can be represented in the form of analog vectors that can be stored along individual columns of the RCM as shown in fig.1. In order to compute the correlation between an input and the stored patterns, input voltages V_i (or currents I_i) corresponding to the input feature can be applied to the horizontal bars. Assuming the outward ends of the in-plane bars grounded, the current coming out of the j_{th}in-plane bar can be visualized as the dot product of the inputs $V_i$and the cross-bar conductance values g_{ij}(fig. 1). Hence, an RCM can directly evaluate correlation between an analog input vector and a number of stored patterns. This technique can be exploited in evaluating the degree of match (DOM) between an input and the stored patterns, the best match being the pattern corresponding to the highest

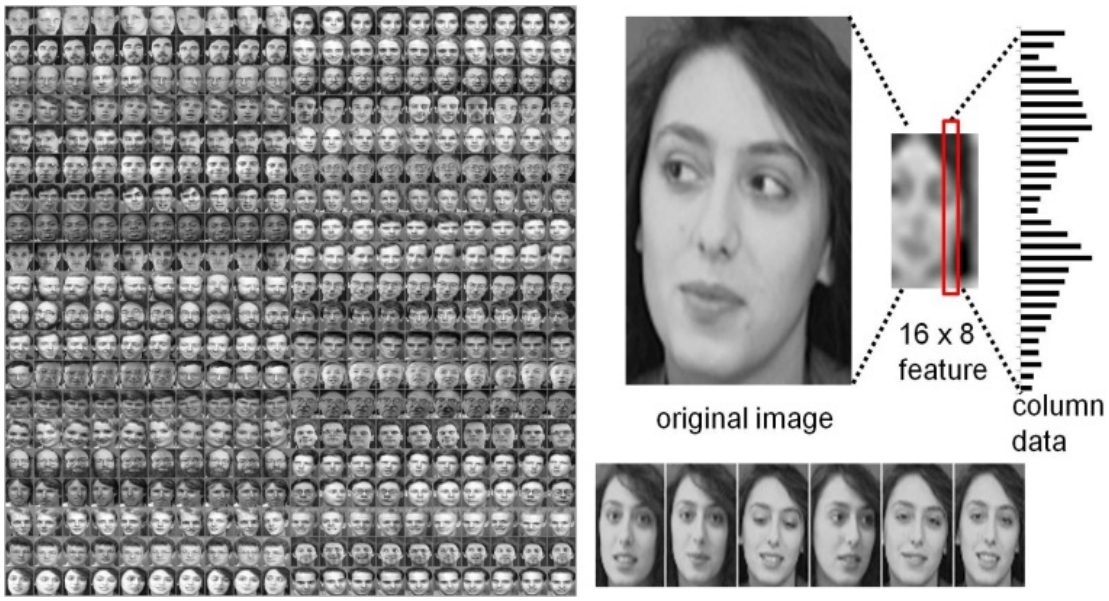


Figure 6.4400 test images of 40 individuals, and the feature reduction method used in this work [109].

correlation magnitude ($\sum_i V_i g_{ij}$). Fig. 4 depicts the feature extraction step for human face-images. In this work, we have used 10 different face-images for 40 individuals, for generating 40 stored data patterns. For an individual, each of the 10 face-images were normalized and down sized from 128x96, 8-bit pixels to 16x8, 5-bit pixels. Pixel wise average of the 10 reduced images was taken to generate 128-element (16x8), 32 level analog patterns (5-bit pixel values) corresponding to the 40 individual faces. The limit of image down-sizing was identified as the scaling factor below which matching accuracy for the 400 test images dropped significantly below the value achieved using the full size image (fig. 5a). For each set of downsizing factor and bit-size, current-mode correlation outputs were obtained using SPICE model of RCM.

Variations in input source as well as memristor values were incorporated to obtain realistic values for the current-outputs. For a given set of stored images, classification

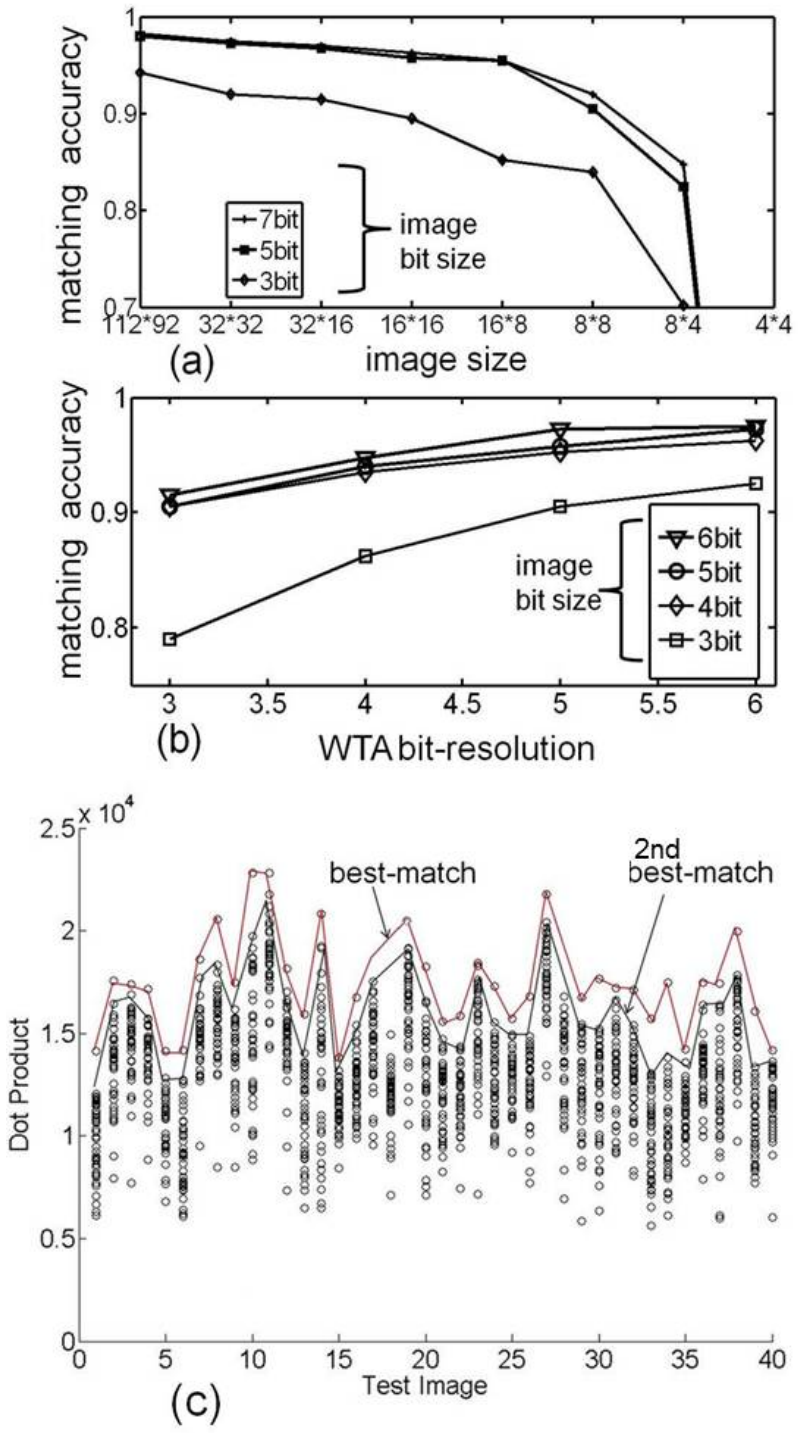


Figure 6.5(a) Training accuracy reduces with image down-sizing, (b) similar trend is obtained for the reducing WTA , (c) dot-products output form the RCM depciting the results for best-match and the second-best match for all 40-template faces when corresponding inut images are provided as input. 3 % σvariation has been used for 32-level analog memristors. A matching accuracy of ~90% was achieved in simulations.

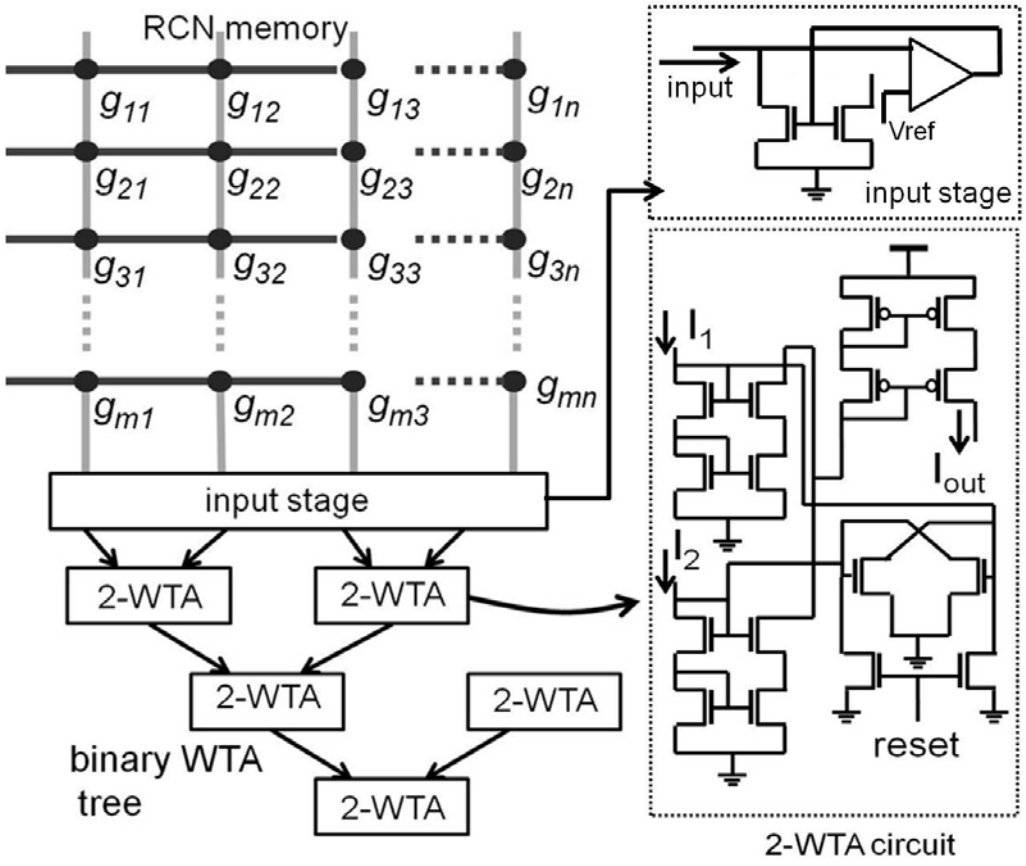


Figure 6.6A standard CMOS solution for associative memory module using binary treewinner-take-all circuit.

accuracy also depends upon the resolution of the detection unit used to determine the DOM figures for all the stored patterns. A resolution (minimum distinguishable difference between analog dot-products outputs) of 4% (5-bit) was chosen based on the observation that up to this value, the classification accuracy remained close to that achievable using ideal comparison (fig. 5b).

Resolving ~4% difference among the current-mode dot product results requires a precision of 5-bits for the detection unit, responsible for identifying the winning pattern. Fig. 6 shows a conventional mixed-signal-CMOS solution for the

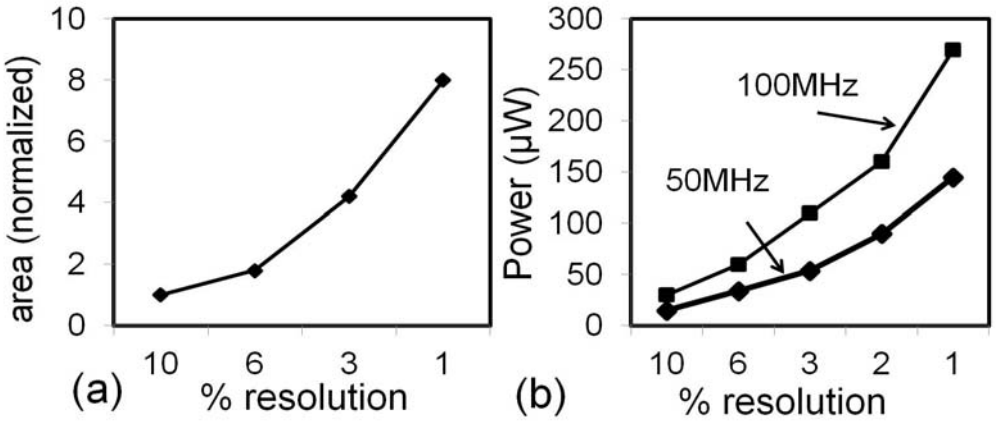


Figure 6.7Design trends for CMOS BT-WTA obtained using SPICE simulation : (a) higher resolution mandates larger cell area (b) for a given bias current, performance trades off with resolution and power consumtion. These results were obtained using SPICE simulation of BT-WTA in [112], with σV_T = 10mV for minimum sized transistors.

detection unit. It constitutes of regulated current mirrors as the input stage that offer low input-impedence and a near constant DC bias to the RCM. Following this, a winner-take-all (WTA) circuit receives the current inputs and determines the 'winner'. Several versions of WTA circuits have been proposed in literature, that can be classified into two broad catagories, current-conveyer WTA (CC-WTA) [111], and binary tree WTA (BT-WTA) [18], the later being more suitable for large number of inputs [111, 112]. BT-WTA employs a binary tree of 2-input comparison stages which involve copying and propagating the larger of the two current inputs to the output (fig 6) [112].

In general, the use of such analog WTA circuits leads to large static power consumption. Infact, the power consumption of an analog WTA unit can be several times

larger than the RCM itself. Morevoer, the performance of such current-mirror based circuits is limited by random mismatches in the constituent transistors and other non-idealites like, channel length modulation, that introduce mismatch in different current paths [113]. In order to maintain a sufficiently high resolution, larger transistor dimensions (both length as well as width) and hence, larger cell area is needed. This is evident from some recent designs [111] -- although the designs used scaled technology, significantly larger channel lengths were used for such circuits. This leads to increased parasitic capacitances and hence, lower operating frequency (fig. 7) for a given static power. Higher frequency and resolution can be achieved at the cost of increased input currents, ie., at the cost of larger power consumption [113]. Special techniques to enhance the precision of current mirrors have been proposed in literature [111], but they introduce significant overhead in terms of power consumption and area complexity. Voltage-mode processing can also be employed in RCM, however it incurrs additional overhead due to current to voltage conversion and subsequent amplifications. This incurs larger mismach, non-linearity and power consumption.

The above discussion suggests that the conventional mixed-signal CMOS design techniques may not be able to leaverage the emerging nano-scale resistive memory technology for memory based computing. This motivates us to look towards alternate device technologies that can be more suitable for this purpose. In the next section we descirbe the spin based neuron model that can lead to efficient computing hardware based on RCM.

6.3 Associative memory module using spin neurons in RCM

In the following subsections, we first describe the design of RCM based correlation unit and its interfacing with domain-wall neurons (DWNs) (fig. 8) presented in chapter-3. This is followed by circuit level description of spin-CMOS hybrid-PE based on DWN that achieves the WTA functionality at ultra-low energy cost. We assumed that regulated

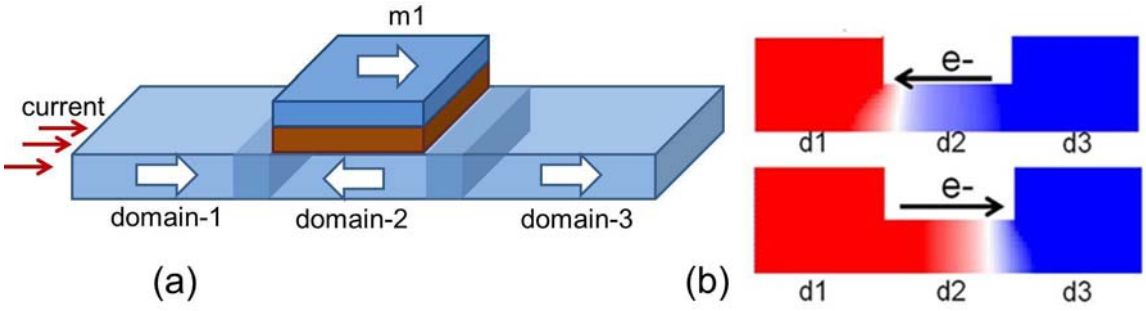


Figure 6.8(a) Spin neuron based on domain wall magnet (b) micro-magnetic simulation for neuron switching.

DC voltages with 1mV accuracy are available [115]. It was shown in chapter 3, that aggressive device-scaling can achieve low switching current (~1μ) and fast switching speed (~1ns) for the DWN. Towards the end of this chapter, the impact DWN threshold on overall performance is presented.

6.3.1 Network Design

Fig. 9a depicts the DWNs with their input (d_1 terminals) connected to RCM outputs. A DC voltage, V, is applied to the d_3 terminals of all the DWNs (access transistors are not shown for simplicity). Owing to the small resistance of the DWN devices, this effectively biases output ends of the RCM (connected to d_1 terminals) to the same voltage. As described in section-2, in order to perform associative matching of an input face-image

with the data stored in the RCM, the input image is down sized to 16x8, 5-bit pixels. Each of the 128 digital values needs to be converted into analog voltages/current levels, to be applied to the RCM input. The low voltage operation of DWN can be exploited to implement, compact and energy efficient current-mode DAC using binary weighted deep-triode current source (DTCS) PMOS transistors, as shown in fig. 9a. A DC supply

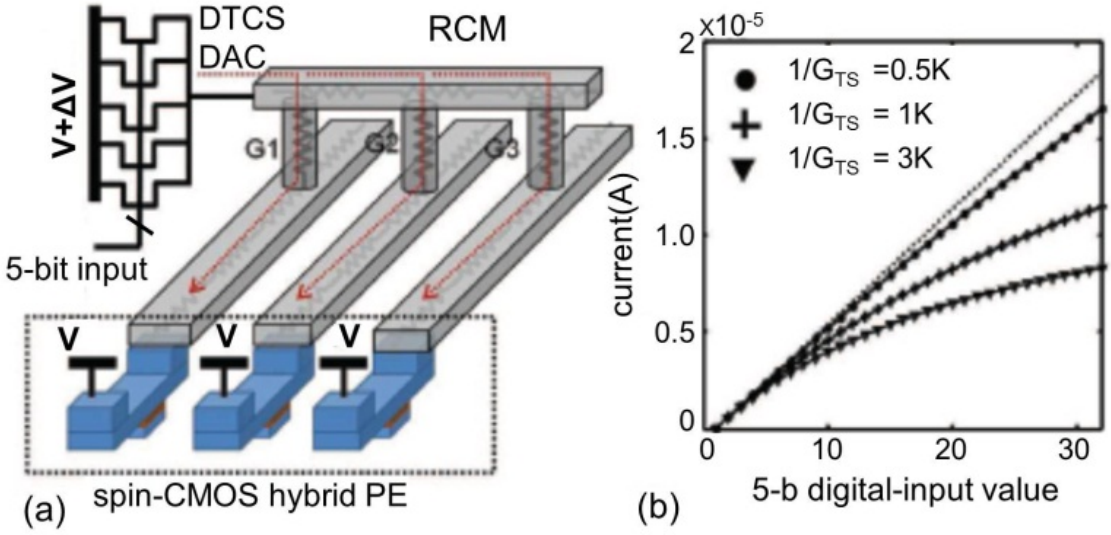


Figure 6.9(a) RCM with a single DTCS input and three receiving DWN, (b) non-linear characteristics of DTCS resulting due to series combination with Gs

of$V$+ΔV is applied to the source terminals of the DTCS, where ΔV is ~50mV. Such a low value of drain to source voltage for the DTCS provide linear *Id* (drain-current)-*Vgs*(gate to source voltage) characteristics that can be exploited for analog-mode driving.

Ignoring the parasitic resistance of the metal crossbar, the drain to source voltage of the DTCS-DAC can be approximated to ΔV. The current I_{in}(i), supplied by the i_{th} DAC can thus be written as $\Delta V.G_T(i)G_{TS}/(G_T(i)+ G_{TS})$, where G_T(i)is the data dependent conductance of the i_{th} DAC and $G_{TS}$is the total conductance (of all the Ag-Si memristors, (including the *ON* resistance of the access transistors if present) connected to a horizontal

bar. Dummy memristors are added for each horizontal input bar such that G_{TS} is equal for all horizontal bars). As a result, the current input through a memristor connecting the i_{th} input bar to the j_{th} output bar (in-plane) can be written as $I(i,j)=\Delta V.G_T(i)G_{TS}/(G_T(i)+G_{TS})(G(i,j)/G_{TS})$, where, $G(i,j)$ is the programmed conductance of the memristor. For accurate dot-product evaluation, the current $I(i,j)$ should be

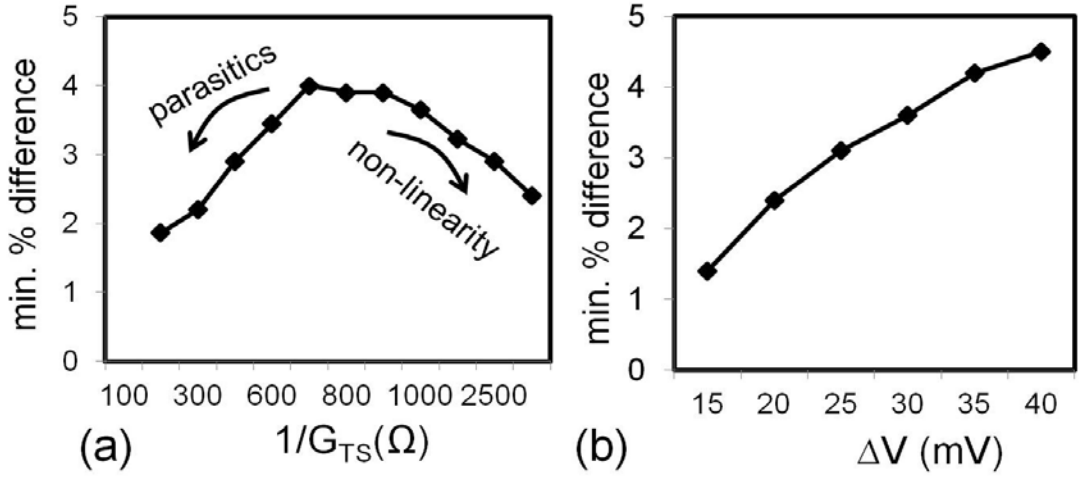


Figure 6.10(a) degradation in detection margin for a given input due to non-linearity (for low G_{TS}) and parasitic voltage drops (for high G_{TS}), (b) degradation in detection margin for the same input, for reducing ΔV, due to parasitic voltage drops.

proportional to the product of G_T (ie, the DTCS conductance, proportional to the input data) and $G(i,j)$. Hence, a low value of G_{TS} (i.e. higher resistance values of the memristors) introduces non-linearity in the DTCS-DAC characteristics (fig. 9b). This leads to reduction in the detection margins (difference between the best and the second best match) for the current-mode dot product outputs for different input images (fig. 10a). As a result, the overall matching accuracy of the network reduces for a given WTA resolution. Ideally, choosing the lowest possible range of values for the memristor resistances (say 200Ω-6.4K Ω, no access transitor being used) would largely overcome the non-linearity

(fig. 10b). However, for higher G(i,j), voltage drop in the metal lines due to parasitic resistances result in corruption of the current signals, once again, leading to degradation in the detection-margin. Hence, the optimal range for the conductance values was found based on the maximum achievable read-margin, as shown in fig. 10a. The Ag-Si memristors can be programmed to low resistance value of ~100Ω. The design parameters like the image compression factor, data bit-width etc, discussed earlier, were therefore determined based on the simulation of RCM model, in order to ensure resolvable detection margin. .

