EVALUATION OF STOCHASTIC MAGNETIC TUNNEL JUNCTIONS AS BUILDING BLOCKS FOR PROBABILISTIC COMPUTING

by

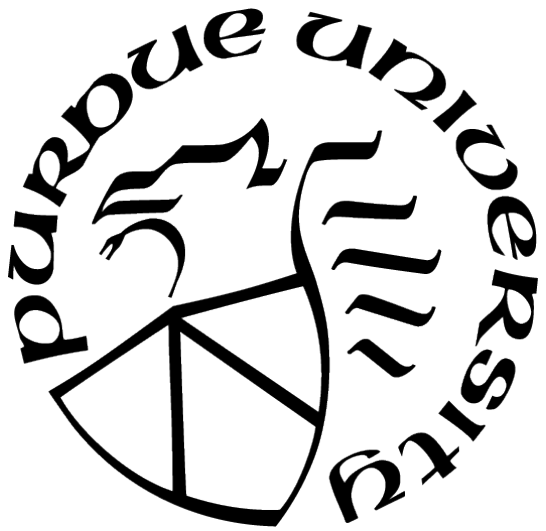
Orchi Hassan

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy



Electrical and Computer Engineering

West Lafayette, Indiana

December 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Surpiyo Datta, Chair

School of Electrical and Computer Engineering

Dr. Joerg Appenzeller

School of Electrical and Computer Engineering

Dr. Zhihong Chen

School of Electrical and Computer Engineering

Dr. Ernesto E. Marinero

School of Materials Engineering

Approved by:

Dr. Dimitrios Peroulis

This thesis is dedicated to *my family*.

ACKNOWLEDGMENTS

Looking back, I am happy to say this has been a rewarding journey of self-growth. I would like to take this opportunity to acknowledge the people who had guided, inspired, and supported me throughout this journey.

It has been an honor and privilege to work with Professor Supriyo Datta, whose guidance and research philosophy have not only shaped this work but have also shaped me as a researcher and a person in the process. Being in touch with an inspiring educator and original-thinker like Professor Datta has been an invaluable experience. Above all, he has taught me the importance of diligence and perseverance and the art of elucidating complex concepts. His incredible patience, understanding, and kindness have been nothing short of a blessing throughout this time.

I also had the good fortune to be guided by a team of expert experimentalists who made up my thesis committee. I would like to thank Professor Joerg Appenzeller and Professor Zhihong Chen for shaping my thoughts on the realities of experimental realizations. They generously gave their time and offered insightful feedback toward improving my work. I would like to thank Professor Ernesto E. Marinero for building my knowledge on magnetism.

I gratefully acknowledge Dr. Jonathan Z. Sun's (IBM) contributions to understanding the fascinating physics of low-barrier magnets. Discussions with him gave life to some of the key contents of this thesis. I would also like to thank Professor Muhammad Ashraful Alam, for guiding me in my early days at Purdue and inspiring me to constantly learn from every aspect of life.

I am thankful to my brilliant set of colleagues: Kerem Y. Camsari, Shehrin Sayed, Rafatul Faria, Ahmed Zeeshan Pervaiz, Brian M. Sutton, Jan Kaiser, Shuvro Chowdhury, Lakshmi Anirudh Ghastasala, Risi Jaiswal, and others for the support, advice, collaborations, and most importantly the stimulating group discussions over the years. I want to especially thank Kerem for being a mentor in every sense of the word.

I am *ever grateful* to my incredible friends at Purdue, who have been a major source of support in everyday life and hope we remain *ever true*. Despite living by myself, I never felt

lonely here. Purdue was home, thanks to you guys. Finally, I am grateful to my family for their unconditional love, support, and encouragement. I am forever indebted to my parents for giving me the opportunities and experiences that have made me who I am today. I am thankful to my aunt, who has done everything she can to make my life easier in the US. One of the hardest parts of grad school was living apart from my sister and my husband. The two 'Missourians' were my greatest confidants, critics, and also my biggest cheerleaders throughout. I dedicate this milestone to all of them, who believed in me and made this journey possible.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