The range of current output from the DTCS-DAC needed is mainly determined by the choice of WTA resolution. If the DWN's are designed to have a threshold of ~1μA, the maximum value of the dot-product output must be greater than 32μA for a 5 bit resolution for the WTA (described later). This in turn, translates to the required range of DAC output current. For 128 element input vectors and 5 bit resolution for the WTA, the maximum value for DAC output required was found to be ~10 μA. This range of current can be obtained using different combination of DTCS sizing and the terminal voltage, ΔV. For a required amount of DAC current, it is desirable to push ΔV to the minimum possible value, in order to reduce the static power consumption in the RCM. This would imply, exploiting the low-voltage operation of the DWNs to the maximum possible extent. The minimum value of ΔV is limited mainly by the parasitic voltage drops that degrades the detection margin and hence the matching accuracy (fig. 10b). For this design (RCM of size 128x40) ΔV of 30mV (with regulated DC supply of 1mV precision [115]) was found to be sufficient to preserve the matching accuracy close to the ideal case (with no-parasitic). The proposed technique effectively biases the RCM across a

small terminal voltage (ΔV), thereby ensures that the static current flow in RCM takes place across a small terminal voltage of ~30mV (between two DC supplies V and $V+\Delta V$).

Above, we noted that the application of DWN in the RCM offers the benefit of ultra low voltage operation that reduces the static power consumption resulting from current-mode, analog computing. Next, we describe the design of spin-CMOS hybrid WTA that performs the winner selection task with negligible static power consumption.

6.3.2 WTA design

The DWN device essentially acts as a low voltage, high speed, high resolution current-mode comparator and hence can be exploited in digitizing analog current levels at ultra low energy cost [24]. The proposed WTA, algorithmically depicted in fig. 11, exploits this fact and clubs a digitization step with a parallel 'winner-tracking' operation.

The first half of the flowchart can be identified as the standard algorithm for successive approximation register (SAR) ADC [24]. The data conversion algorithm employed in an SAR-ADC can be explained as follows. To begin the conversion, the approximation register (that stores the digitization result) is initialized to the midscale (i.e., all but the most significant bit is set to 0). At every cycle a digital to analog converter (DAC) produces an analog level corresponding to the digital value stored in the SAR and a comparator compares it with the analog input using an analog comparator. If the comparator output is high, the current bit remains high, else it is turned low and the next lower bit is turned high. The process is repeated for all the bits. At the end of conversion, the SAR stores the digitized value corresponding to the analog input.

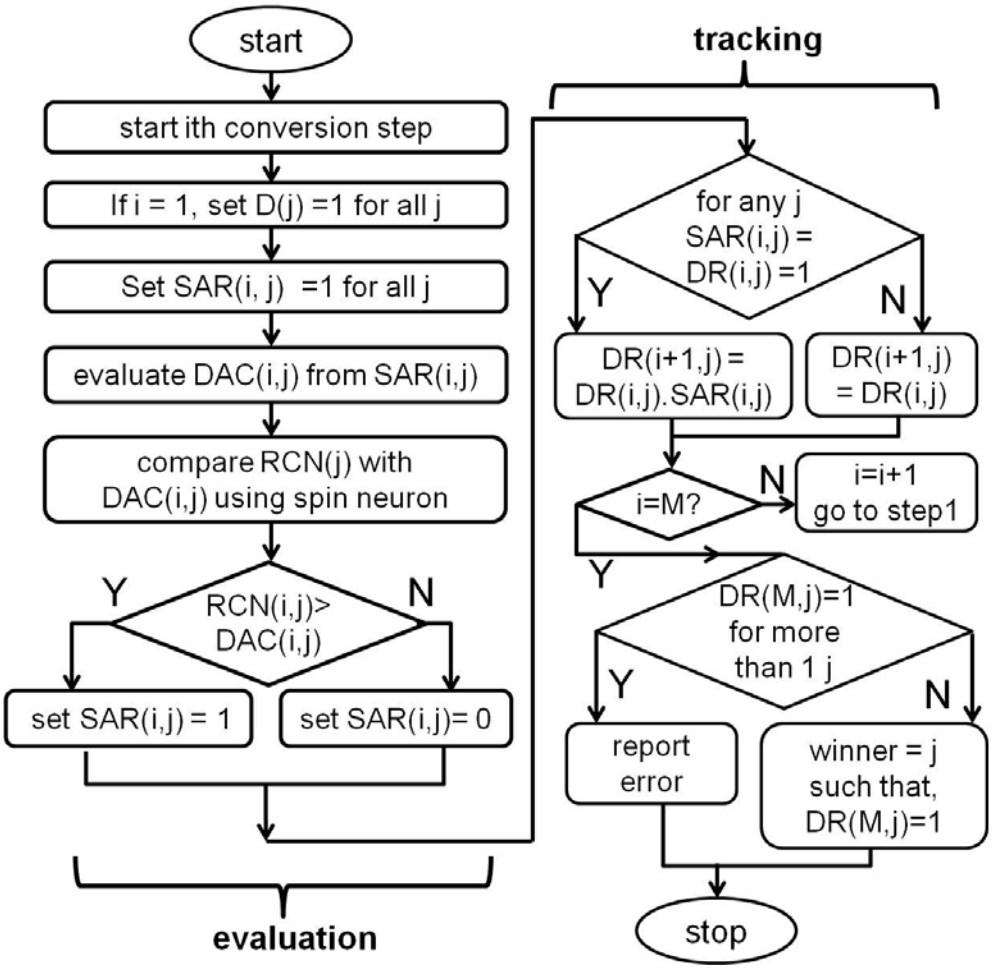


Figure 6.11WTA algorithm used in this work

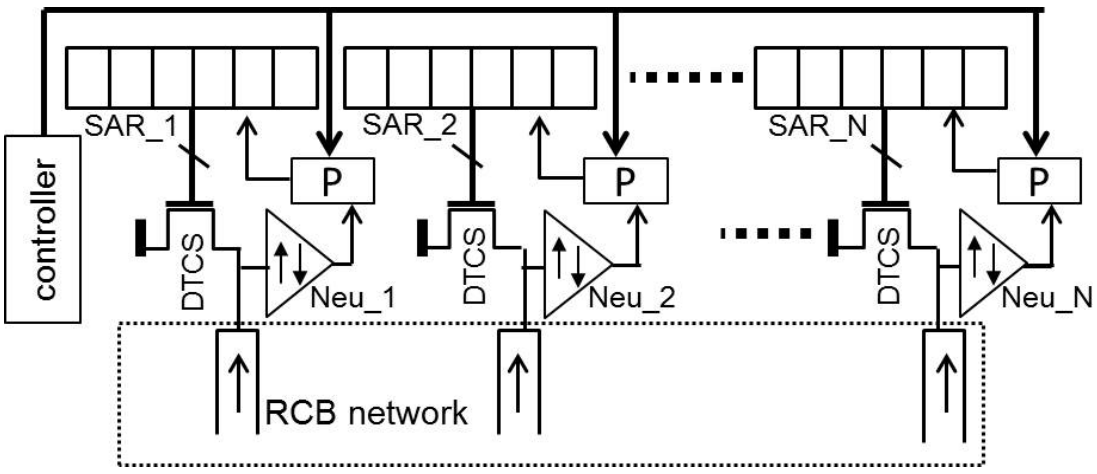


Figure 6.12Block diagram for SAR operation of the WTA circuit

The circuit realization of this operation using DWN's is shown in fig. 12.Output currents of the RCM columns (in this case 40 columns storing the pattern vectors of 40 face-images) are received by individual DWN input nodes that are effectively clamped at a DC supply *V,* as described earlier. Each DWN has an associated DTCS-DAC, which is driven by the corresponding successive approximation register. The drain terminals of the

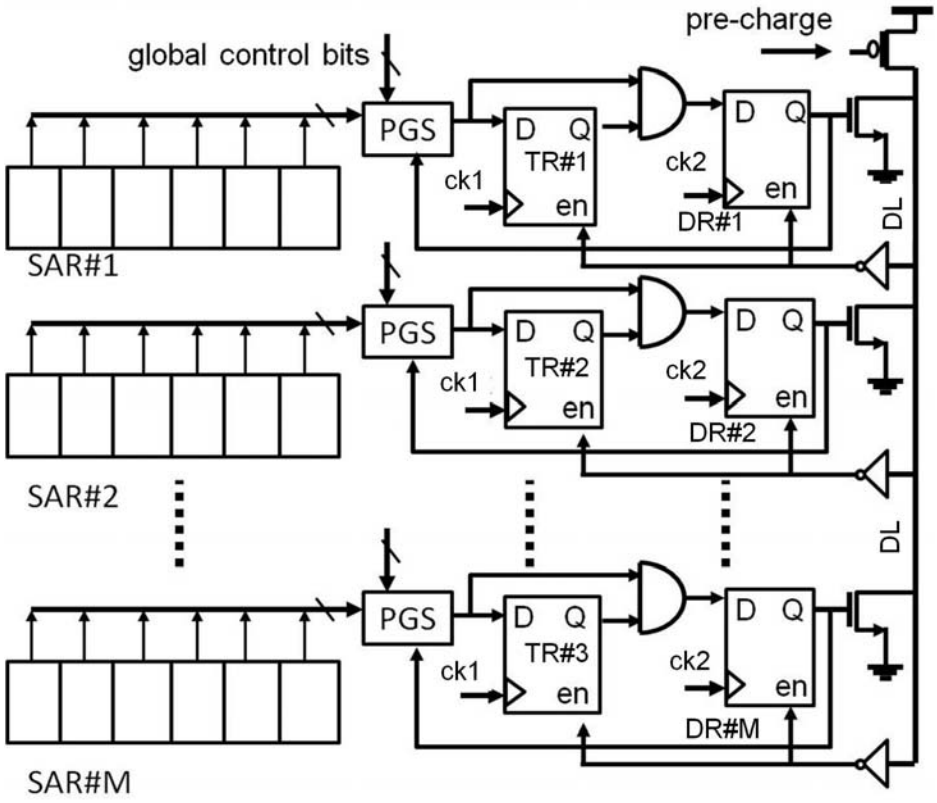


Figure 6.13Circuit operation for the tracking part of the WTA algorithm.

DTCS transistors are a DC voltage *V-ΔV*. In each conversion cycle, the DWN device essentially compares the RCM output and the DAC output (and hence acts as the comparator of the SAR block). The comparison result is detected by the latch described in chapter 3, and the result is used to modify the SAR logic using the scheme described above (though pass-gate based multiplexers *P*, driven by a global controller). In the

overall scheme, the component of RCM output current sunk by the DTCS in the ADC's flow through across a DC level of $2\Delta V$. Note that for a neuron resistance of ~100Ω , the current injection into it towards the final conversion steps (more MSB's) will be less than 5μA (note, only the difference between DAC output and the dot-product output enters the neuron). This leads to a voltage drop of less than 0.6mV which is small as compared to 30mV used in this work.

The second half of the WTA algorithm operates in parallel with the first (i.e., the ADC operation). It can be explained with the help of the corresponding circuit diagram shown in fig. 13. 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, the first and second the 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

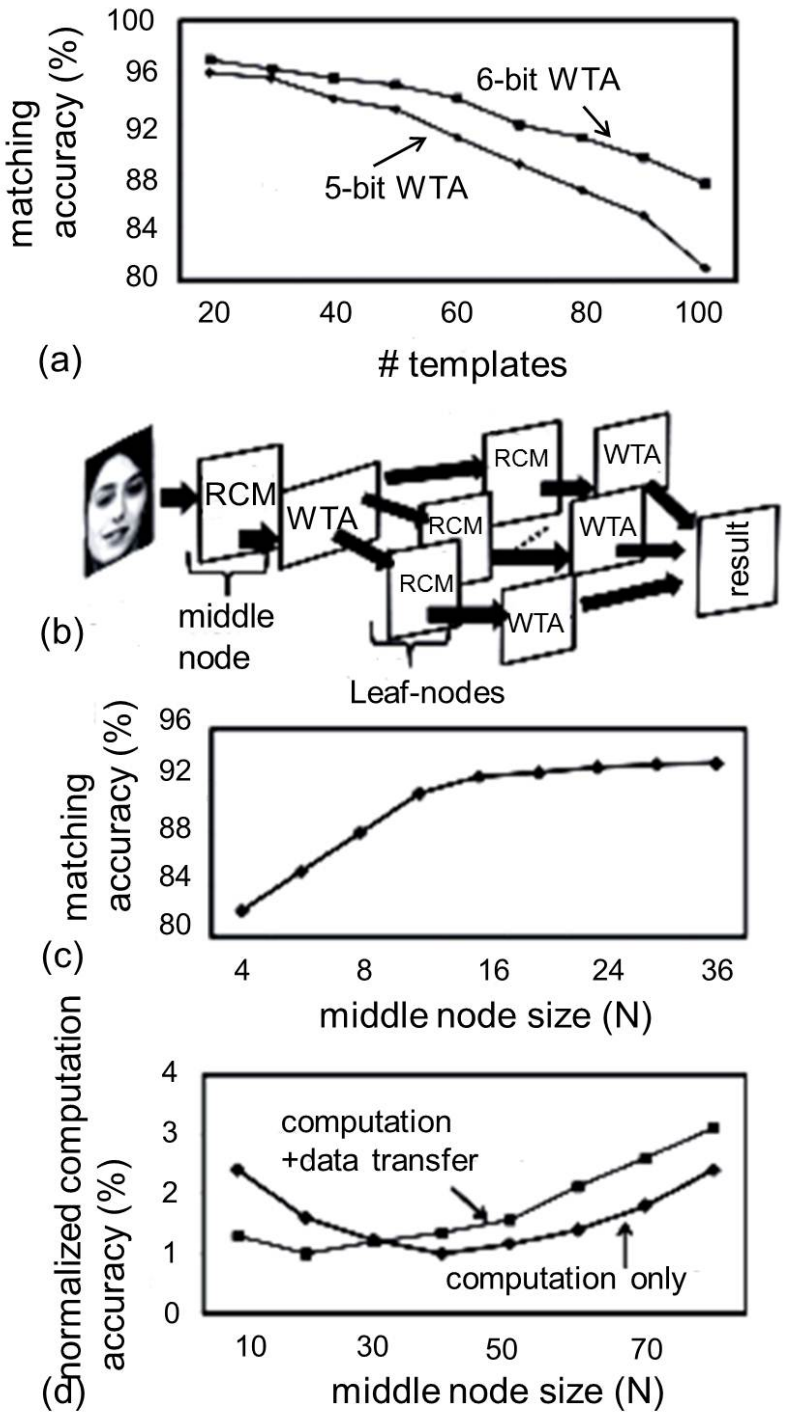


Figure 6.14(a) Degradation in matching accuracy with increasing number of templates for single step matching (for a given WTA resolution), (b) 2-level search tree obtained using K-mean clustering, (c) matching accuracy vs. size of the middle node for a training set with 3000 images, (d) computation energy for different number of middle nodes obtained using K-means clustering for 3000 images.

identified as the winner and the corresponding SAR value is effectively the degree of match (DOM). In case a random image is input to the hardware, the proposed scheme will still identify the 'winning' pattern. But if the DOM is lower than a predetermined threshold, the winner is discarded, implying that the input image does not belong to the stored data set.

The winner-tracking circuitry described above is fully digital and does not consume any static power. Moreover, owing to the global digital control, it is easily scalable with the number of inputs as well as the required bit precision. For the data set of 400 individual images (with 40 mean templates stored in the array), the propsoed WTA design of 5-bit resolution resulted in ~90% matching accuracy.

The overall power consumption in the proposed design is drastically reduced as compared to a MS-CMOS realization (described in section-2), due to two main reasons. First, the power consumption in the RCM itself is significantly lowered due to low voltage operation, and second, the fully digital WTA avoids any additional static power consumption. Note that the proposed WTA implemented in MS-CMOS would result in large power consumption, resulting from conventional ADC's. The low-voltage current-mode switching characterisitcs of DWN however, provides a compact and ultra low power digitization technique.

6.3.3 Large Scale Associative Computing System Using Spin-RCM Associative Modules

The hybrid associative module described above can be used to realize a generic, large-scale data-mining system, using appropriate synthesis techniques. As an example, let us consider the design of a face recognition module with large number of stored images.

Using additional image data set in [110], we created a set of ~3000 images of ~200 individuals.

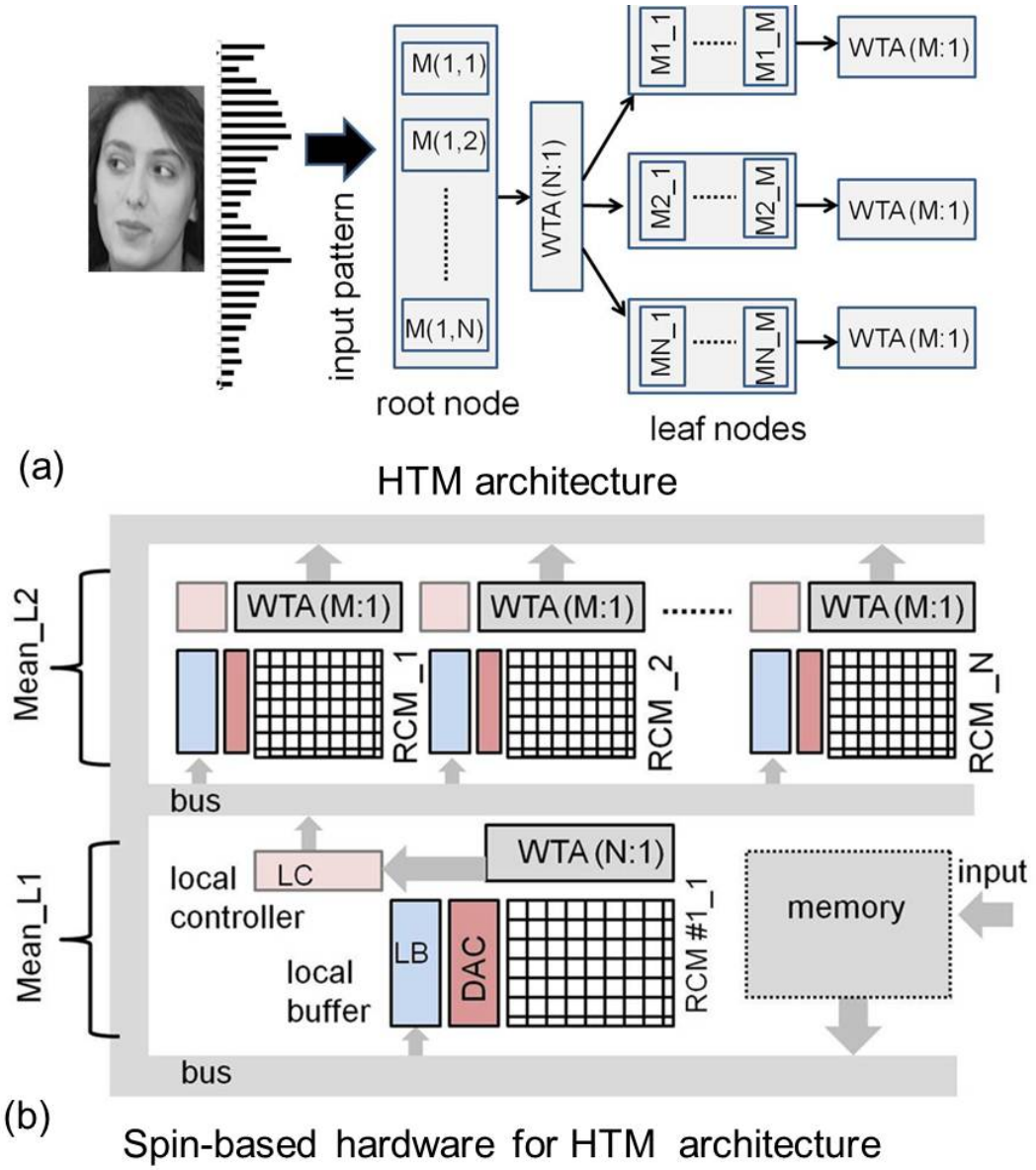


Figure 6.15(a) Hierarchical HTM architecture, (b) HTM-based associative computing architecture based on the proposed spin-based hardware.

Fig.14 a shows that with increasing number of individual images, the matching accuracy reduces steeply, for direct, 1-step matching operations. Higher matching accuracy for a large data set can be obtained adopting a two-step search. *K*-means clustering can be used to design such a network [116]. Using this method, *N* mean-images are obtained from the actual 3000 images that are stored at first-stage RCM module (called the middle-node). The input image is first compared with these *N* mean-images in the middle-node. Depending upon the result of first-stage matching, the input-image is routed to one of the *N* 'leaf'-nodes, i.e., the RCM modules that store the actual images (that can include the means of images of the same individuals). Note that, the total number of images in all the *N* leaf nodes maybe lower than the actual number of images used in the *K*-means clustering algorithm to form the network. This is because, the images of the same persons may be averaged and combined if they fall into the same leaf-node. Fig. 14b pictorially depicts the two-step associative matching procedure described above.

The optimum choice of *N* depends upon the computation accuracy as well as overall computation energy. For small values of *N*, matching accuracy is low (fig. 14c). For large values of *N*, computation energy starts increasing due to larger size of the leaf nodes as well as larger energy overhead due to data-communication (fig. 14d). Computation energy plots for two different cases are depicted in fig. 14d. While considering the energy dissipation only due to the RCM-based computation, the optimal number of leaf-nodes (*N*) was found to be higher. However, as will be discussed later, for larger number of nodes, energy dissipation due to data communication starts dominating. Hence, while considering both the energy components, namely, computation energy and data

communication energy, the optimum value of *N* was found to be lower (fig. 14d). For the computation load considered in this work, the optimum number of leaf nodes was found to be ~16, where each leaf node stored around ~40 image templates. Note that the energy-dissipation due to data transport can be drastically reduced by the use of recently proposed spin-torque based interconnect technique [24].

Fig.15 depicts a possible hybrid associative computing hardware based on the proposed scheme. The RCM blocks store the templates corresponding to the different nodes in a multi-level search tree. The input signal can be obtained directly from an image-sensor input using energy-efficient spin-torque interconnect. The CMOS units are responsible for routing the data and control signals. In order to access the area and data-transmission overhead, memristor cells as well as CMOS-WTA modules were laid out in 45nm CMOS technology. Minimum-sized access transistors were employed for the layout of the schematic shown in fig. 3a. The area density per-cell is mainly limited by access transistors. The larger cell-area can, however, be exploited for wider metal-cross-bars. This would reduce the parasitic resistance of the metal lines, thereby allowing a larger array size and lower voltage for computing. The area of the CMOS-WTA was found to be slightly larger than the cross-bar networks of size 40x128 (128 being the input bus-width). This corresponds to an area of ~700μm^2 in a 45nm technology node. The area of the proposed design can be compared with that of mixed analog-digital CMOS WTA presented in [111]. The design presented in this work would consume ~0.018 mm^2 for 40 input WTA. For 45nm CMOS this would translate to ~1200μm^2 , assuming linear scaling of the proposed design with technology node. However, analog designs do not necessarily scale down with technology node, due to matching

considerations and increased process-variations [113]. The area estimated for a 5-bit resolution, 40-input WTA at 45 nm technology node was larger than ~0.005mm^2, which is ~7x higher than that achieved by the proposed design.

6.4 Performance and Prospects

In order to compare the performance of the proposed design with state of the art mixed signal (MS) CMOS design, we simulated two different CMOS BT-WTA topologies proposed in [112] and [111] respctively, using 45nm CMOS technology models. The first design is the standard BT-WTA, whereas, the second is a recently proposed modification. We also simulated a 45nm digital CMOS design that employed multiply and accumulate operations for evaluating the correlation between the 5-bit 128 element digital templates and input features of the same size.

Simulations for MS-CMOS designs show that the power consumption for the WTA unit dominates the total power. On the other hand, for the proposed scheme, there is negligible static power consumption in the WTA operation. However, since, the static power consumption in RCM is also significantly lowered, it becomes comparable to the dynamic switching power in the WTA. This is evident from the trend shown in fig. 16a. It also shows that the static power consumption in the DWN-based design can be significantly reduced by futher lowering the DWN switching threshold. However, the dynamic power remains almost constant and starts to dominate for reduced DWN thresholds.

Plot in fig. 16b shows the impact of transistor process variations upon MS-CMOS designs. The power-delay products are plotted for a WTA resolution of 4%. Note that in the proposed WTA, the impact of transistor-variations in the DTCS-DAC is limited to

just a single step, whereas, the MS-CMOS circuits suffer more due to the cumulative effect of multiple transistors in the signal path. As discussed in section-2, with larger variations, the accuracy and resolution of MS-CMOS circuits like, current-mirrors

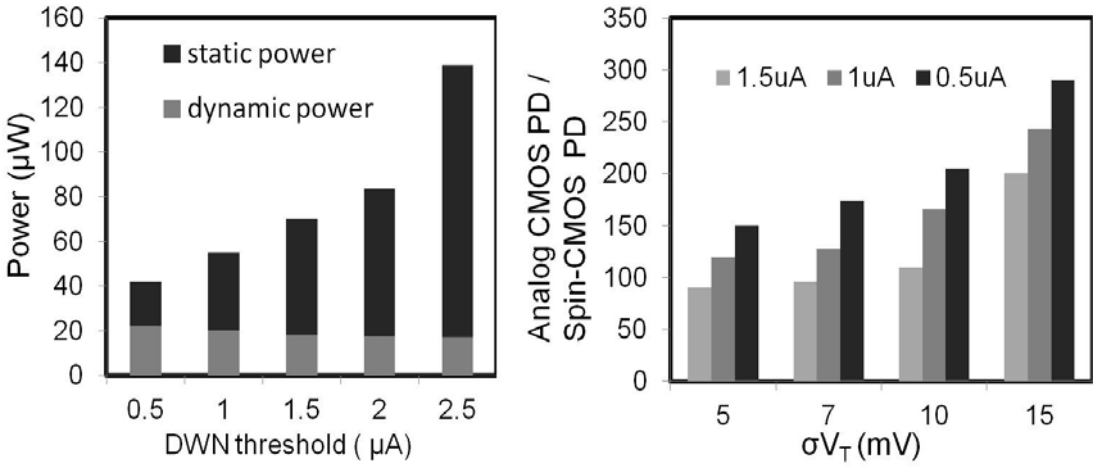


Figure 6.16(a) Power consumption of the proposed design (for 1-step matching for 40 individual templates) swith its static and dynamic components, for different values of DWN threshold, (b) ratio of power-delay (PD) product of MS-CMOS and the proposed design for increasing transistor variations.

decreases steeply, necessitating the use of larger devices, which impairs the circuit performance.

Table in fig. 17 compares the proposed spin-CMOS design with MS-CMOS designs in [111] and [112], and with the 45 nm digital CMOS design. The device parameters for the proposed design used for table in fig. 17 are given in table of fig. 18. The results shown are for σV_T =5mV for minimum sized transistors, which is a near ideal case for MS-CMOS circuits. Results for three different WTA resolutions are given which show similar energy benefits of the proposed scheme, even for smaller WTA resolution. For analog designs, lower resolution constrain allows smaller transistors and hence, better performance. Power consumption for the DWN based design, also reduces with

resolution. Lower WTA resolution allows smaller DAC currents, resulting in reduced static power and lower switched capacitance for the smaller WTA blocks, leading to reduced dynamic power.

		spin-CMOS PE	[18]	[17]	45nm Digital CMOS
Power	5-bit	65μW	5.5mW	8mW	4mW
	4-bit	45μW	2.9mW	5.0mW	2.8mW
	3-bit	32μW	2.3mW	3.2mW	1.2mW
Frequency		100 MHz	50MHz	50MHz	2.5MHz
Energy	5-bit	1	160	215	2460
	4-bit	1	140	221	2300
	3-bit	1	155	210	1100

Figure 6.17 Table for comparison between the proposed design and CMOS hardware, analog and digital

Most interestingly, results for comparison with 45nm digital hardware shows ~1000x lower computing energy for the proposed design. Note that, this comparison does not include the overhead due to memory read in the digital design. As discussed earlier, digital hardware in general prove inefficient for the class of computation considered in this work. Another important point to be noted is that, the use of MS-CMOS circuits in RCM barely perform ~10x better than the digital implementation and hence, achieve far less energy efficiency as compared to the proposed design. Thus, ultra-low energy analog computing using spin neurons can significantly enhance the prospect of RCM technology for computational hardware.

As discussed above, the basic AAM unit discussed in this work can be extended to a more generic, large scale data-mining architecture. The proposed design scheme can be applicable to a wide class of non-Boolean computing architectures that also include different categories of neural networks. For instance, the spin-RCM based correlation

Template size	16x8, 5-bit	Magnet material	NiFe
# template	40	free-layer size	1.5x16x45nm^3
comparator resolution	5-bit	Ms	400 emu/cc
Input data rate	100MHz	Ku2V	20KT
crossbar parasitics	1Ω/μm, 0.4fF/μm	Ic	1.5μA
Crossbar material	Cu	T_{switch}	1.5ns
memristor material	Ag-aSi	Cross-bar material	Cu
		Resistance range	1kΩ to 32kΩ

Figure 6.18 Table for design parameters used in this work

modules presented in this work can provide energy efficient hardware solution to convolutional neural networks that are attractive for cognitive computing tasks, but involve very high computational cost.

6.5 Summary

Emerging RCM technology holds great potentials for non-Boolean computing hardware. However, conventional mixed signal CMOS circuits may fail to leverage the benefits of RCM due to their large power consumption and poor scalability. We showed that the critical analog functionality needed in RCM based computing tasks can be provided by

magneto-metallic "spin-neurons" at ultra-low energy cost. The resulting design can achieve more than two orders of magnitude lower energy consumption as compared to analog CMOS design

7. ENERGY-EFFICIENT AND ROBUST ASSOCIATIVE COMPUTING WITH INJECTION-LOCKED DUAL PILLAR SPIN-TORQUE OSCILLATORS

7.1 Introduction

In this chapter we propose the application of nano-scale Spin Torque Oscillators (STO) for building energy-efficient processing blocks for associative computing [92-99]. STOs are based on magnetic spin-valves that constitute of a 'fixed' and a 'free' magnetic-layer [92]. The spin-polarity of the free-layer (FL) can be set into sustained oscillations by injecting charge-current through the device, under appropriate bias conditions and device configurations. An input-dependent shift in the bias state of a set of phase-synchronized STOs can be employed for pattern-matching applications [92, 93]. However, the choice of the device-configuration and the synchronization-technique, can heavily impact the design-feasibility and the overall benefits of STO-based computing modules.

We propose the application of three-terminal, Dual-Pillar-STO (DP-STO) for associative computing [93]. DP-STO offers an ultra-low-voltage, low-resistance biasing-path leading to low-biasing power. It also provides a high-resistance output-port, providing large output voltage-swing (corresponding to the free-layer (FL) magnetization-state), thereby, minimizing the overhead for sensing the spin-mode oscillations. Injection-locking is used for robust and low-power synchronization that can offer high-immunity to device-noise and parameter-variations [99].

We employ the injection locked DP-STO-arrays in modeling energy-efficient hybrid circuits for hamming-distance (HD) evaluations required for associative-computing. Apart from low-power operation, DP-STO proves amenable for robust injection-locking due to the isolation between input and the output RF signal that it offers. The above factors combined together render DP-STO an attractive technology for realizing, large-scale and ultra-low-energy associative-computing blocks.

7.2 Dual Pillar Spin torque Oscillator for low power operation

In this section we first present the standard, 2-Terminal-STO (2T-STO) and the basic design-conflicts associated with its application in low power associative computing. Following this, DP-STO is presented as an alternative device that can overcome the limitations of 2T-STO for computing applications.

7.2.1 2Terminal-STO:

A standard 2T-STO [6, 7], shown in fig. 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). layer. The ferromagnetic layers have two stable spin-polarization states, depending upon magnetic anisotropy [1]. The magnetization of one of the layers is fixed, while that of the other (free-layer) can be influenced by a charge current passing through the device or by an applied magnetic field. The high-polarity fixed magnetic-layer spin-polarizes the electrons constituting the charge-current, which in turn exert spin transfer torque (STT) on the free-layer [2]. The dynamics of the free-layer is governed by Landau-Lifshitz-Gilbert equation (LLG) [8], as shown in fig.1. It includes a precession term induced by a static magnetic field

H_{eff}(applied perpendicular to the magnetization-plane), a current-induced STT term and an intrinsic damping torque which opposes the STT-induced deflection in the free-layer

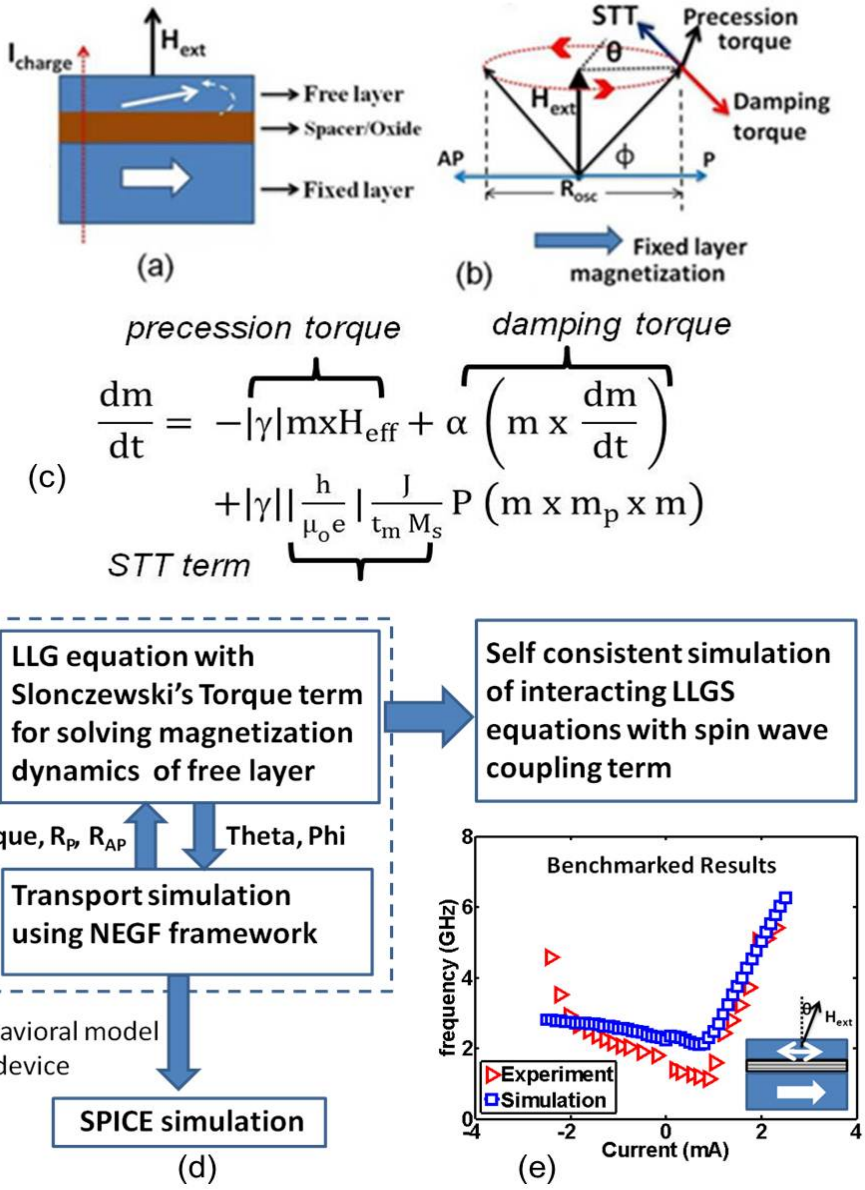


Figure 7.1(a) 2-T STO, (b) different torque terms acting on the free-layer, in presense of a charge-current-*J,* and and external magnetic field H_{eff} , (c) LLG governinig the free-layer magnetization *m* (γ is the Gyromagnetic ratio, α is the damping constant, h is the Plank-constant, t_m is the FL-thickness, M_s is the saturation magntization of the magnet, *P* is the polarization constant and m_p is the spin-polarization of the fixed-layer), (d) self-consistent solution of LLG and NEGF spin-transport for modelling STO, (e) frequency versus bias current plot benchmarked with the experimental data presented in [148].

magnetization. For a given static magnetic field, the free-layer can achieve sustained spin-precession at an angle φ (formed with the plane of ground-state magnetization), at which the STT and the damping torque balance out each other. (Fig 1(a)) [6-8]. The resistance of the spin-valve can be expressed as a function of relative angle (θ) between the spin-polarization of the two ferromagnetic layers as:

$$R = \left(\frac{R_P + R_{AP}}{2} \right) + \left(\frac{R_P - R_{AP}}{2} \right) \cos\theta \tag{7.1}$$

Where, R_P and R_{AP} denote the resistance when the two layers are parallel (θ = 0) and antiparallel (θ = 180). The absolute resistance of a GMR device is much smaller than that of a TMR device (less than ~1 ohm). A GMR-STO, being fully metallic, can be operated with very low voltage (~10 mV). However, the sensed signal amplitude is very low which requires complex sensing circuitry to amplify the signal, leading to high power consumption [93]. 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 large bias voltage, leading to energy inefficiency at the device level. We proposed a Dual-Pillar-STO that can overcome the aforementioned bottleneck and can be suitable for energy-efficient computing [93].

7.2.2 Dual-Pillar-STO

A Dual-Pillar-STO (DP-STO) [93], shown in fig. 2b, clubs the best of a GMR and a TMR-based STO and hence, overcomes the limitations of both the 2-terminal devices. The three-terminal DP-STO employs an extended free layer magnet-m1. Towards the

right it forms a low-resistance GMR interface with one fixed magnet layer-m2, and, a TMR interface with another fixed magnet layer-m3. A simple CMOS

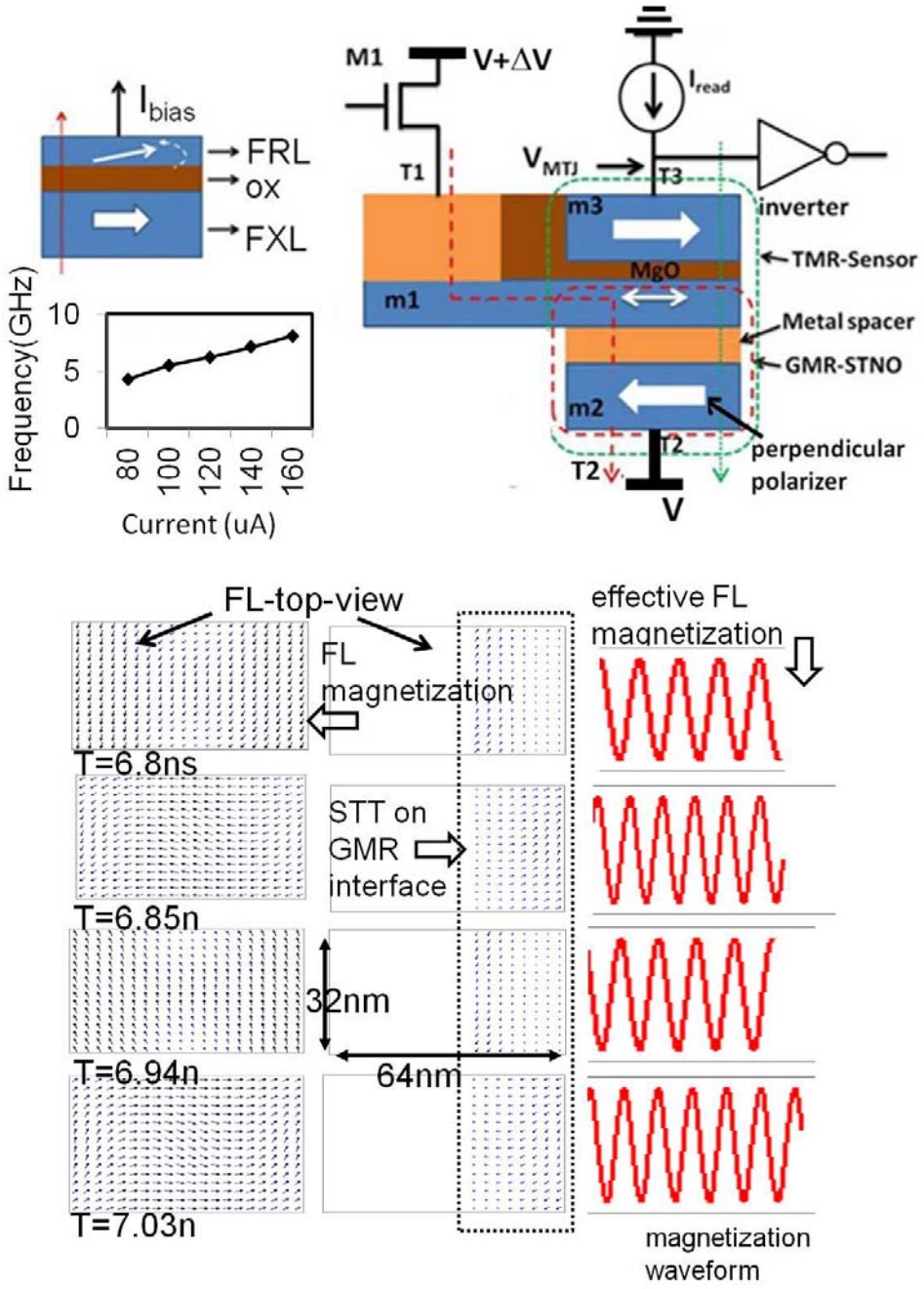


Figure 7.2(a) Conventional 2T-STO (FRL: free-layer, FXL: fixed layer, ox: oxide), (b) DP-STO with perpendicular polarizer and associated biasing and sensing circuits, m1 is the free layer with dimensions: 44x22x2 nm3, (c) micro-magnetic OOMMF simulation plots for DP-STO, (d) freq. vs. DC bias current for DP-STO in fig. 4b.

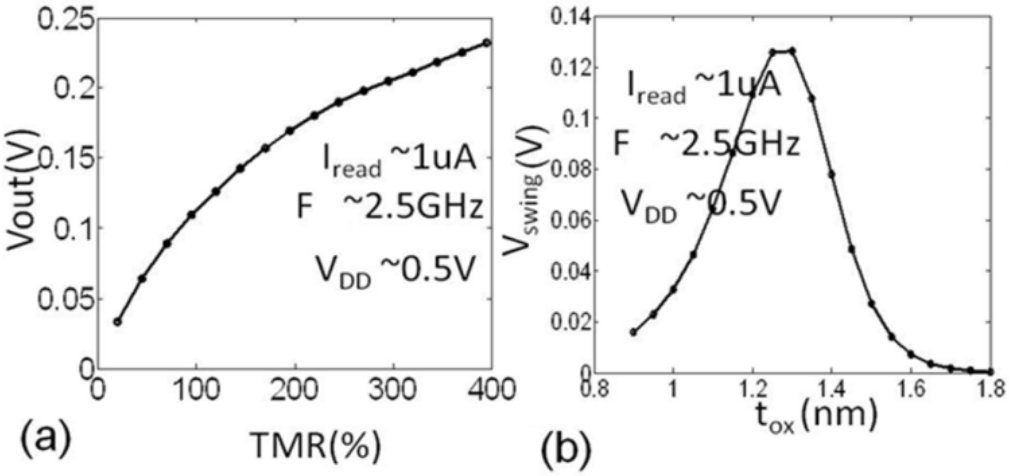


Figure 7.3(a) increase in output swing with TMR, (b) Effect of t_{ox} on MTJ output swing

interface circuitry for biasing the DP-STO and sensing the oscillations is also shown in Fig 4b. Input bias current which sets the free layer in oscillation is applied between terminals T1 and T2 using transistor M1 (dashed line in Fig 2b). Owing to the low resistance magneto-metallic GMR channel, the bias-current can be applied through transistor M1, with a very small drain-to-source voltage, ΔV, (transistor operating in deep triode region). This current induces spin torque on the portion of free layer in contact with GMR interface and sets the magnetization of the free layer into sustained oscillations. Fig. 2c shows the plots for room temperature micro-magnetic simulations for DP-STO with perpendicular-polarizer, biased with ~100μA DC current.

The spin-state of the oscillating free-layer can be sensed by injecting a small read-current into the magnetic-tunnel junction (MTJ) formed between the free-layer m1 and a fixed , reference-layer m3 (fig. 2b). The resistance ratio of an MTJ is defined in terms of tunnel magneto-resistance ratio (TMR) as: $(R_{AP}-R_P)/R_P$x100, where $R_{AP}$and $R_P$are the anti-parallel and the parallel-state resistances of the MTJ respectively. For a

TMR of ~200%, a voltage swing of ~V/3can obtained using the voltage divider (where V is the supply-voltage). Such an output swing can be directly detected by a simple CMOS inverter. Higher TMR may provide higher output-swing and hence better robustness (fig. 3a). High oxide thickness (t_{ox}) for the MTJ provides higher absolute resistance for the voltage-divider, minimizing the read-current and hence the static-power associated with the sensing-operation. However, too high value for MTJ resistance diminishes the output-swing for high-frequency operation, due to low-pass filtering effect (fig.3b).

Table-7.1 Comparison of power consumption for 2T-STO and DP-STO

STO-type	GMR	TMR	DP
Bias-voltage	30mV	0.7V	30mV
Bias power	2.7μW	63μW	2.7μW
Sensing power	2.5mW	0.14μW	0.14μW
TMR(/GMR)	20%	200%	200%
common parameters			
Free-layer size : 22x44x2nm^3 :α: 0.01 ; A = 20pJ/m ; Ms : 400emu/cc, DC bias : 100μA (perpendicular polarizer) Bias frequency : ~5GHz ; Eb : 40K_BT			

Thus, a DP-STO leads to low biasing power due to low voltage GMR port and at the same time provides large output signal through the high resistance TMR port [93]. The latter is conducive to compact and low power CMOS interface for sensing the output signal. For a TMR of ~200% and higher a simple CMOS inverter may be used for sensing the output. Table-I compares the power consumption of DP-STO with two terminal STOs based on GMR and TMR devices [93]. It shows that for GMR STO, the

sensing power dominates due the requirement of large amplification. On the other hand, for TMR STO, the biasing power is dominant because of larger biasing-voltage required for the high resistance device. THE DP-STO on the other hand achieves low power for biasing as well as sensing.

Multiple DP-STOs can be phase synchronized through electrical [94] or magnetic coupling techniques [95-96]. Dynamics of phase-synchronous DP-STOs can be utilized in associative pattern-matching operations as discussed in the next section. For a practical associative pattern-matching hardware integration of a large number of STOs might be essential. The DP-STO can facilitate such a large-scale integration, due the simplified CMOS-interface and low-power operation it offers.

7.3 Associative computing using synchronized STOs

Associative pattern-matching operation can be achieved using arrays of synchronized STOs by exploiting their input-dependent locking characteristics[92]. The synchronization can be achieved through magnetic-interaction between the STO-free-layer, or by using different forms of electrical-connectivity. Fig 4(a) shows the transient plot of two coupled STOs (solid and dashed lines) lock over time. In Fig. 4(b) current through one of the STOs is kept constant at 100μA and the current through the second STO is increased from 90 μA to 120 μA. Constant current through the first STO generates a constant frequency of oscillation, whereas, the frequency of the second device increases with its input current. When the frequencies of STOs are far apart, they oscillate independently. They acquire phase and frequency-lock when their frequencies lie in ‘locking-range’, as depicted in Fig 4(b). The locking range can be defined as the

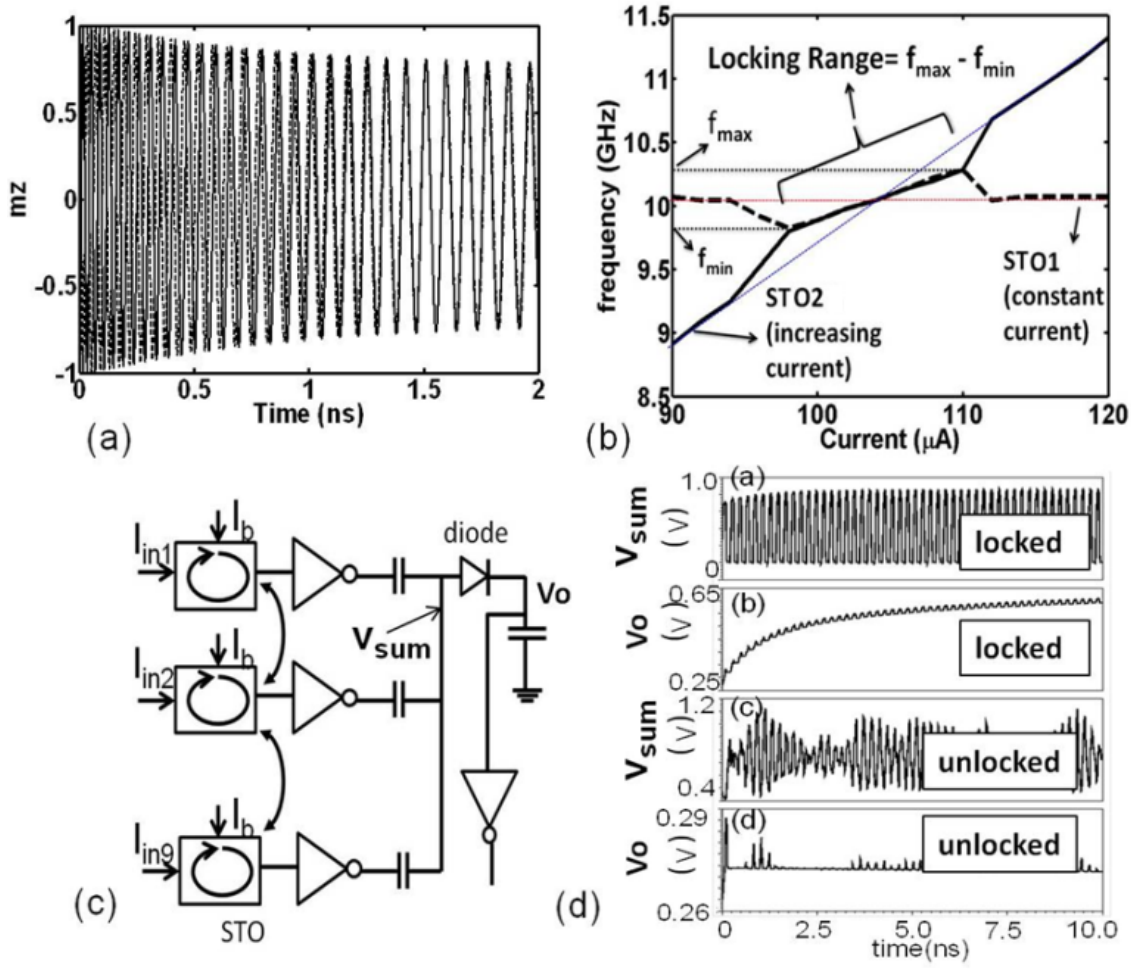


Figure 7.4(a) Transient-plot of phase-frequency locking between two STOs coupled using dipolar-interaction, (b) frequency locking range of two STOs using mono-domain simulation, matched closely with multi-domain micro-magnetic simulations (c) averager and peak-detector circuit for detecting edge-map, (d) transient response of edge-detection circuit for locked and unlocked case.

maximum difference between the DC biases of the two STOs for which phase-lock is retained.

Coupled STOs can be used to evaluate the degree of match between two analog vectors. Fig. 4c shows the circuit for an STO-based associative-module (AM) that achieves this functionality [117]. In this circuit, all the STOs are coupled and are biased with the same DC input. This enforces phase-locked oscillation of all the STOs in the

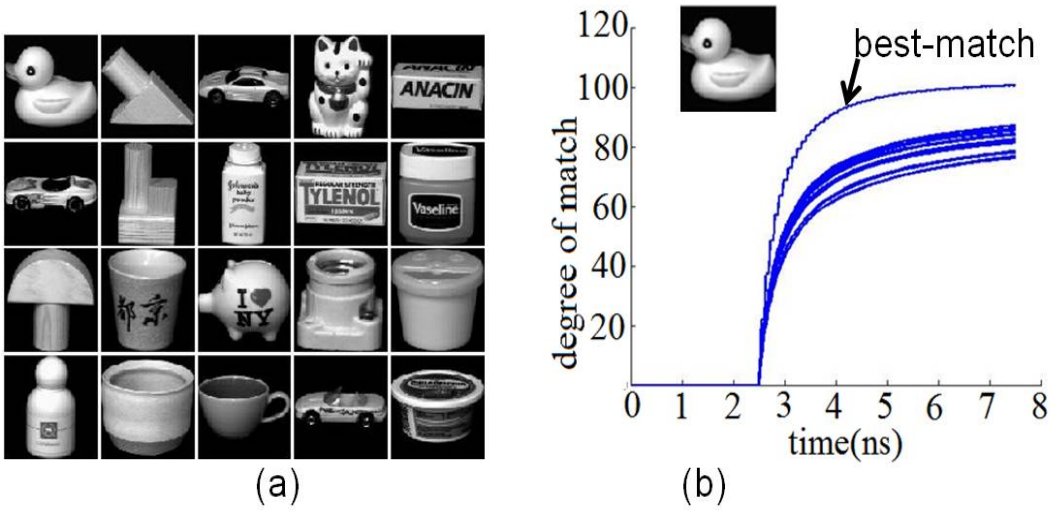


Figure 7.5(a) Image-data-set used in simulation: pixel values corresponding to the individual images were stored as 1-D analog templates, (b) integrator outputs for a particular input image compared with all the other template images.

AM. To compute the associative matching between two analog vectors of *N* elements, current-inputs proportional to the element-wise difference of the two vectors are injected into *N* coupled STOs. If the two vectors closely match each other, the inputs to the STOs are too small to bring them out of the locking range. The STOs therefore retain phase and frequency lock. 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. The circuit shown in fig. 4c performs a capacitive summation of the individual STO waveforms of the AM, and applies the sum to an integrator formed by a diode-capacitor combination [117]. 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. 4d). On the other hand, in the case of un-locked STOs, the summation is an irregular and low amplitude waveform which leads to lower or negligible charging of the output. Thus the case of match

between an input-vector and a template-vector can be identified by comparing the integrator output.

In order to simulate the matching operation for 16x16 pixel images in fig. 5a. The pixel-wise difference between the images and the stored templates were injected into the STOs in different AMs with 8-STOs each (requiring 256/8 = 32 clusters in total). The integrator outputs of all the associative modules were summed and the result was considered as the degree of match (DOM). Higher value of the integrator output implied closer match and vice-versa.

Next we compare two different coupling mechanisms for STOs, namely, magnetic and electrical, for associative computing, with respect to variation and noise tolerance.

7.4 Synchronization mechanisms for STOs

The mechanism for STO-phase-synchronization employed for associative computing, can play an important role in robustness and design-feasibility. In the following sub-sections we discuss and compare two different synchronization techniques for STOs , namely, magnetic-coupling and injection-locking.

7.4.1 Magnetic Coupling

Magnetic coupling may be achieved through spin-wave interaction [96] or dipolar-coupling [94, 95]. Spin-wave coupling may involve interaction through exchange as well as dipolar fields of oscillating magnetic domains, through a shared magnetic-substrate or channel . Dipolar interaction on the other hand, can facilitate locking of physically isolated DP-STNOs lying in close proximity. In this work we employ dipolar-field interaction for coupling multiple DP-STNOs.

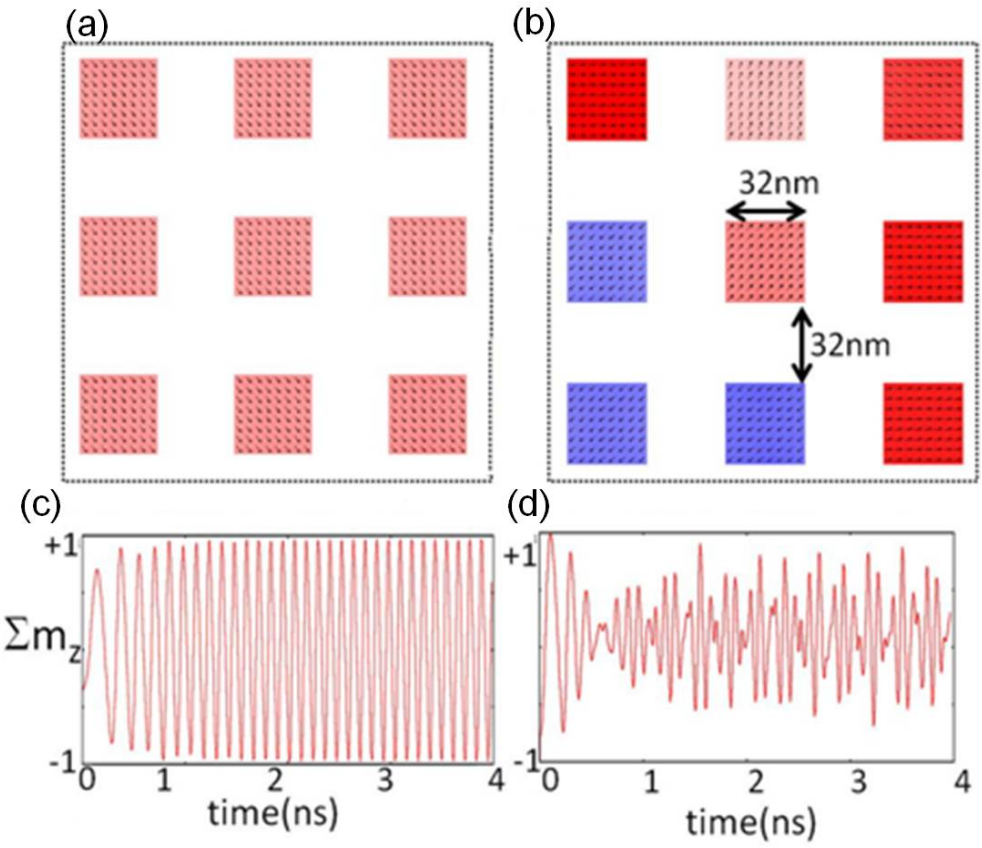


Figure 7.6Micro-magnetic simulation plots for a 3x3 STO array with dipolar coupling (a) for locked case, (b) unlocked case; evolution of average magnetization for the cluster (c) in Fig.6a, (d) in fig. 6b.

Fig. 6a and fig. 6b show the micro-magnetic simulation plots for the locked and the un-locked cases for dipolar-field coupled STOs respectively. In fig. 6a, showing the locked case, the inputs are small and hence fail to disturb the locking due to a common DC bias and near-neighbor dipolar-filed interaction. The average magnetization for this case is shown in fig. 6c. The inputs in the case of the unlocked oscillations, shown in fig. 6b are large enough to overcome the locking, resulting in irregular average waveform, as shown in fig. 6d.

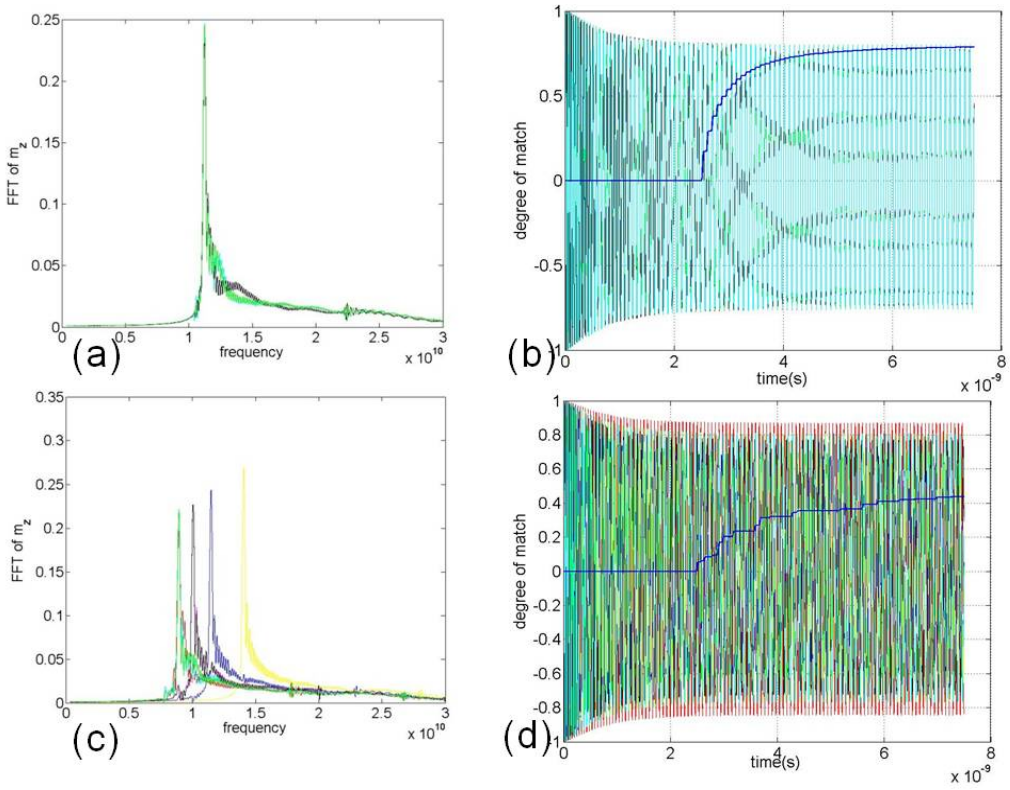


Figure 7.7(a) FFT of 9 magnetically coupled STOs with identical device parameters, (b) overlapped transient waveforms with same DC bias and integrator output (deep-blue curve), (c) FFT of 9 magnetically coupled STOs with 20% spread in Ms and α , (d) STO waveforms and integrator output corresponding to part-c.

We estimated the impact of parameter variation by introducing Gaussian spread in the critical STO parameters like the saturation magnetization M_s and the Gilbert damping constant α. These parameters can have significant spread across multiple device-samples and hence it is important to evaluate the impact of spread in these parameters upon the dynamics of coupled STOs. Towards this end, we simulated associative pattern-matching circuitry based on 9-coupled STOs as described in section-II. Fig. 7 shows that there is effectively no locking for 20% spread in these parameters, for a cluster of 9 coupled

STOs. The integrator outputs for the case of parameter-spread are also compared with that of the ideal case.

The associative matching operation was simulated for the image-set in fig. 5, as describe in section-III. Multiple clusters of magnetically coupled 9-STOs were used to evaluate the DOM (which are effectively the integrator outputs of the individual clusters) for groups of 9 pixels each. The DOM of individual AMs (formed by the 9-STO clusters) were merged to get the overall DOM for the entire image. Fig. 8 a shows the effect of parameter variation on the AM outputs. It shows results for four different degrees of parameter variations. For the ideal case (with zero parameter variations), the best match-case (when the input image matches the template) is clearly distinguishable from the non-matching cases and hence can be easily detected by a coarse-comparator. With the addition of ~10% parameter variation, the best-match case was still correct (i.e, obtained the highest value), but it is too close to the rest of the outputs to be reliability detected. For further higher variations, the best-matching result was found to be incorrect. Fig. 8b shows the difference between the best and the second best matches with increasing parameter variations. The thick lines denote correct match (i.e., the best match being the correct template), whereas the thinner lines connect the points with wrong match. The plot shows that, even for zero-temperature simulations the AM based on magnetically coupled STO fails to perform correctly beyond 5% variations in α and $M_{s.}$ Stochastic Landau-Lifshitz-Gilbert (LLG) formulation was used to incorporate the effect of thermal noise in the STO-dynamics [93]. The corresponding transient plots for AM outputs are given in fig. 9, which show that the best match case was indistinguishable beyond 2% parameter variation.

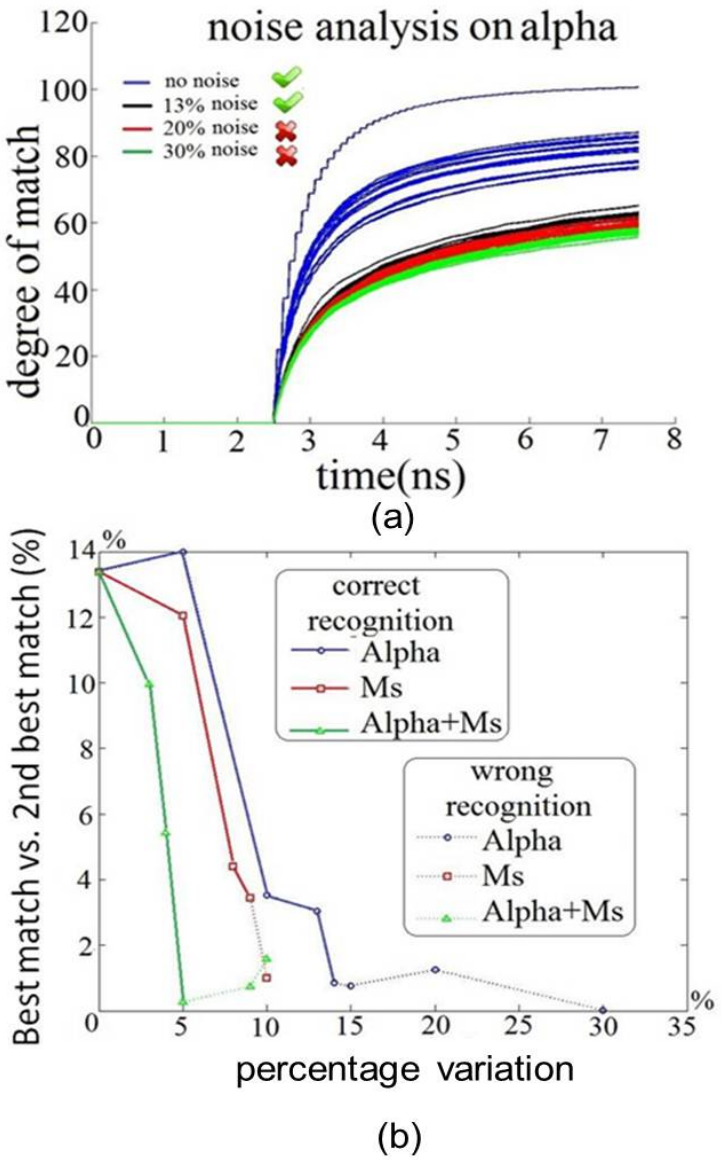


Figure 7.8(a) integrator outputs for three different degrees of parameter-spread using zero-temperature simulation, (b) % difference between the best and the second-best match of the integrator output, for increasing % variations.

The reason behind the high sensitivity of magnetic-coupling to parameter-variations and thermal noise can be visualized by observing the LLG equations governing the dynamics of coupled STOs (assuming mono-domain behavior for each STO) .

$$\frac{dm}{dt} = -|\gamma| m x H_{eff} + \alpha \left(m \times \frac{dm}{dt} \right) + |\gamma| \left| \frac{h}{\mu_o e} \right| \frac{J}{t_m M_s} P \left(m \times m_p \times m \right) \quad (7.2)$$

Here, m is magnetization of the free-layer of the STO, γ is the Gilbert gyromagnetic ration, α the damping constant, t_m is free-layer thickness, P is a constant proportional to effective spin polarization of the fixed magnet and $m_p$is the magnetization vector of the fixed magnet. $H_{eff}$is the effective magnetic field acting on the STO-free layer whose components are expressed in eq.3:

$$H_{eff} = H_{ext} + H_{ani} + H_M + H_{int} + H_{noise} \quad (7.3)$$

Here, H_{ext} denotes an external magnetic field, H_{ani} corresponds to the free-layer's anisotropy field, H_M is the magnetostatic field which is proportional to the component of the free-layer magnetization along its easy-axis. H_{int} denotes the effective magnetic field experienced by the STO-free layer due to its interaction with the neighboring free-layers. (Note that the field due to fixed reference layers in single pillar STO and DP-STO act as static fields and can be coupled with H_{ext}). H_{noise} denotes the noise-term that models the thermal fluctuations. The first term in the LLG equation denotes the 'precession term' resulting from a static applied-field. The second term denotes the 'damping-term' whereas the third corresponds to the spin-torque term. As mentioned earlier, steady oscillation of the STO free-layer is effected when the spin-torque term cancels out the damping term. Note that, this condition is determined by critical device parameters like Ms, α and P, all of which can vary significantly from device to device. As a result, the oscillation frequency and phase of the individual STO free-layers are sensitive to these

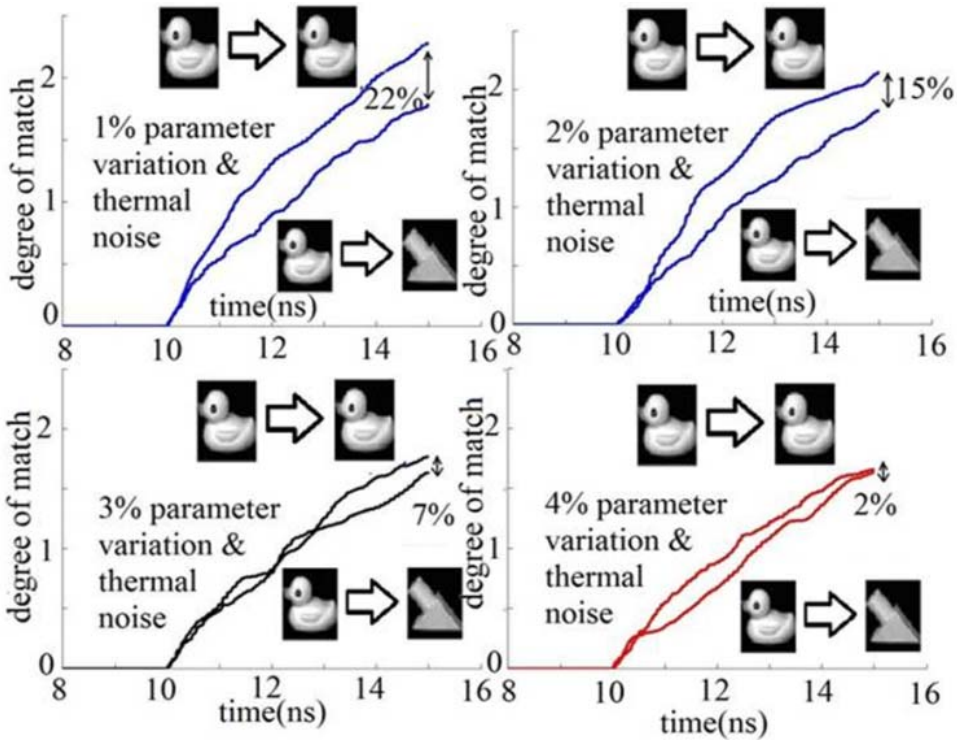


Figure 7.9Integrator waveform for best and second-best match for magnetic coupling.

parameter variations for a given DC bias current *J*. This affects the robustness of the dipolar-filed interaction determined by the field-term $H_{int.}$ H_{int} is proportional to $\sum_{i=1}^{N} C_i m_i$, where m_i denote the magnetization of *N* neighboring free-layers (to which the STO –free-layer is coupled) and C_i denote the corresponding coupling-strength, dependent upon geometry and device parameters [92]. For dipolar coupling used in this work C_i is simply dependent upon the dipolar filed of the i_{th} neighbor and its special coordinate with the STO under consideration [95, 96]. Both C_i as well as m_i are affected by the parameter spread as well as the stochastic thermal noise in the individual free-layers. Thus, the individual oscillation frequencies (without coupling) as well as the magnetic interactions based on dipolar or exchange interactions (in case of spin-wave

coupling) are inherently prone to thermal noise and parameter variations. This leads to weaker-coupling strength and higher susceptibility to these effects.

The foregoing analysis indicates that it might be challenging to build robust associative modules with magnetically coupled STOs due their weak immunity to thermal noise and parameter-variations. We explored an alternate coupling mechanism for STOs that can possibly offer higher robustness. This method, based on RF-injection locking is discussed next.

7.4.2 Injection Locking

In order to establish electrical locking a common RF signal can be injected into a larger number of oscillators [94]. If the RF frequency is close to that of the bias frequency of the STOs (determined by the DC bias), they acquire phase-lock to the injected signal. Fig. 10a pictorially depicts this scheme for two STOs. In this circuit, both the STOs are biased with identical DC voltages, along with identical AC signals.

The frequency of the AC signal is chosen to be close to that of the STO oscillation produced with the DC bias alone. For a significantly wide range of AC amplitudes of the global RF signal, the STOs were found to phase lock with it, at a constant phase-difference (same for all STOs). The phase difference among the different STOs however was close to zero under ideal conditions (zero noise and parameter variations). This implied an effective mutual synchronization and phase-locking among the STOs. Fig. 10b shows the circuit for injection locked DP-STO clusters. In this scheme, a global RF voltage-signal is used to drive the gates of biasing transistors associated with each DP-STO in the cluster. For RF injection, there is no significant

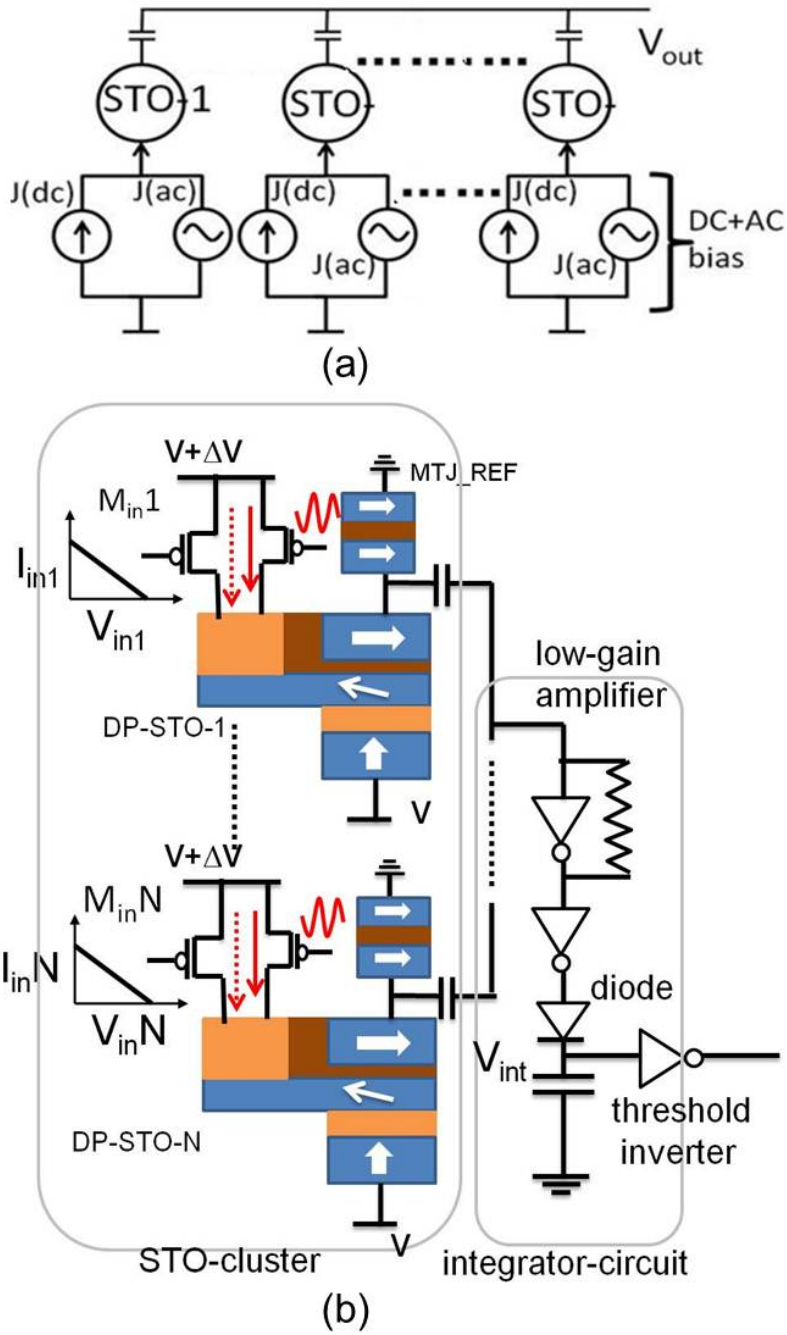


Figure 7.10(a) Two STOs with electrical coupling, (b) transient waveforms for the two STOs showing acquisition of phase-lock, (c) table showing increase in DC and AC locking range with increase in AC amplitude.

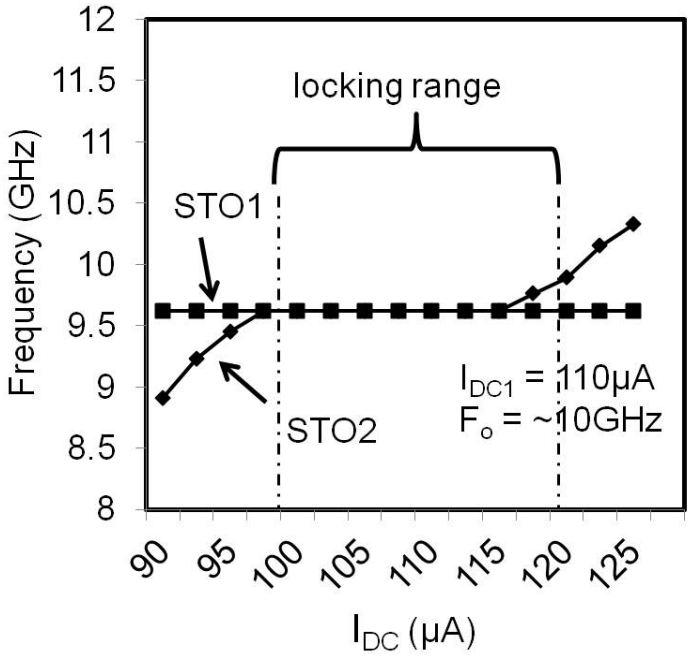


Figure 7.11Oscillation frequency vs. DC bias for an injection locked STO, showing locking range

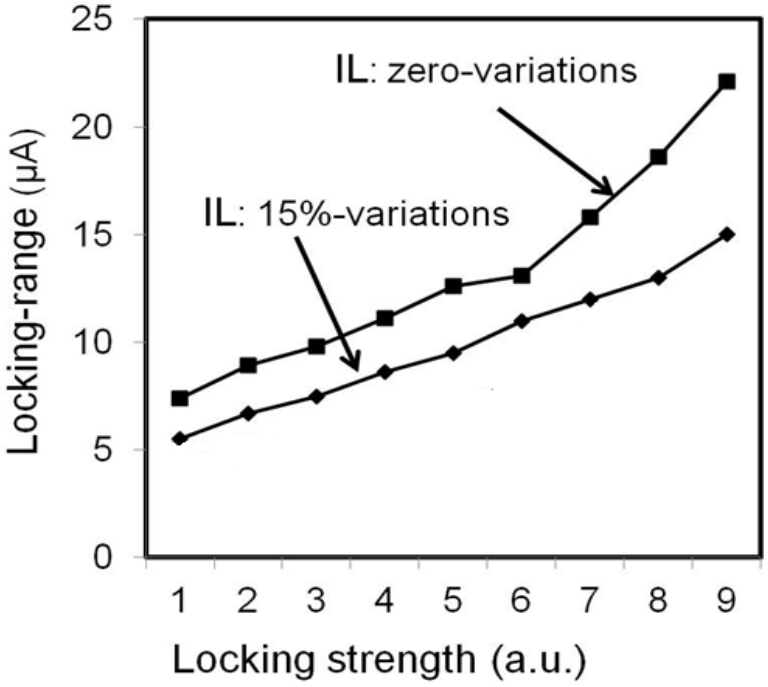


Figure 7.12Increase in locking range with the strength of injection locking , locking strength on the x-axis is proportional to the amplitude of RF injection signal.

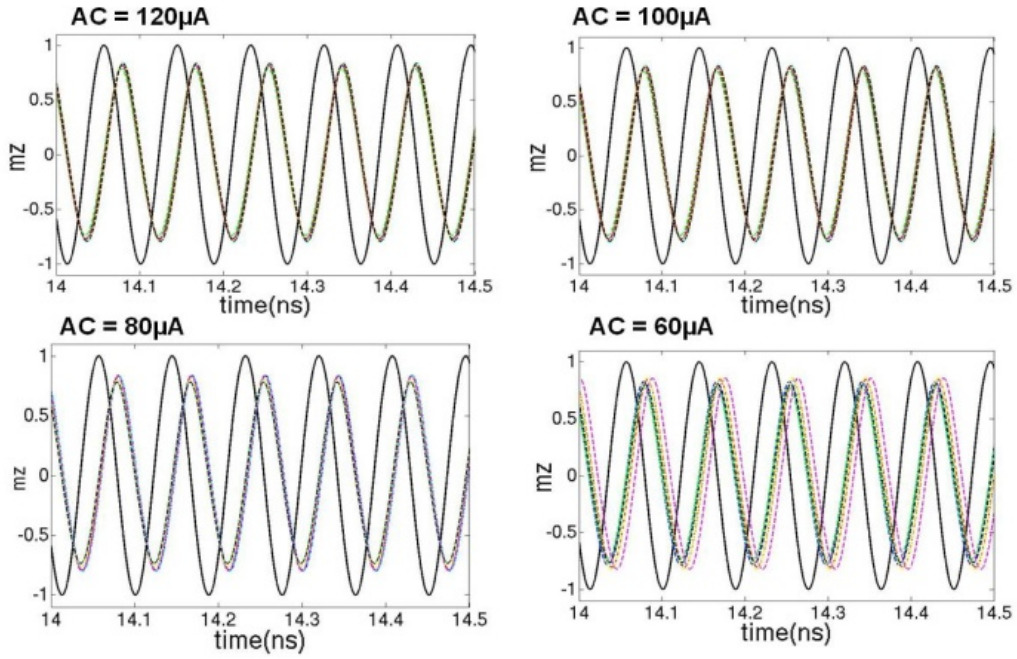


Figure 7.13Transient plots for 8 electrically coupled STOs with 5% parameter variation and thermal noise for different AC amplitudes.

overhead in terms of static power as long as the AC signals has effectively zero DC component. The CV^2 power (switched capacitance power) dissipated for such an AC drive was found to be negligible as compared to the static power due to DC-biasing..

The DC locking range of an injection locked STO can be defined as the maximum difference between the DC inputs of the two STOs for which the phase-lock is retained. The table in fig. 11 depicts the locking range of an injection locked STO (STO2). STO1 is biased a fixed DC current (110μA) along with a RF signal of frequency ~9.5GHz. STO2 is biased with the same RF signal, however, the DC bias for STO2 is swept from 90μA and 125μA. STO2 retains lock to the injected RF signal for DC biases between 100μA-115 μA, thereby, offering a locking range of ~15μA. The locking range can be

improved by increasing the strength of RF-injection, as shown in fig. 12. Thus, the effect of parameter variation and thermal noise can be suppressed by applying stronger RF-bias to the injection locked STOs.

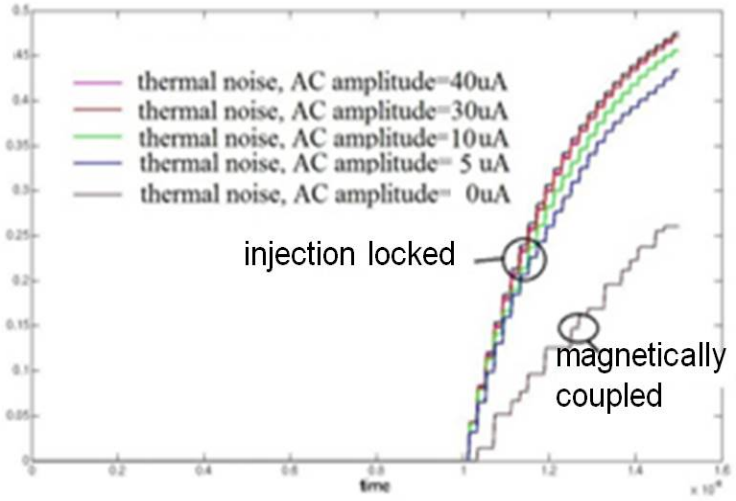


Figure 7.14Effect of increasing RF injection on integrator output.

The effect of increasing RF bias on the STO-transient is shown in fig. 13. The figure shows the output signals for 8 injection locked DP-STOs, biased with a DC current of ~200μA. The plots show reduction in jitter and phase noise with increase in the amplitude, thereby leading to stronger phase synchronization. The effect of increasing AC bias on the locking of 8 electrically coupled STOs is shown in fig. 9, under 5% parameter variation and thermal noise. The solid-line corresponds to the reference AC signal (normalized). The oscillation waveforms for the 8 STOs are plotted using dotted lines.

Fig. 14 shows the output of the integrator circuit for the injection-locked STOs. The increase in the output value results from stronger synchronization and hence cleaner

averaged waveform (obtained by adding the individual STO-waveforms). The injection-locking method depends upon the spin-torque term, specifically the RF component of the bias current J in eq2. The RF component of J, namely, J_{RF} is a global signal which is not

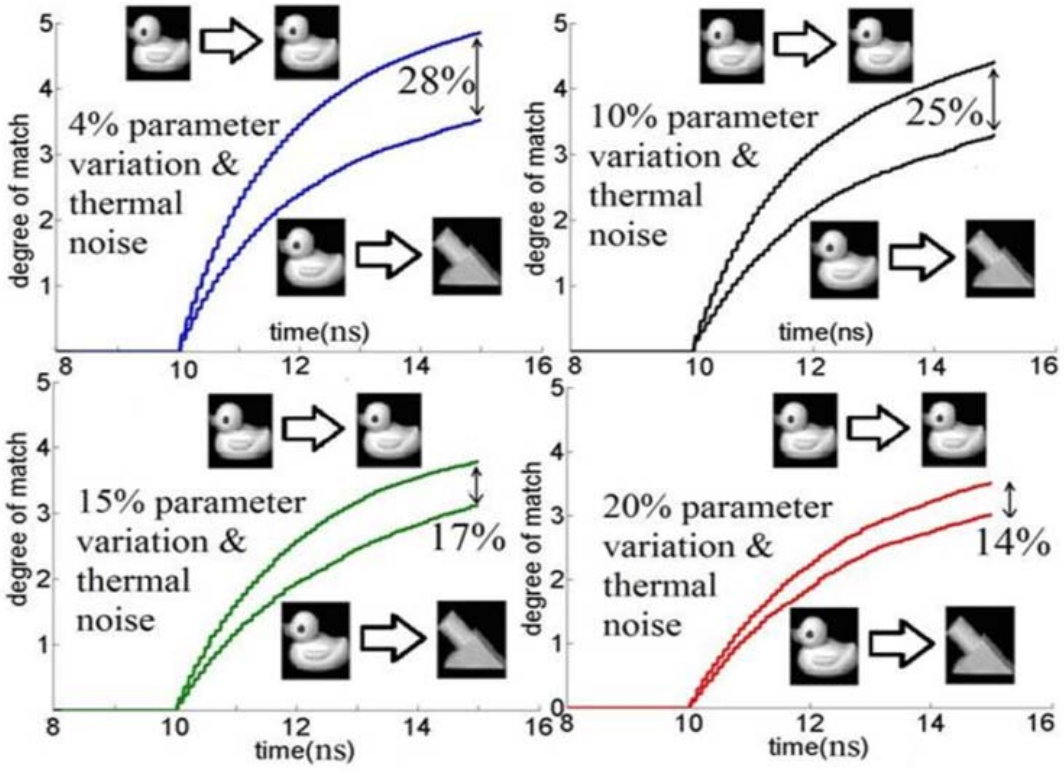


Figure 7.15Integrator waveform for best and second-best match for electrical coupling

affected by the noise of individual magnets. A stronger J_{RF} effectively suppresses the impacts of thermal noise and parameter spread in the dynamics of individual STOs, resulting in stronger injection locking to the external RF signal. Thus, stronger RF-bias improves the tolerance to parameter variation and thermal noise. These results indicate the superiority of the electrical coupling method over the magnetic coupling techniques. The key factor behind this advantage is the use of a common global RF signal in the case

of electrical coupling, which is not influenced by the thermal noise and parameter variations of individual STOs.

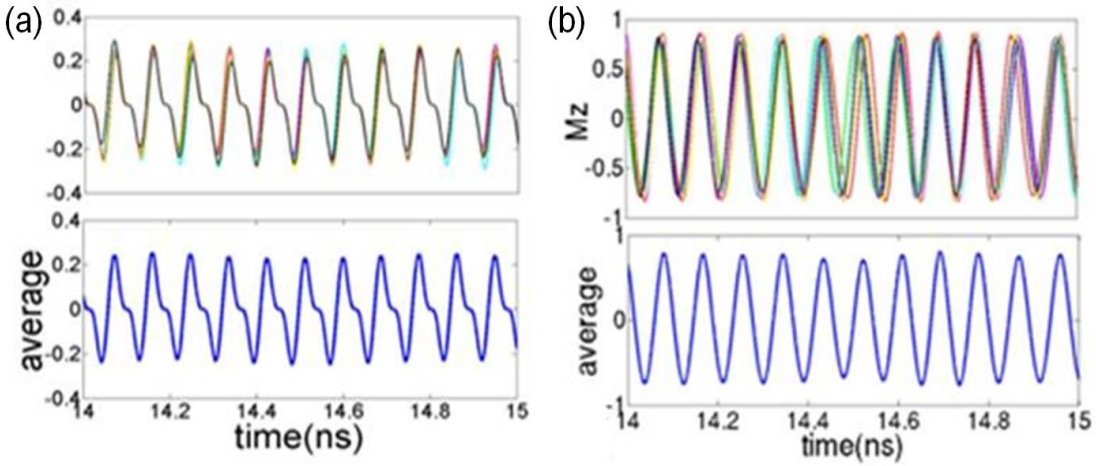


Figure 7.16(a) waveforms for electrically single-pillar coupled STOs, (b) waveforms for electrically 2-dual-pillar coupled STOs

The integrator outputs for the best and the second-best match for AM based on electrically coupled STOs are shown in fig. 15. The plots show that the associative modules could provide distinguishable outputs for up to ~20 % parameter variations, for room-temperature simulations. We used this method to couple up to 32 STOs. No significant degradation in variation tolerance was observed with increasing number of STOs. In contrast, the number of STO that 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 be therefore limited to 9 for a configuration such as shown in fig. 6. Moreover, the effect of thermal noise and parameter variation on individual bias frequencies and the coupling interaction degrades the synchronization strength for such a coupling scheme as discussed earlier. However,

injection locking provides an additional degree of freedom , namely the RF injection amplitude, that can be tuned to achieve desirable degree of synchronization for STO-based associative cluster discussed in section-III.

Apart from low power consumption, another important advantage offered by DP-STO is higher robustness for injection-locking. As mentioned earlier, the proposed electrical-coupling method results in a finite but constant phase difference between the global RF signal and the coupled STOs. For a 2-terminal STOs this results in a distorted output, due the mixing of the RF bias and the STO's own oscillations (which have a constant phase offset). The corresponding plots are shown in fig. 16a. As a result of this distortion the amplitude of the summed output of an STO cluster is found to be significantly lower (~50%) and has lower noise immunity. The DP-STO on the other hand provides isolated paths for the RF bias and the sensed output which is a clean sinusoid, as shown in fig. 16b. Thus, these advantages of DP-STO may be attractive for the implementation of robust and low-power associative modules.

7.5 Summary

We analyzed the impact of parameter-variation and thermal-noise on magnetic and electrical coupling mechanisms for STOs for their prospective application in non-Boolean/associative computing. Results indicate that the injection-locking can be significantly more robust as compared to magnetic coupling techniques. We proposed and analyzed low-power Dual-Pillar STO for low power and compact CMOS interface. We observed that DP-STO can better exploit the electrical coupling technique by due to separation between the biasing RF signal and its own RF output.

8. EXPLORING SPINTRONIC SWITCHES FOR ULTRA LOW ENERGY GLOBAL INTERCONNECTS

8.1 Introduction

The ever increasing demand for higher computing-capabilities has necessitated the integration of multiple processing cores and larger memory-blocks, resulting in increasingly busy inter-chip links and complex, power-hungry input/output (I/O) interfaces for microprocessors [100]. The same is true with respect to on-chip global interconnects like, multi-byte buses and connection-networks for on-chip memory-read and long-distance inter-block links. Moreover, with the scaling of CMOS technology, energy efficiency and performance of the on-chip global-interconnect degrades due to increase in per-unit length resistance of long metal-lines [149]. As a result, the design of inter-chip and on-chip global interconnects has emerged as a major challenge for high-speed computing systems.

Solutions at technology, circuit, and system level have been explored to address the aforementioned design challenges pertaining to interconnects [149-155]. For instance, the use of current-mode signaling for long distance links has been shown to offer reduced power consumption and enhanced bandwidth [151] (fig. 1). This is because current-mode transmission reduces the voltage swing on the metal-lines, thereby reducing the

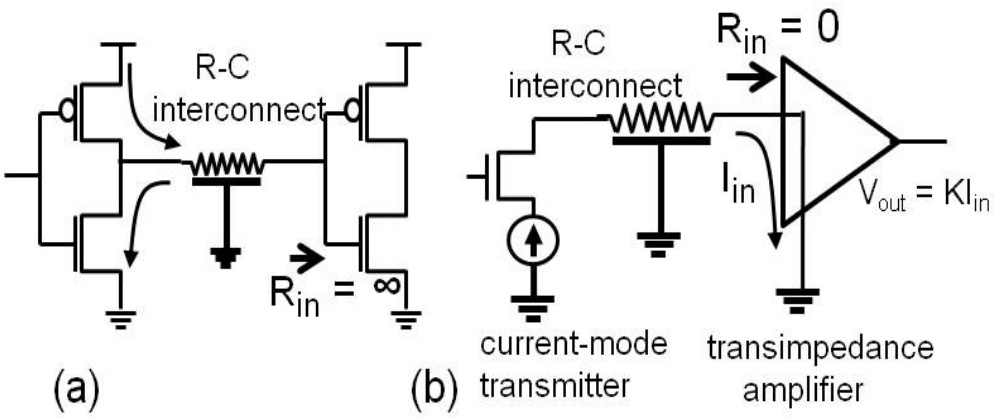


Figure 8.1(a) Voltage-mode interconnect that involves capacitive switching and offers high input impedance to the link (b) current-mode interconnect with a low input-impedance receiver

capacitive switching power. Also, the receiver for current-mode links are designed to provide minimal input impedance to the transmission line (as opposed to voltage-mode links, which provide high impedance capacitive-load). This results in higher bandwidth, as compared to voltage-mode signaling. Increased bandwidth alleviates the need of equalization at the receiver end to a significant extent. However, analog-based current-mode transceivers are more complex than simple inverters, used for voltage-mode links, and add significantly to static-power consumption as well as area complexity, at the I/O interfaces [151, 152]. As a technology solution, use of optical interconnects for inter-chip [153, 154] as well as on-chip data links [155], has been proposed. However, optical modulators (at the transmitting side) and receivers consume large amount of power and area that can eschew their overall benefits [155].

In this work we propose an alternate technology solution that can potentially lead to ultra-low energy, high-speed data links with highly simplified I/O interfaces. Recent experiments have shown that spin-polarity of nano-scale magnets can be flipped

at sub-nanosecond speed using small charge currents [156-160]. Application of current-induced spin-torque switching of nano-magnets for memory and logic-design has been proposed in literature [159, 26, 16]. In this work we explore the possibility of applying such nano-scale spin-torque switches to the design of ultra low-voltage, current-mode on-chip and inter-chip transmission links. Magneto-metallic spin-torque devices, like domain-wall magnet [26], and spin-valves [9], can act as ideal receivers for current-mode signals, owing to their small resistance and the possibility of low-current high-speed switching [156, 12]. Such low resistance receiver ports can allow ultra-low voltage biasing of the entire communication link, thereby reducing the static power consumption (resulting from direct current-paths between supply rails) due to current-mode signaling. Moreover such devices can facilitate easy conversion of current-mode signal into full-swing on-chip voltage levels, through the use of magnetic tunnel junctions (MTJ). As a result, nano-scale spin torque devices can lead to very compact and highly energy-efficient on-chip and inter-chip interconnects for large-scale parallel-computing systems.

8.2 Spin-Torque Switch for Current mode Interconnect

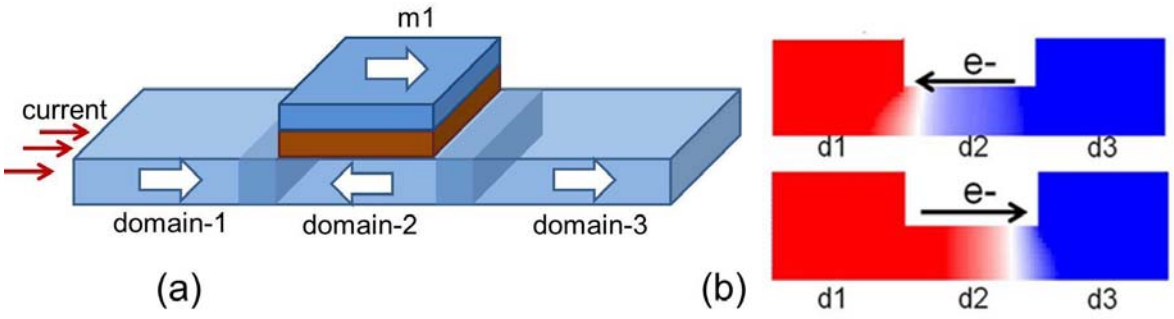


Figure 8.2(a) Spin neuron based on domain wall magnet (b) micro-magnetic simulation for neuron switching.

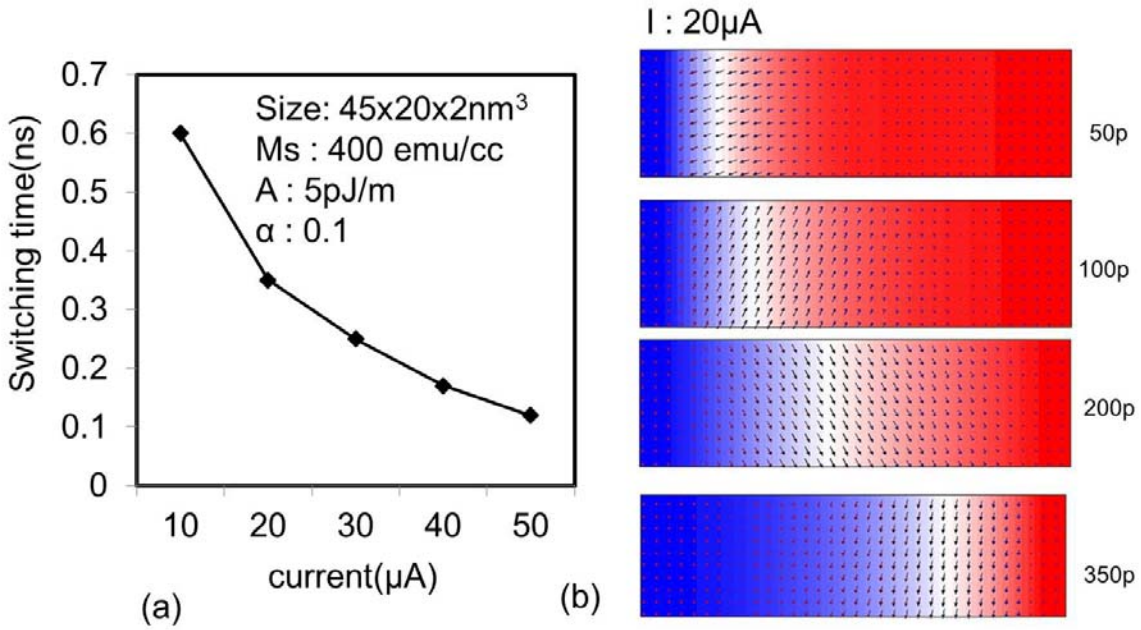


Figure 8.3 (a) Switching time vs. input current for given DWM parameters, (b) micro-magnetic simulation plots for 20μA input current.

In this section we describe design of current mode interconnect using spin-torque switch based on domain wall magnet (discussed in earlier chapters as spin neuron).A brief discussion on the applicability of other spin-torque devices is given in a later section.

Recent experiments have achieved domain-wall (DW) motion in magnetic nano-strips with a critical current density of the order of $10^6 A/cm^2$ [156-158]. Spin-orbit coupling in multi-layer PMA nano-strips can further reduce the switching current, for a given switching time [103]. Such mechanisms can also mitigate the Walker breakdown phenomena, observed for large-current injection in DWM strips, which has been known to limit the maximum achievable speed for domain wall motion [103]. Thus, scaled magnetic nano-strips can be employed to design low-current, high-speed, magneto-metallic switches applicable to high-speed current-mode signal processing.

The device structure for such a domain-wall-switch (DWS) is shown in fig. 2a. It constitutes of a thin and short *magnetic* domain, *d2* ('free-domain') connecting two anti-parallel *magnetic* domains of fixed polarity, *d1* (domain-1) and *d3* (domain-3) (fixed through exchange coupling to larger magnets [5]). Domain-1 forms the input port. Spin-polarity of the free-domain (*d2*) can be written parallel to *d1* or *d3* by injecting charge current along it from *d3*to *d1*and vice-versa. Thus, the DWS can detect the polarity of the current flow into its input node. Hence, it acts as an ultra-low-voltage and compact current comparator [26] that can be employed for recovering data from a bipolar, current-mode signal received at its input. Fig. 3a shows that aggressive scaling of the DWS free layer can achieve low current, high speed switching. As mentioned earlier, application of emerging spin-torque phenomena like spin-orbital coupling can be exploited for lowering the amount of current required for a given switching speed [103].

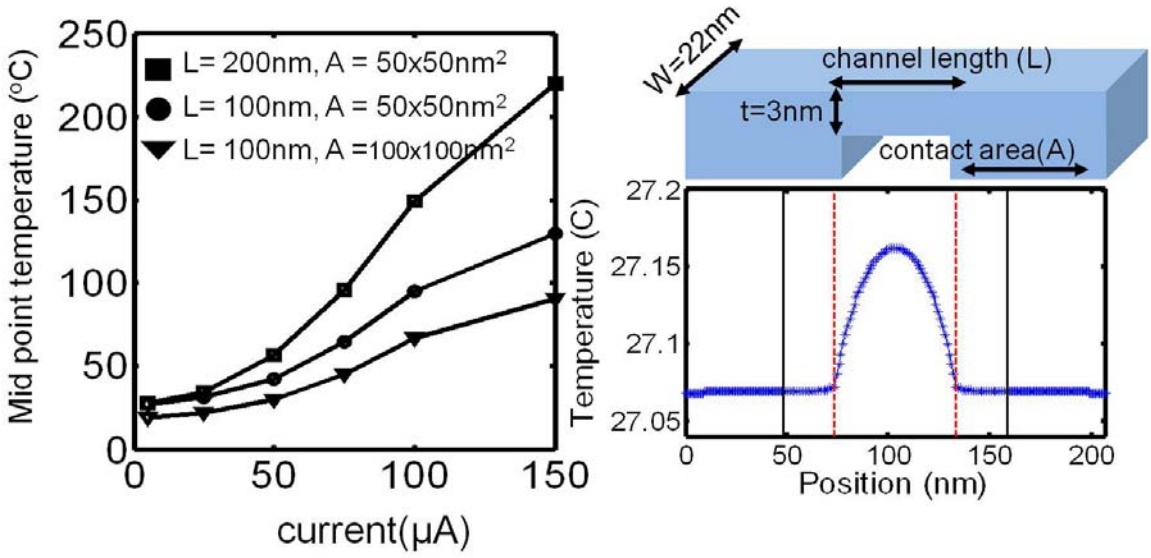


Figure 8.4(a) COMSOL simulation for temperature rise in the DWS device for different device dimensions, (b) plot showing temperature profile along the device for a small input current of ~1μA.

The upper limit upon the permissible current density and hence upon the switching speed may be determined by the Joule-heating effect in the DWS. The effect of Joule heating in the device was simulated using finite-element simulation through COMSOL [161]. The thin and short central free-domain of the device is the most critical portion with respect to current driven heating (fig.4a). Plot in fig. 4 shows that the heating in the device can be reduced by choosing larger contact area of the two fixed domains. Also, shorter free domain results in smaller heating. Thus, the current handling capacity of the device can be increased by appropriate structural optimization.

In order to read the state of the free domain *d2* of the DWS, an MTJ formed between a fixed polarity magnet *m1* and *d2* is employed. The effective resistance of the MTJ is smaller when *m1* and *d2* have the same spin-polarity and vice-versa. A large ratio between these two resistance states, defined in terms of tunnel-magnetoresistance-ratio (TMR) can facilitate simplified read operation. A simple CMOS inverter can be employed to convert the spin-mode information received by the DWS into binary voltage levels.

8.3 Interconnect Design using DWS

Owing toits low-resistance (~100Ω), current-mode switching channel, the DWS can act as an ideal current-mode receiver and can simultaneously facilitate low-voltage (~50mV) biasing of the entire transceiver-link, as shown in Fig. 5a. On the transmitter-side linear region transistors biased at a source potential of +/- ΔV , relative to the DWS are used for supplying the data dependent current. The use of small ΔV (~50mV) achieves low static power dissipation per-bit. Signaling-energy of the proposed interconnect can be

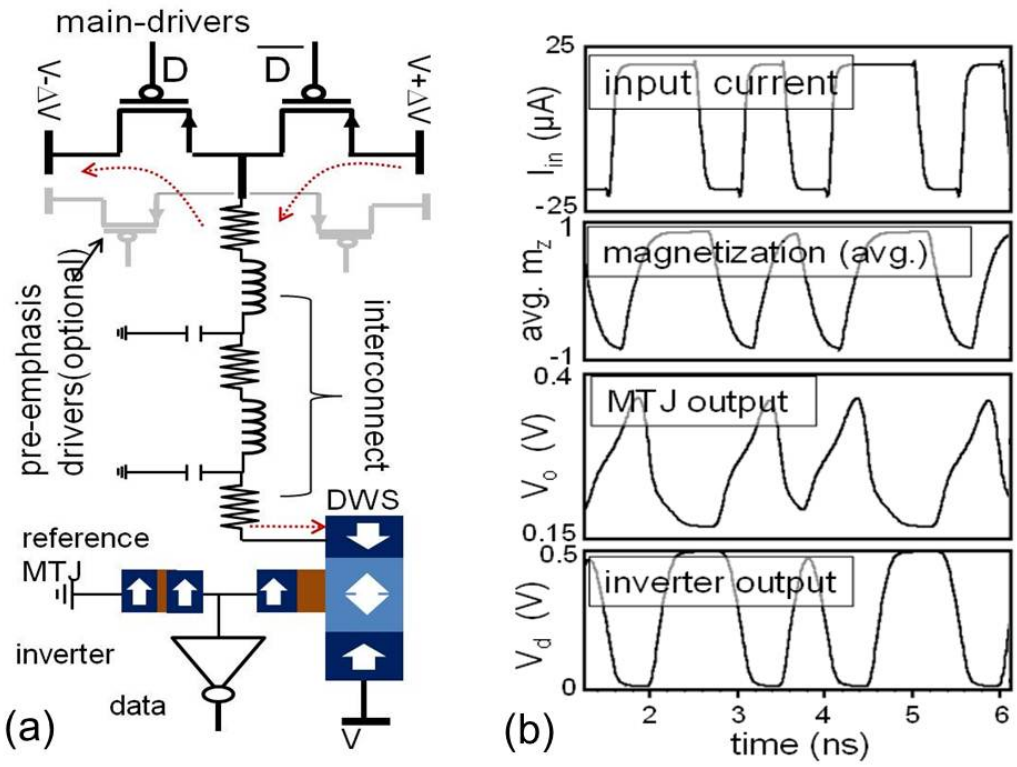


Figure 8.5 (a) Interconnect design using UDWS, (b) transient simulation plots for DWS-based interconnect at 2Gbps signaling-speed

optimized by the appropriate choice of signaling voltage ΔV and the driver-size (fig. 6a). For increasing driver size, dynamic switching power increases while, the required ΔV and hence, the static power reduces. For a given signaling speed, the signaling-current increases with interconnect-length due to frequency dependent attenuation of the signal. For longer data-link the total channel resistance increases, requiring further increase in signaling voltage and hence, the signaling energy (fig. 6b).

At the receiver-end, the MTJ associated with the DWS free-layerallows conversion of the spin-mode information into binary-on-chip voltage-levels through a

resistive voltage-divider that it forms with a reference MTJ. The ratio of parallel and anti-parallel spin-states of an MTJ is defined in terms of tunnel magneto-resistance ratio

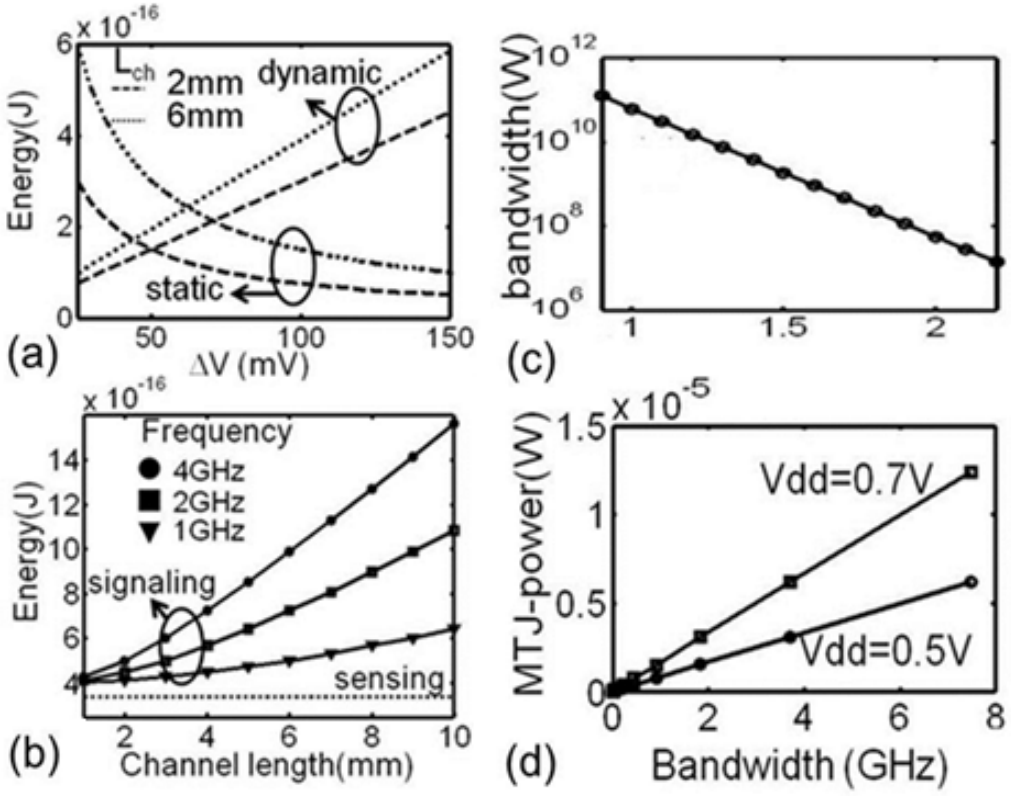


Figure 8.6(a) Dynamic and static energy components for the proposed interconnect vs. bandwidth of sensing node vs ΔV (minimum dynamic power corresponds to minimum size transistor in 45nm CMOS),(b)energy-dissipation as a function of channel length, (c) bandwidth of sensing node vs t_{ox}, (b) TMR vs. t_{ox}, (d) static-power in the MTJ vs. node bandwidth.

(TMR) [26]. A TMR of ~200% (corresponding to resistance ratio of ~4) can provide a voltage swing close to VDD/3 for at the voltage divider output that can be sensed by a minimum-size CMOS inverter (fig. 5a). Thus the DWS can acts as a high-gain (converting ~20μA of switching current-signal into digital voltage levels), ultra-low power, and compact trans-impedance amplifier (TIA) that can facilitate the design of energy-efficient current-mode global interconnects [102]. Simulation-waveforms for

MTJ-based transimpedance conversion are shown in fig. 5b. Depending upon the target signaling-frequency, the oxide thickness t_{ox} of the MTJ can be optimized for minimizing sensing-current and hence, the associated static-power (fig. 6c). With increasing $t_{ox,}$ the static current in the read-path reduces, but the bandwidth at the voltage-divider output reduces (fig. 6d). Note that decoupled read-write paths in the DWS allow high values of t_{ox} that can help achieve high-TMR along with low sensing-power, without sacrificing write-energy [102].

8.4 Performance and Prospects

In the proposed interconnect-design, the energy consumption per-bit transmitted can be evaluated as the sum of the components related to static power dissipation across the transmission line E_{int}, and the components resulting from the power consumption in the conversion circuit E_{conv}, at the receiver. The dynamic switching power for the small-size digital driver at the transmitter can be negligibly small as compared to the aforementioned components.

The DWS facilitates ultra-low voltage biasing of the entire transmission link, such that the static current flows across a small terminal voltage of 2ΔV. A 10 mm long on-chip interconnect (parameters given in [151]) would offer a resistance of ~500Ω and an effective capacitance of ~2.5pF. Transmission of data at 2Gbps (data-period T_d =0.5 ns) speed over such a link may require a current-amplitude (I_d) of ~20μA in order to be able to switch the DWS. This current magnitude can be supplied by minimum size driving transistors (with effective resistance of ~1kΩ in 45nm CMOS technology) with a ΔV of ~30mV. The component E_{int}can be therefore calculated as E_{int} = 2ΔV x I_d x$T_{d,}$ which evaluates to ~0.6fJ. The power consumption in the detection unit can be minimized with

the optimal choice of t_{ox} as discussed earlier. A TMR of 200% was used for the MTJs. For 2Gbps operation, the power consumption in the optimized detection circuit was

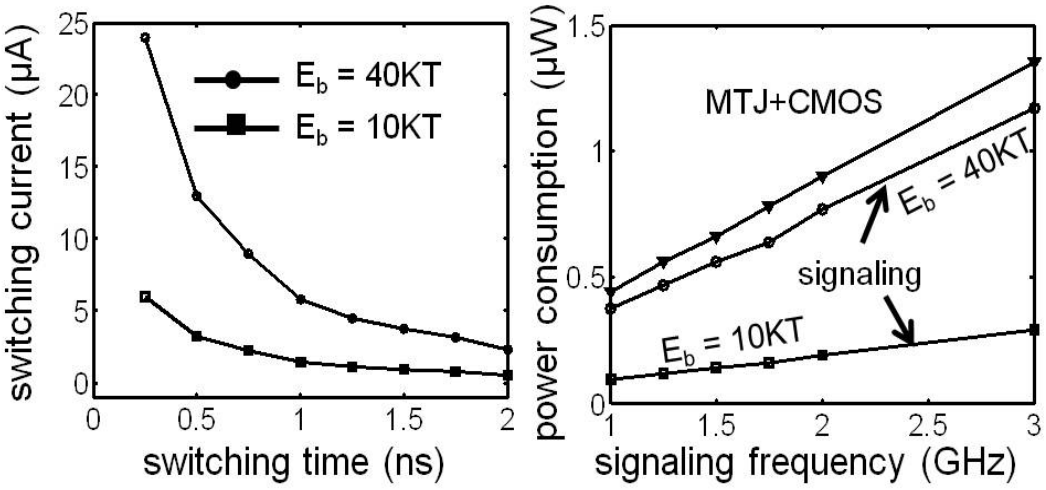


Figure 8.7(a) Switching time vs. switching current for two different anisotropy barriers, (b) power consumption in current-mode signaling and in driver and receiver circuits (including MTJs) increases linearly with signaling frequency, signaling (involving DWS switching) accounts for smaller power consumption as compared to driver and receiver circuits for wide range of DWS energy barrier (E_b). The oxide thickness of MTJ has been reduced for increasing frequency, in order to allow faster sensing.

found to be ~0.8μW. This translates to a value of ~0.4fJ for E_{conv}. Thus, the overall energy dissipated per-bit can be ~1fJ which is around two order of magnitude less than that reported in a recent mixed-signal CMOS implementation [102].

As mentioned earlier, DWS switching current and hence the signaling power can be reduced by using lower anisotropy barrier (fig. 7) or by employing a device structure with spin-orbit coupling. As mentioned earlier, the later method can also be conducive to high domain wall velocities of the order of ~1000m/s [103]. Such device -optimizations may facilitate more than 10GHz signaling with less than 100μA current, provided simultaneous requirements of device scaling and reduced current density (as

demonstrated for relatively larger devices) are achieved.

The proposed spin-CMOS hybrid interconnect can be compact and area-efficient as compared to conventional mixed signal CMOS current-mode I/O interfaces. Thus the spin-torque based I/O interfaces can emerge as a very attractive solution to the design challenges associated with on-chip and inter-chip interconnects. Other spin-torque switches can also be employed in the proposed scheme. In the following sections we introduce alternate spin-torque device structures that can be suitable for the design of current-mode interconnects

8.5 Alternate spin devices for interconnect design

8.5.1 Interconnect design using Bipolar Domain Wall switch

A 3-terminal, bipolar domain wall switch (BDWS) is shown in fig. 8a.Our proposed device consists of two fixed-domains of opposite magnetization (domain-2 and domain-3) that act as micromagnetic simulation plots for the BDWS at three-time stepsinput-ports and to polarize the input currents. The third domain (domain-1) is a free-domain. The spin-polarity of the current injected into the free-domain is the difference between the current inputs I_1 and I_2entering through the two inputs. The free-domain can switch parallel to either of the two fixed input domains depending on which of the two inputs currents is larger and hence, this device acts as a current-comparator. The minimum difference between the two inputs the BDWS can detect depends on the critical current density for domain-wall shift in the free-domain. A difference of few micro-amperes may be detected using a 15x2 nm^2 domain cross-section, with a critical current density of the order of 10^6 A/cm^2. Micro-magnetic simulation results for two inputs of 5μA and 10μA are given in fig. 8c. Spin-orbital coupling can be applied to the free-layer for achieving

enhanced DW-motion and hence higher switching-speed in the free-domain. The state of free-domain (domain-3) is read through the MTJ formed at its top.

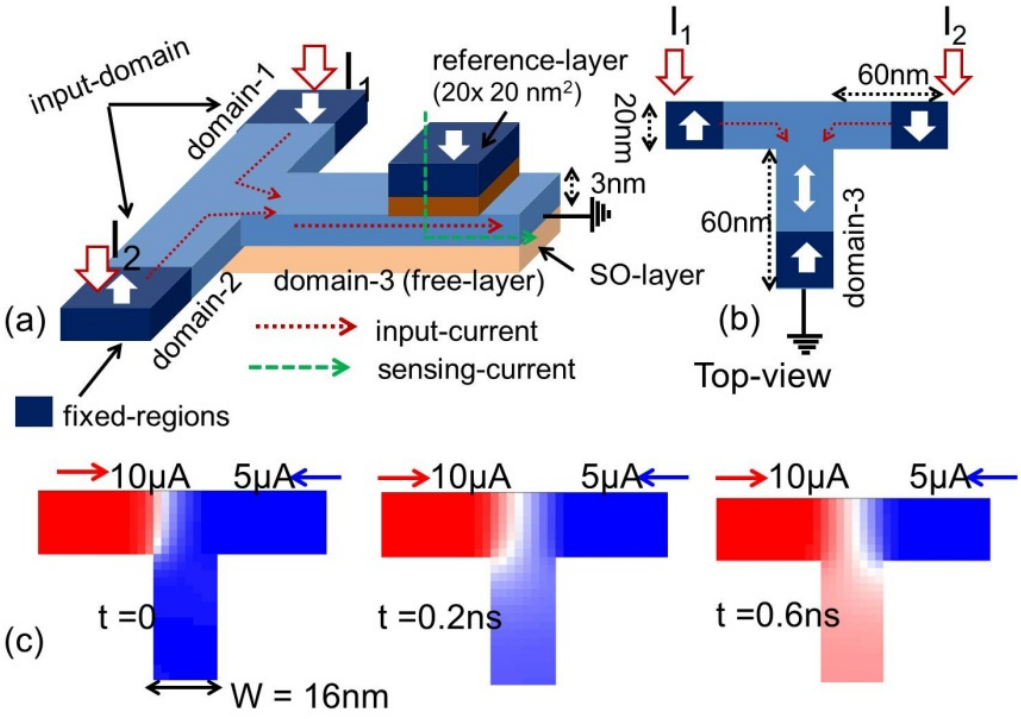


Figure 8.8(a) BDWS based on domain-wall-switching, with a possible spin-orbital coupling (SO) coupling applied to the free layer (b) top-view of the device, (c) micro-magnetic simulation plots for the bipolar DWS.

Fig. 9 depicts the circuit for a current-mode data interconnect employing an STS-based receiver. At the transmitter end, a linear region PMOS transistor *M1* is driven by a voltage-mode data-signal. Its source terminal is connected to a DC-voltage $V+\Delta V$, where V is 0.5V and ΔV can be less than ~50mV. On the receiver side, the DWS is biased at a voltage V, as shown in the figure. A bias transistor, *M2,* on the receiver-end injects a constant DC current (with half the amplitude of the input signal) into one of the two inputs of the DWS, which gets subtracted from the data-signal entering into the other input. This results in data-dependent flipping of the DWS free-domain. The received data

can be detected using a high-resistance voltage divider formed between the SWS-MTJ and a reference-MTJ, as show in fig. 9. A high TMR for the MTJ can

provide a voltage-swing large enough to be sensed by a simple CMOS inverter.

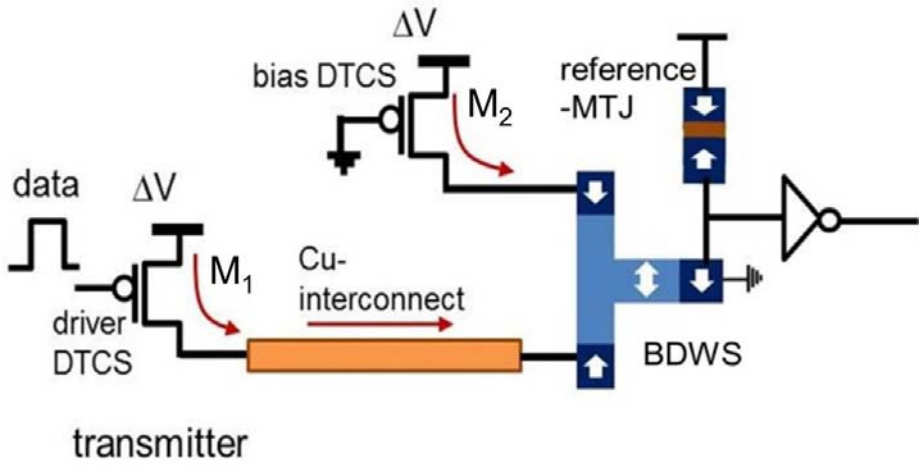


Figure 8.9Circuit for on-chip and inter-chip interconnect using BDWS; DTCS width ~0.2

8.5.2 Interconnect using switches based on Lateral Spin Valve

Recently high-speed switching of nano-magnets in spin-valves (SV) has been demonstrated [160, 9]. High-speed magnetization switching can be obtained with the help of combined in-plane and out-of plane spin-torque. This phenomenon may involve current-injection through a fixed-layer with easy-axis orthogonal to that of the free-layer. Such a fixed-layer injects orthogonal-spins into the free-layer that lower the effective energy-barrier for switching. The use of this phenomenon in lateral spin valve (LSV) was proposed in [9] to implement current-mode Bennett-Clocking (CBC). Although, both, unipolar as well as bipolar device models for LSV switches may be employed [26], in this chapter we limit our discussion to the later.

Fig. 10a shows the device structure for bipolar spin neuron. It constitutes of an output magnet $m_1$with MTJ based read-port (using a reference magnet m_5), and two anti-parallel input magnets m_2 andm_3,with their'easy-axis' parallel to that of m_1. A preset-

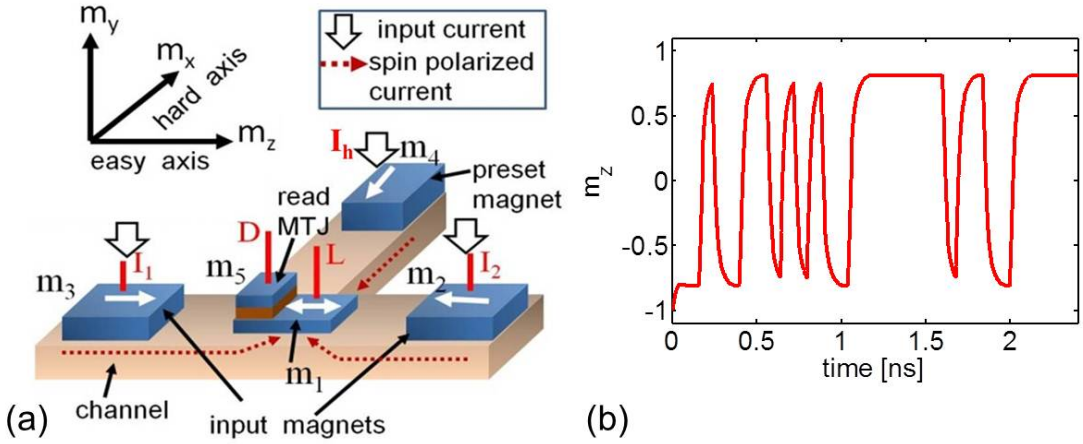


Figure 8.10(a) Bipolar LSV-switch for high-speed interconnect design, (b)10Gps Switching of Bipolar LSV switch with I_h = 100μA, I_1-I_2 = {0, 50μA), free-layer size: 30x15x1nm^3.

magnetm_4, with an orthogonal easy-axis, is used to implement current-mode Bennett-clocking (CBC). A current pulse input through m_4, forces the output magnet, m_1, along its hard-axis. The preset is overlapped with the input current pulses received through the magnets m_2 and m_3. In presence of sufficiently strong input current-pulses, $m_1$switches back to its easy-axis depending upon the input. The spin-polarity of $m_1$under a given current input depends upon the sign of the difference ΔI, between the current inputs through $m_2$and m_3. In order to operate the BLSV switch as an interconnect device, one of the two complementary inputs can be fixed to reference value, whereas the other input can receive the data signal with two levels, above and below the reference input. The data-dependent spin-state of the output magnet can be detected using a magnetic tunnel junction formed at its top. Fig. 3b shows the transient simulation plots for a BLSV

switch with a free-layer size of 30x15x1 nm^3, for 10Gbps input data. High-injection efficiency for the input interface has been assumed (~90%).

8.6 Summary

In this work we proposed to explore a novel technology solution for on-chip and inter-chip interconnect design using spin-torque switches. Magneto-metallic spin-torque switches act as ideal, low-impedance current-mode receivers, allow ultra-low-voltage biasing of the I/O interconnect and facilitate easy conversion from spin to charge using an MTJ interface. As a result, the proposed technique for high-speed current-mode interconnect design can be highly compact and more than two orders of magnitude energy-efficient and, as compared to state of the art technology solutions for on-chip (global) and inter-chip data links. The proposed technique can provide an attractive technology solution to the inter-connect bottleneck faced by high performance computing systems. In future we plan to do more rigorous analysis and comparison of different spin-device device models used in this work for interconnect design. This will include, analysis for scalability, reliability and variation tolerance. Accurate modeling of interconnects will also be required to estimate the advantages of the proposed schemes more accurately.

9. CONLCUSION AND FUTURE WORK

9.1 Conclusion

Current-induced spin-torque switching of nano-magnet, as a phenomenon, is widely accepted to be very useful for on chip-memory memory design. However, benefits of spin-torque devices for computational hardware are still being explored. Several device-models and circuit-design techniques have been proposed for applying spin-torque devices like spin-valves and domain-wall-magnets in computational hardware. However, most of them have been focused on digital logic. Ultra-low voltage, current-mode switching of magneto-metallic spin-torque devices can potentially be more suitable for non-Boolean computation schemes that can exploit current-mode analog-processing. Such schemes may not essentially be projected as drop-in replacement of CMOS. But such techniques can certainly be attractive for enhancing the functionality of CMOS by assisting it in tasks where is does not fare well. As a part of our work we proposed device models for 'spin-neurons' that can act as the fundamental building blocks of such non-Boolean computing blocks. Device circuit co-design for different classes of non-Boolean-architectures using spin-torque based neuron models in spin-CMOS hybrid circuits show that the spin-based non-Boolean designs can achieve large energy savings for generic

Computing applications like, image-processing, data-conversion, cognitive-computing, pattern matching and programmable-logic and global interconnect circuits as compared to state of art CMOS designs.

9.2 Future Work

Following tasks have been planned to be accomplished as future work:

9.2.1 Modeling and analysis spin-torque based clocking latches :

Emerging spin-torque (ST) phenomena may lead to ultra-low-voltage, high-speed nano-magnetic switches. Such current-based-switches can be attractive for designing low-swing global-interconnects like clocking-networks. We propose the design of such interconnects using functionality-enhanced ST-switches. For clocking-networks, Spin-Hall-Effect (SHE) can be used to produce an assist-field for fast ST-switching using global-mesh-clock with less than 100mV swing. The ST-switch acts as a compact-latch, written by ultra-low-voltage input-pulses. The data is read using a high-resistance tunnel-junction. Owing to low-voltage, current-mode operation, the proposed scheme can achieve low-power for clocking.

9.2.2 Modeling and analysis of spin-torque based current sensor for MRAM:

Spin-Transfer-Torque Magnetic Random Access Memory (STT-MRAM) is a promising candidate for future on-chip memory, owing to its high-density, zero-leakage and energy efficiency. In a conventional STT-MRAM cache write operations consume larger energy as compared to read, due to relatively large write-current requirement. In recent years

novel spin-torque based write schemes have been proposed for MRAM that can bring large reduction in write energy, such that the read-energy now becomes dominant. Conventional read schemes based on CMOS sense amplifiers may not offer commensurate reduction in read energy, owing to their poor scalability and limited speed. We propose a spin-torque based sensing technique for MRAM that employs nano-scale spin-torque switches for low-voltage, low current read-operations in STT-MRAM. Such a sensing-scheme can achieve improved-scalability, simplified-design for read peripherals, high-speed read-operations and 90% lower read energy. As a result more than ~80% reduction in overall energy can be obtained for STT-MRAM based caches.

9.2.3 Spiking Neural Network (SNN) for cognitive computing:

With special circuit techniques the spin-memristor crossbar design can be used to model bio-mimicking spiking neural networks [90]. We expect to achieve more than three orders of magnitude reduction in computation energy due to low power operation of the spin neurons.

9.2.4 Exploring on-chip (global) interconnect topologies for spin-based design:

The proposed spin-torque switches can be applicable to ultra low energy interconnect design for long distance on-chip and inter-chip signaling. However, appropriate interconnect topologies need to be explored that can maximally leverage the spin-torque switches. More rigorous device as well as circuit level analysis will also be performed along with the topology/system level design-exploration.

9.2.5 Physics based modeling of Memristors :

As discussed in the last chapter, developing physics based device model for memristor is essential to assess some important performance metrics for resistive crossbar based designs. We plan to benchmark physical models for some chosen memristive devices with the corresponding experimental data. Following this, compact device models will be generated form the calibrated physical model to facilitate large scale circuit level simulations.

9.2.6 Exploring the feasibility of ultra-low voltage supply distribution for the proposed spin-based hybrid computing scheme:

Some of the applications explored so far and of those to be explored in future, may not require very precise voltage levels. For instance simple image processing applications like edge detection half-toning etc, and,SNN-based cognitive computing circuits may be able to operate with noisy supply. However others like ADC, mixed-mode computing blocks (like filters) and threshold logic would require precise supply generation. We plan to explore two specific techniques towards this end. First, modeling of power supply grid with dedicated on-chip voltage regulators will be considered and the resulting energy and area overhead will be estimated. Second, we plan to explore the use of trench capacitor for decoupling power supplies in order reduce the requirement of excessive regulation. Trench capacitors are fully CMOS compatible and their application as decoupling capacitors has been proposed earlier [89]. Trench capacitors can achieve more than two order of magnitude higher capacitance per unit area. Hence their use can be conducive to lower supply noise.

LIST OF REFERENCES

LIST OF REFERENCES

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

[2] M. T. Alam, M. J. Siddiq, G. H. Bernstein, M. Niemier, W. Porod, and X. S. Hu, "On-chip clocking for nanomagnet logic devices," *IEEE Trans. Nanotech.* vol. 9, no. 3, pp. 348-351, May 2010.

[3] T. Kimura, Y. Otani, and J. Hamrle., "Switching magnetization of a nanoscale ferromagnetic particle using nonlocal spin injection," *Phys. Rev. Lett.*, vol. 96, iss. 3, pp. 037201-1-037201-4, Jan. 2006.

[4] J. Z. Sun, M. C. Gaidis, E. J. O'Sullivan, E. A. Joseph, G. Hu, D. W. Abraham, J. J. Nowak, P. L. Trouilloud, Yu Lu, S. L. Brown, D. C. Worledge, and W. J. Gallagher, "A three-terminal spin-torque-driven magnetic switch", *Appl. Phys. Lett.* vol. 95, iss. 8, pp. 083506-1-083506-3, Aug. 2009.

[5] M. Yamanouchi, D. Chiba, F. Matsukura, T. Dietl, and H. Ohno, "Velocity of domain-wall motion induced by electrical current in the ferromagnetic semiconductor," *Phys. Rev. Lett.* vol.96, iss. 9, pp. 096601-1-096601-4, Mar. 2006.

[6] D. Chiba, G. Yamada, T. Koyama, K. Ueda, H. Tanigawa, S. Fukami, T. Suzuki, N. Ohshima, N. Ishiwata, Y. Nakatani, and T. Ono, "Control of multiple magnetic domain walls by current in a Co/Ni nano-wire," *Appl. Phys. Express* vol. 3, pp. 073004-1-073004-3, Jul. 2010.

[7] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano, "A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram," in *Proc. of IEEE Int. Electron Devices Meeting (IEDM)*, pg. 459-462, Dec. 2005.

[8] F. M. Spedalieri, A. P. Jacob, D. E. Nikonov, and V. P. Roychowdhury, "Performance of magnetic quantum cellular automata and limitations due to thermal noise," *IEEE Trans. Nanotech.* vol.10, iss. 3, pp. 537-546, May 2011.

[9] B. Behin-Aein, D. Datta, S. Salahuddin, and S. Datta, "Proposal for an all-spin logic device with built-in memory," *Nature Nanotechnology* vol. 5, pp. 266-270, Feb. 2010.

[10] S. Srinivasan, A. Sarkar, B. Behin-Aein, and S. Datta, "All-spin logic device with inbuilt nonreciprocity," *IEEE. Trans. Magnetics* vol. 47, iss. 10, pp. 4026-4032, Oct. 2011.

[11] 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. 123510-1-123510-3, Mar. 2011.

[12] 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 2011 IEEE/ACM Int. Symp. on Nanoscale Architectures (NANOARCH)*, pp. 129-136, Jun. 2011.

[13] 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," *Appl. Phys. Lett.* vol. 84, iss. 15, pp. 2820-1-2820-3, Feb. 2004.

[14] D.-T. Ngo, K. Ikeda, and H. Awano, "Direct observation of domain wall motion induced by low-current density in TbFeCo wires," *Appl. Phys. Express* vol. 4, pp. 093002-1-093002-3, Aug. 2011.

[15] S. Fukami, T. Suzuki, K. Nagahara, N. Ohshima, Y. Ozaki, S. Saito, R. Nebashi, N. Sakimura, H. Honjo, K. Mori, C. Igarashi, S. Miura, N. Ishiwata, and T. Sugibayashi, "Low-current perpendicular domain wall motion cell for scalable high-speed MRAM," in *Proc. of Symp. on VLSI Tech.*, pp. 230-231, Jun 2009.[16] J. A. Currivan, J. Youngman,

[16] M. D. Mascaro, M. A. Baldo, and C. A. Ross, "Low energy magnetic domain wall logic in short, narrow, ferromagnetic wires," *IEEE Mag. Lett.* vol. 3, pp. 3000104-1-3000104-4, Apr. 2012.

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

[18] D. Morris, D. Bromberg, J. Zhu, and L. Pileggi, "mLogic: ultra-low voltage non-volatile logic circuits using STT-MTJ devices," in *Proc. of ACM/IEEE Design Automation Conference (DAC)*, pp. 486-492, Jun. 2012.

[19] J. Liu, and M. Brooke, "Fully parallel on-chip learning hardware neural network for real-time control," in *Proc. of IEEE Int. Symp. on Circuits and Sys. (ISCAS)* vol.5, pp. V-371-V-374, Jun. 1999.

[20] R. Dlugosz, T. Talaska, and W. Pedrycz, "Current-mode analog adaptive mechanism for ultra-low-power neural networks," *IEEE Trans. Circuits and Systems II: Express Briefs* vol.58, iss. 1, pp. 31-35, Jan. 2011.

[21] Bermak et al., "A highly scalable 3D chip for binary neural network classification applications," in *Proc. of IEEE Int. Symp. on Circuits and Sys. (ISCAS)* vol. 5, pp. V-685-V-688, May 2003.

[22] J. Burr, "Digital neural network implementations," in *Neural Networks, Concepts, Applications, and Implementations*. Prentice-Hall, 1991, vol. III.

[23] J. M. Cruz, and L. O. Chua, "A 16x16 cellular neural network universal chip: the first complete single-chip dynamic computer array with distributed memory and with gray-scale input-output," *Analog Integrated Circuits and Signal Processing* vol. 15, no. 3, pp. 227-237, Mar. 1998.

[24] P. Dudek, and P. J. Hicks, "A general-purpose processor-per-pixel analog SIMD vision chip," *IEEE Trans. Circuits and Sys. I: Reg. Papers* vol. 52, iss. 1, pp. 13-20, Jan. 2005.

[25] J. Li, P. Ndai, A. Goel, S. Salahuddin, and K. Roy, "Design paradigm for robust spin-torque transfer magnetic RAM (STT MRAM) from circuit/architecture perspective," *IEEE Trans. on VLSI Sys.* vol. 18, iss. 12, pp. 1710 - 1723, Dec. 2010.

[26] M. Sharad, C. Augustine, and K. Roy, "Boolean and non-boolean computation with spin devices," in *Proc. of IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2012.

[27] M. Sharad, G. Panagopoulos, C. Augustine, and K. Roy, "Spin-based neuron model with domain wall magnets as synapse,"*IEEE Trans. Nanotech.* vol. 11, no. 4, pp. 843-853, Jul. 2012.

[28] M. Sharad, C. Augustine, G. Panagopoulos, and K. Roy, "Cognitive computing with spin based neural networks," in *Proc. of ACM/IEEE Design Automation Conference (DAC)*, pp. 1262-1263, Jun. 2012.

[29] M. Sharad, G. Panagopoulos, and K. Roy, "Spin neurons for ultra low power computational hardware," in *Proc. of IEEE Device Research Conf. (DRC)*, pp. 221-222, Jun. 2012.

[30] 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 *Proc. of IEEE Int. Conf. On Simulation of Semiconductor Processes and Devices (SISPAD)*, pp. 51-54, Sep. 2011.

[31] C. Augustine, A. Raychowdhury, D. Somasekhar, J. Tschanz, K. Roy, and V. K. De, "Numerical analysis of typical STT-MTJ stacks for 1T-1R memory arrays," in *Proc. of IEEE Int. Electron Devices Meeting (IEDM)*, pp. 22.7.1-22.7.4, Dec. 2010.

[32] C. Augustine, N. N. Mojumder, X. Fong, S. H. Choday, S. P. Park, and K. Roy, "Spin-transfer torque MRAMs for low power memories: perspective and prospective," *IEEE Sensors J.* vol. 12, iss. 4, pp. 756-766, Apr. 2012.

[33] M. Sharad, D. Fan, and K. Roy, "Design of ultra high density and low power computation blocks using nano-magnets," to appear in *IEEE Int. Symp. on Quality Electronic Design (ISQED)*, 2013.

[34] R. Venkatesan, M. Sharad, K. Roy, and A. Raghunathan, "DWM-TAPESTRI - an energy efficient cache design using spin memory with domain wall shift based writes," to appear in *IEEE Conf. on Design Automation & Test in Europe (DATE*), 2013.

[35] M. Sharad, G. Panagopoulos, C. Augustine, and K. Roy, "*NLSTT-MRAM: robust spin transfer torque MRAM using non-local spin injection for write," in Proc. of IEEE Device Research Conf. (DRC), pp. 97-98, Jun. 2012.*

[36] C. Augustine, B. Behin-Aein, X. Fong, K. Roy, "A design methodology and device/circuit/architecture compatible simulation framework for low-power magnetic quantum cellular automata systems," in *Proc. of ACM/IEEE Asia & South Pacific Design Automation Conf.(ASP-DAC)*, pp. 847-852, Jan. 2009.

[37] J. Li, C. Augustine, S. Salahuddin, and K. Roy, "Modeling of failure probability and statistical design of spin-torque transfer magnetic random access memory (STT MRAM) array for yield enhancement," in *Proc. of ACM/IEEE Design Automation Conf. (DAC)*, pp. 278-283, Jun. 2008.

[38] S. Shin, K. Kim, and S.-M. Kang, "Memristor-based fine resolution programmable resistance and its applications," in *Proc. ofIEEE Int. Conf. on Comm., Circuits and Systems (ICCCAS),* pg. 948-951, Jul. 2009.

[39] R. Berdan, T. Prodromakis, and C. Toumazou, "High precision analogue memristor state tuning," *Electronics Lett.* vol. 48, iss. 18, pp. 1105-1107, Aug. 2012.

[40] F. Miao, W. Yi, I. Goldfarb, J. Joshua Yang, M. Zhang, M. Pickett, J. Strachan, G. Medeiros-Ribeiro, and R. Williams, "Continuous electrical tuning of the chemical composition of TaOx-based memristors," *ACS Nano* vol. 6, no. 3, pp 2312–2318, Feb. 2012.

[41] K. K. Likharev, "Biologically inspired computing in CMOL CrossNets," in *Proc. of AAAI Fall Symp.*, pp. 90, Oct. 2009.

[42] O. Turel, J. H. Lee, X. Ma, and K. K. Likharev, "Neuromorphic architectures for nanoelectronic circuits," *Int. J. Circ. Theory and Appl.* vol. 32, iss. 5, pp. 277–302, Sep. 2004.

[43] S. H. Jo, K.-H. Kim, and W. Lu, "High-density crossbar arrays based on a Si memristive system", *ACS NanoLett.* vol. 9, iss. 2, pp. 870–874, Jan. 2009.

[44] S. H. Jo, and W. Lu, "CMOS compatible nanoscale nonvolatile resistance switching memory," *ACS NanoLett.* vol. 8, no. 2, pp. 392–397, Jan. 2008.

[45] L. Gao, F. Alibart, and D. B. Strukov, "Analog-input analog-weight dot-product operation with Ag/a-Si/Pt memristive devices," in *Proc. of VLSI-SoC*, Oct. 2012.

[46] B. Mouttet, "Proposal for memristors in signal processing," *Nano-Net* vol. 3, pp. 11-13, 2009.

[47] L. Hai, W. Qing, and G. S. Rose, "Hardware realization of BSB recall function using memristor crossbar arrays," in *Proc. of ACM/IEEE Design Automation Conf. (DAC)*, pp. 498-503, Jun. 2012.

[48] L. O. Chua, and L. Yang, "Cellular neural networks: applications" *IEEE Trans. Circuits and Sys.* vol. 35, iss. 10, pp. 1237-1290, Oct. 1988.

[49] H. Harrer, and J. A. Nossek, "Discrete-time cellular neural networks," *Int. J. of Circuit Theory and App.* vol. 20, iss. 5, pp. 453-467, Sep. 1992.

[50] W. Miao, Q. Lin, W. Zhang, and N. Wu, "A programmable SIMD vision chip for real-time vision applications," *IEEE J. Solid-State Circuits* vol. 46, iss. 6, pp. 1470-1479, Jun. 2008.

[51] T. Komuro, I. Ishii, M. Ishikawa, and A. Yoshida, "A digital vision chip specialized for high-speed target tracking," *IEEE Trans. Electron Devices*, vol. 50, iss. 1, pp. 191–199, Jan. 2003.

[52] R. Dominguez-Castro, S. Espejo, A. Rodriguez-Vazquez, R. A. Carmona, P. Foldesy, A. Zarandy, P. Szolgay, T. Sziranyi, and T. Roska, "A 0.8-µm CMOS two-dimensional programmable mixed-signal focal-plane array processor with on-chip binary imaging and instructions storage," *IEEE J. Solid-State Circuits* vol. 22, iss. 7, pp. 1013-1026, Jul. 1997.

[53] T. Kumoro, S. Kagami, and M. Ishikawa, "A dynamically reconfigurable SIMD processor for a vision chip," *IEEE J. Solid-State Circuits* vol. 39, iss. 1, pp. 265-268, Jan. 2004.

[54] El Gamal, and H. Eltoukhy, "CMOS image sensors," *IEEE Circuits and Systems Magazine* vol. 21, iss. 3, pp. 6-20, May. 2005.

[55] Zarandy, and C. Rekeczky, "Bi-i: a standalone ultra high speed cellular vision system," *IEEE Circuits and Systems Magazine* vol. 5, iss. 2, pp. 36-45, 2005.

[56] Rodriguez-Vazquez, G. Linan-Cembrano, L. Carranza, E. Roca-Moreno, R. Carmona-Galan, F. Jimenez-Garrido, R. Dominguez-Castro, and S. E. Meana, "ACE16k: the third generation of mixed-signal SIMD-CNN ACE chips toward VSoCs," *IEEE Trans. Circuits and Sys.* vol. 51, iss. 5, pp. 851-863, May 2004.

[57] Durpet, J. O. Klein, and A. Nshare, "A programmable vision chip for CNN based algorithms," in *Proc. of IEEE Int. Workshop on Cellular Neural Networks and their App. (CNNA)*, pp. 207-212, May 2000.

[58] S. S. Rajput, and S. S. Jamuar, "Low voltage analog circuit design techniques," *IEEE Circuits and Sys. Magazine* vol. 2, iss. 1, pp. 24-42, 2002.

[59] A.J. Annema, "Analog circuit performance and process scaling," *IEEE Trans on Circuits and Sys.-II: Analog and Digital Signal Processing* vol. 46, iss. 6, pp. 711-725, Jun. 1999.

[60] M. van Elzakker, E. van Tuijl, P. Geraedts, D. Schinkel, E. Klumperink, and B. Nauta, "A 1.9W 4.4fJ/conversion-step 10b 1MS/s charge-redistribution ADC," in *Proc. of IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pp. 244-610, Jan. 2008.

[61] S. Kawahito, J. Park, K. Isobe, S. Shafie, T. Lida, and T. Mizota, "A CMOS Image Sensor Integrating Column-Parallel Cyclic ADCs with On-Chip Digital Error Correction Circuits," in *Proc. of IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pp. 56-595, Jan. 2008.

[62] D. Biolek, R. Senani, V. Biolkova, and Z. Kolka "Active elements for analog signal processing: classification, review, and new proposals," *Radioengineering* vol. 17, no. 4, pp. 15-32, Dec. 2008.

[63] P. Kignet, and M. Steyaer, "An analog parallel array processor for real-time sensor signal processing," in *Proc. of IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pp. 92-93, Feb. 1996.

[64] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons: 2000.

[65] Brataas, G. E. W. Bauer and P. J. Kelly, "Non-collinear magnetoelectronics," *Phys. Rep.* vol. 427, iss. 4, pp. 157-255, Apr. 2006.

[66] T. Valet, and A. Fert, "Theory of perpendicular magnetoresistance in magnetic multilayers," *Phys. Rev. B* vol. 48, iss. 10, pp. 7099-7113, Sep. 1993.

[67] D. B. Strukov, J. L. Borghetti, and R. S. Williams, "Coupled ionic and electronic transport model of thin-film semiconductor memristivebehaviour," *Small* vol. 5, iss. 9, pp. 1058-1063, May 2009.

[68] D. Nikonov and I. Young, "Uniform Methodology for Benchmarking Beyond-CMOS Logic Devices," *International Electron Devices Meeting*, December 2012.

[69] M. Suri, O. Bichler, D. Querlioz, B. Traoré, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, B. DeSalvo, "Physical aspects of low power synapses based on phase change memory devices," J. Appl. Phys. **112**, 054904 (2012), 10 pages.

[70] D. Kuzum, R. Jeyasingh, B. Lee, and H.-S. Philip Wong, "Nanoelectronic Programmable Synapses Based on Phase Change Materials for Brain-Inspired Computing," *ACS Nano Letters*, 2011.

[71] Yamaguchi, . Ono, S. Nasu, K. Miyake, K. Mibu, T. Shinjo, "Real-Space Observation of Current-Driven Domain Wall Motion in Submicron Magnetic Wires," *Phys. Rev. Lett.* 92, 077205 (2004), 4 pages.

[72] Y. Bengio, ``Learning Deep Architectures for AI," Foundations and Trends in Machine Learning, vol. 2(1), pp. 1-127, Jan. 2009.

[73] K. Fukushima, ``Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,"*Biological Cybernetics*, 36(4), pp. 93-202, 1980.

[74] D. H. Hubel and T. N. Wiesel, ``Brain mechanisms of vision," *Scientific American*, 241(3), 1979.

[75] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, ``Gradient-Based Learning Applied to Document Recognition,” *Proceedings of the IEEE*, 86(11), pp. 2278-2324, Nov. 1998.

[76] G. E. Hinton, S. Osindero, and Y.-W. The, ``A fast learning algorithm for deep belief nets,” *Neural Computation*, 18(7), pp. 1527-1554, July 2006.

[77] Y. Bengio and Y. LeCun, ``Scaling learning algorithms towards AI,” in L. Bottou, O. Chapelle, D. DeCoste, and J. Weston (Eds), Large-Scale Kernel Machines, MIT Press, 2007.

[78] Q. Le, M. A. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, A. Y. Ng., ``Building high-level features using large scale unsupervised learning,”, *Int. Conf. on Machine Learning*, 2012.

[79] C. Farabet, C. Poulet, J. Y. Han, and Y. LeCun, ``CNP: An FPGA-based Processor for Convolutional Networks,” in *Proc. International Conference on Field Programmable Logic and Applications*, 2009.

[80] S. Chakradhar, M. Sankaradas, V. Jakkula, and S. Cadambi, ``A dynamically configurable coprocessor for convolutional neural networks,” *Prof. International Symposium on Computer Architecture*, pp. 247-257, 2010.

[81] Paul Merolla, John Arthur, FilippAkopyan, Nabil Imam, RajitManohar, Dharmendra S. Modha, "A Digital Neurosynaptic Core using Embedded Crossbar Memory with 45pJ per spike in 45nm," IEEE Custom Integrated Circuits Conference, pg. 1-4, 2011.

[82] Jae-sun Seo, Bernard Brezzo, Yong Liu, Benjamin D. Parker, Steven K. Esser, Robert K. Montoye, BipinRajendran, Jose A. Tierno, Leland Chang, Dharmendra S. Modha, and Daniel J. Friedman "A 45nm CMOS Neuromorphic Chip with a Scalable Architecture for Learning in Networks of Spiking Neurons," IEEE Custom Integrated Circuits Conference, pg. 1-4, 2011

[83] Bryan L. Jackson, BipinRajendran, Gregory S. Corrado, Matthew Breitwisch, Geoffrey W. Burr, Roger Cheek, KailashGopalakrishnan, Simone Raoux, Charles T. Rettner, Alex G. Schrott, Rohit S. Shenoy, Bulent N. Kurdi, Chung H. Lam, and Dharmendra S. Modha, "Nano-Scale Electronic Synapses using Phase Change Devices," JETC, 2011.

[84] K. Kim, S. Gaba, D. Wheeler, J. M. Cruz-Albrecht, T. Hussain, N. Srinivasa, W. Lu, "A Functional Hybrid Memristor Crossbar-Array/CMOS System for Data Storage and Neuromorphic Applications", NanoLett., vol. 12, iss. 1, pg. 389–395, 2012

[85] J. M. Cruz-Albrecht, M. W. Yung, N. Srinivasa, "Energy-Efficient Neuron, Synapse and STDP Integrated Circuits", IEEE Trans. on Biomedical Circuits and Sys., vol. 6, iss. 3, pg. 246-256, 2012

[86] G S Snider, "Self-Organized Computation With Unreliable, memristivenanodevices", IOP, Nanotechnology, vol. 18, 2007

[87] J. Rajendran et al., "Memristor based Programmable Threshold Logic Array", Nanoarch, 2013

[88] T. Tran et al., "Reconfigurable Threshold Logic Gates using Memristive Devices", IEEE Subthreshold Microelectronics Conference,2012

[89] F. Roozeboom et al., " ALD Options for Si-integrated Ultrahigh-density Decoupling Capacitors in Pore and Trench Designs", ECS Transactions, 3 (15) 173-181 (2007)

[90] Afifi, A., A. Ayatollahi, and F. Raissi. "Implementation of biologically plausible spiking neural network models on the memristor crossbar-based CMOS/nano circuits." *Circuit Theory and Design, 2009. ECCTD 2009. European Conference on*. IEEE, 2009.

[91] Sharad et al., "Spintronic Switches for Ultra Low Energy on Chip and Inter-chip Current-mode Interconnects", Submitted to Electron Device Letters.

[92] G. Csabaet. al, "Modeling of Coupled Spin Torque Oscillators for Application in Associative Memory", IEEE conf. on Nanotechnology, 2012

[93] M. Sharadet. al., "Dual Pillar Spin Torque Nano Oscillator", APL, 2013

[94] MR Pufallet. al., "Electrical Measurement of Spin-Wave Interactions of Proximate Spin Transfer Nanooscillators", Physical Review Letters, 2006

[95] Slavin, A. N., and V. S. Tiberkevich,*Physical Review B* 72.9 (2005): 092407.

[96] T. Moriyama, G. Finocchio, M. Carpentieri, B. Azzerboni, D. C. Ralph, R. A. Buhrman, *Physical Review B*, *86*(6), 060411, 2012

[97] J. Persson et. al., "Phase-Locked Spin Torque Oscillators: Impact of Device Variability and Time Delay", JAP, 2007

[98] Shibata et al., "CMOS supporting circuitries for nano-oscillator-based associative memories, CNNA 2012.

[99] M. Sharad et al., "Energy-Efficient and Robust Associative Computing with Injection-Locked Dual Pillar Spin-Torque Oscillators", Submitted to TMAG, 2014

[100] Owens et al., IEEE Micro*,*27.5, 96-108, 2007

[101] Tzartzanis et al., JSSC, *40.*11, 2141-2147, 2005

[102] Sharad et al., IEEE Electron Device Lett., 34.8, 1068-1070, 2013

[103] V. Khvalkovskiy et al., "Matching domain-wall configuration and spin-orbit torques for efficient domain-wall motion", Physical Review B, 87.2, 020402, 2013

[104] L. Gao et. al., "Analog-Input Analog-Weight Dot-Product Operation with Ag/a-Si/Pt Memristive Devices", VLSISOC, 88-93, 2012.

[105] H. Manem et al., "A Read-Monitored Write Circuit for 1T1M Multi-Level Memristor Memories", ISCAS, 2938-2941, 2012.

[106] C. Jung et al., "Two-Step Write Scheme for Reducing Sneak-Path Leakage in Complementary Memristor Array", IEEE Trans. on Nanotech, 1.3, 611-618, 2012.

[107] Sangho Shin et. al, "Memristor-Based Fine Resolution Programmable Resistance and Its Applications", ISCAS, pp. 948-951, 2009

[108] K. Kim et al., "A Functional Hybrid Memristor Crossbar-Array/CMOS System forData Storage and Neuromorphic Applications", Nano Letters, 12.1, 389-395, 2011

[109] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[110] http://cswww.essex.ac.uk/mv/allfaces/index.html

[111] R. DŁugosz et. al., " Low power current-mode binary-tree asynchronous Min/Max circuit", Microelectronics Journal, 41.1, 64-73., 2009.

[112] Demosthenous et. al., "A CMOS Analog Winner-Take-All Network for Large-Scale Applications" , IEEE Trans. on Fundamental Theory and App., 45(3), 300-304, 1998

[113] P. R. Kinget, "Device Mismatch and Tradeoffs in the Design of Analog Circuits" JSSC, 40.6, 1212-1224, 2005

[114] Zhao, Weisheng, et al. "High speed, high stability and low power sensing amplifier for MTJ/CMOS hybrid logic circuits." IEEE Transactions on Magnetics*,* 45.10, 3784-3787, 2009

[115] Saberkari et al., “Fast transient current-steering CMOS LDO regulator based on current feedback amplifier”, Integration, the VLSI Journal, *46.*2, 165-171, 2013

[116] Duda et al., "*Pattern Classification*; JohnWiley" & Sons: 2000.

[117] Shibata et al., "CMOS supporting circuitries for nano-oscillator-based associative memories, CNNA 2012

[118] S. G. Narendra and A. P. Chandrakasan., "Leakage in Nanometer CMOS Technologies," Springer 2005.

[119] The growing power density (measured in W/cm2) of Intel□s microchip processor families", Intel -2007.

[120] W. Abadeer, and W. Ellis, "Behavior of NBTI Under AC Dynamic Circuit Conditions", Proc. of International Reliability Physics Symposium, pp. 17-22, May, 2003.

[121] W. Wang, J. Tao and P. Fang, "Dependence of HCI mechanism on Temperature for 0.18μm technology and beyond", IEEE International Reliability Workshop, Final Report, pp. 66-68, October, 1999.

[122] R. Frankovic and G. H. Bernstein, "Temperature dependence of electromigration threshold in Cu", Appl. Phys. Lett., vol. 81, pp. 1604-1605, 1997.

[123] Bohr, Mark. "The evolution of scaling from the homogeneous era to the heterogeneous era." *Electron Devices Meeting (IEDM), 2011 IEEE International*. IEEE, 2011.

[124] Hutchby, J. A., Bourianoff, G. I., Zhirnov, V. V., & Brewer, J. E. (2002). Extending the road beyond CMOS. *Circuits and Devices Magazine, IEEE*, *18*(2), 28-41.

[125] Huai, Yiming. "Spin-transfer torque MRAM (STT-MRAM): Challenges and prospects." *AAPPS Bulletin* 18.6 (2008): 33-40.

[126] Wong, H. S., Raoux, S., Kim, S., Liang, J., Reifenberg, J. P., Rajendran, B., ... & Goodson, K. E. (2010). Phase change memory. *Proceedings of the IEEE*, *98*(12), 2201-2227.

[127] Emre Kultursay et al., **"**Performance enhancement under power constraints using heterogeneous CMOS-TFET multicores", CODES+ISSS, 2012

[128] Taubenblatt, M. A. (2012). Optical interconnects for high-performance computing. *Journal of Lightwave Technology*, *30*(4), 448-457.

[129] Chen, Xiangyu, et al. "Fully integrated graphene and carbon nanotube interconnects for gigahertz high-speed CMOS electronics." *Electron Devices, IEEE Transactions on* 57.11 (2010): 3137-3143127.

[130] Misra, Janardan, and IndranilSaha. "Artificial neural networks in hardware: A survey of two decades of progress." *Neurocomputing* 74.1 (2010): 239-255

[131] Chittaro, Luca. "Visualizing information on mobile devices." *Computer* 39.3 (2006): 40-45.

[132] Schemmel, Johannes, Johannes Fieres, and Karlheinz Meier. "Wafer-scale integration of analog neural networks." *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. IEEE, 2008.

[133] Wolf, Stuart A., et al. "The promise of nanomagnetics and spintronics for future logic and universal memory." *Proceedings of the IEEE* 98.12 (2010): 2155-2168.

[134] Apalkov, D. M., and P. B. Visscher. "Spin-torque switching: Fokker-Planck rate calculation." *Physical Review B* 72.18 (2005): 180405.

[135] Durlam, Mark, et al. "Nonvolatile RAM based on magnetic tunnel junction elements." *Solid-State Circuits Conference, 2000. Digest of Technical Papers. ISSCC. 2000 IEEE International*. IEEE, 2000.

[136] Grollier, Julie. "The Spin Torque Lego-from spin torque nano-devices to advanced computing architectures." *APS March Meeting Abstracts*. Vol. 1. 2013.

[137] Imre, A. et al. Majority logic gate for magnetic quantum-dot cellular automata Science 311, 205–208, 2006.

[138] Alam, M.T.; Siddiq, M.J.; Bernstein, G.H.; Niemier, M.; Porod, W.; Hu, X.S.; , "On-Chip Clocking for Nanomagnet Logic Devices,"

[139] N. D'Souza et. al., "Four-state nanomagnetic logic using multiferroics", IOP Science, 2011

[140] Zhao, Weisheng, et al. "New non-volatile logic based on spin-MTJ." *physica status solidi (a)* 205.6 (2008): 1373-1377.

[141] Ian Young , “Mapping a Path to the Beyond-CMOS Technology for Computation”, DRC, 2012

[142] Likharev, Konstantin, et al. "CrossNets: High-Performance Neuromorphic Architectures for CMOL Circuits." *Annals of the New York Academy of Sciences* 1006.1 (2003): 146-163.

[143] Bernstein, K., Cavin, R. K., Porod, W., Seabaugh, A., &Welser, J. (2010). Device and architecture outlook for beyond CMOS switches. *Proceedings of the IEEE*, *98*(12), 2169-2184.

[144] Yao, X., Harms, J., Lyle, A., Ebrahimi, F., Zhang, Y., & Wang, J. P. (2012). Magnetic tunnel junction-based spintronic logic units operated by spin transfer torque. *Nanotechnology, IEEE Transactions on*, *11*(1), 120-126.

[145] Tombros, N., Jozsa, C., Popinciuc, M., Jonkman, H. T., & Van Wees, B. J. (2007). Electronic spin transport and spin precession in single graphene layers at room temperature. *Nature*, *448*(7153), 571-574

[146] M. Sharadet. al., "Dual Pillar Spin Torque Nano Oscillator", APL, 2013

[147] B. Georges et. al., "Coupling Efficiency for Phase Locking of A Spin Transfer Nano-Oscillator to A Microwave Current", Phys. Rev. Lett., 2008

[148] T. Wada, T. Yamane, T. Seki, T. Nozaki, Y. Suzuki, H. Kubota, A. Fukushima, S. Yuasa, H. Maehara, Y. Nagamine, K. Tsunekawa, D. D. Dyayaprawira and N. Watanabe, Physical Review B, 81, 104410 , 2010.

[149] R. Dobkin et al.,"Parallel vs. Serial On-Chip Communication" SLIP, 2008

[150] F. Mahony et al., "A 47×10Gb/s 1.4mW/(Gb/s) Parallel Interface in 45nm CMOS", ISSCC, 2010.

[151] S. Lee et al., "A 95fJ/b Current-Mode Transceiver for 10mm On-Chip Interconnect", ISSCC, 2013

[152] B. Kim et al., "A 4gb/s/ch 356fj/b 10mm equalized on-chip interconnect with nonlinear charge-injecting transmit filter and transimpedance receiver in 90nm cmos." ISSCC, 2009.

[153] P. Pepeljugoski et al. "Low power and high density optical interconnects for future supercomputers." Optical Fiber Communication (OFC), 2010

[154] C. L. Schow et al., "Low-Power 16 x 10 Gb/s Bi-Directional Single Chip CMOS Optical Transceivers Operating at < 5 mW/Gb/s/link", JSSC, 2009

[155] R. Dokania rt al, "Analysis of Challenges for On-Chip Optical Interconnects", GLSVLSI, 2009

[156] C. K. Lim, "Domain wall displacement induced by subnanosecond pulsed current", App. Phy. Lett., 2004

[157] Ngo et al., " Direct Observation of Domain Wall Motion Induced by Low-Current Density in TbFeCo Wires", Applied Physics Express,2011

[158] J. Vogel et. al., "Direct Observation of Massless Domain Wall Dynamics in Nanostripes with Perpendicular Magnetic Anisotropy", arXiv:1206.4967v1, 2012

[159] S. Fukami et al., "Low-current perpendicular domain wall motion cell for scalable high-speed MRAM," VLSI Tech. Symp, 2009

[160] H. Zhao et al., "Sub-200 ps spin transfer torque switching in in-plane magnetic tunnel junctions with interface perpendicular anisotropy" Journal of App. Phys., 2012

[161] Zhang et al., "Thermoelectric performance of silicon nanowires" App. Phys. Lett., 2009.

[162] R. H. Koch, J. A. Katine, and J. Z. Sun, “Time-Resolved Reversal of Spin-Transfer Switching in a Nanomagnet”, Phys. Rev. Lett., v. 92, 088302 (2004).

VITA

VITA

MrigankSharad received the B. Tech and M. Tech degree in Electronics and Electrical Communication Engineering from Indian Institute of Technology, India, in 2010, where he specialized in Microelectronics and VLSI Design. Currently he is working toward the Ph.D. degree in Electrical and Computer Engineering at Purdue University. His primary research interests include low-power digital/mixed-signal circuit design. His current research is focused on device-circuit co-design for low power logic and memory, with emphasis on exploration of post-CMOS technologies like, spin-devices. He has also worked on application of spin-torque devices in approximate computing hardware, interconnect and memory design. Mrigank was awarded Prime Minister of India Gold Medal for his academic performance by IIT Kharagpur. He received Andrews Fellowship from Purdue University in 2010 and Student Innovator Award from Burton D. Morgan Center for Entrepreneurship in 2014. Mrigank has authored more than 40 papers in international journals and conferences during his PhD.