BSN	binary stochastic neuron
FM	ferromagnet(ic)
fps	flips per second
GSHE	giant spin hall effect
IMA	in-plane magnetic anisotropy
LBM	low barrier magnet
LLG	Landau-Lifshitz-Gilbert (equation)
MCMC	markov chain monte carlo
ME	magnetoelectric
MRAM	magnetic random access memory
MTJ	magnetic tunnel junction
p-bit	probabilistic bit
PE	piezoelectric
PMA	perpendicular magnetic anisotropy
PSL	probabilistic spin logic
RNG	random number generator
SA	simulated annealing
sLLG	stochastic LLG
s-MTJ	stochastic magnetic tunnel junction
SOT	spin-orbit torque
SPICE	simulation program with integrated circuit emphasis
SR	stochastic resistance
STT	spin-transfer torque
TMR	tunnel magnetoresistance
TRNG	true random number generator

ABSTRACT

Probabilistic computing has been proposed as an attractive alternative for bridging the computational gap between the classical computers of today and the quantum computers of tomorrow. It offers to accelerate the solution to many combinatorial optimization and machine learning problems of interest today, motivating the development of dedicated hardware. Similar to the 'bit' of classical computing or 'q-bit' of quantum computing, probabilistic bit or 'p-bit' serve as a fundamental building-block for probabilistic hardware. p-bits are robust classical quantities, fluctuating rapidly between its two states, envisioned as three-terminal devices with a stochastic output controlled by its input. It is possible to implement fast and efficient hardware p-bits by modifying the present day magnetic random access memory (MRAM) technology. In this dissertation, we evaluate the design and performance of low-barrier magnet (LBM) based p-bit realizations.

LBMs can be realized from perpendicular magnets designed to be close to the in-plane transition or from circular in-plane magnets. Magnetic tunnel junctions (MTJs) built using these LBMs as free layers can be integrated with standard transistors to implement the three-terminal p-bit units. A crucial parameter that determines the response of these devices is the correlation-time of magnetization. We show that for magnets with low energy barriers ($\Delta \leq k_B T$) the circular disk magnets with in-plane magnetic anisotropy (IMA) can lead to correlation-times in *sub-ns* timescales; two orders of magnitude smaller compared to magnets having perpendicular magnetic anisotropy (PMA). We show that this striking difference is due to a novel precession-like fluctuation mechanism that is enabled by the large demagnetization field in mono-domain circular disk magnets. Our predictions on fast fluctuations in LBM magnets have recently received experimental confirmation as well.

We provide a detailed energy-delay performance evaluation of the stochastic MTJ (s-MTJ) based p-bit hardware. We analyze the hardware using benchmarked SPICE multi-physics modules and classify the necessary and sufficient conditions for designing them. We connect our device performance analysis to systems-level metrics by emphasizing problem and substrate independent figures-of-merit such as flips per second and dissipated energy per flip that can be used to classify probabilistic hardware.

1. INTRODUCTION

The future of computing beyond Moore's Law is in building heterogeneous computing platforms that can handle diverse workloads. This entails introducing new computing paradigms and architectures tailored for addressing specific applications [1] complementary to the general-purpose computing architecture. Conventional computing has been dominating the computing paradigm for decades fueled by the continuous improvements in computing performance following the observation made by Gordon Moore over 50 years ago [2]. This observation that the number of components (semiconductor transistors) in an integrated circuit doubles every two years, dubbed Moore's Law by Carver Mead [3], underpinned by Dennard scaling [4] had come to shape the modern-society itself. Even after the end of Dennard scaling in 2004, new device-physics and changes in computer architecture enabled conventional computing performance to continue to increase exponentially. However, as semiconductor device scaling reaches its physical limits the exponential growth rate is finally tapering. But on the other hand, thanks to the internet of things and our own digital footprint the amount of data available to analyze is increasing at unprecedented rates everyday [5]. As the demand to capitalize from this plethora of data increases, technological and economic forces are propelling the computing paradigm to shift from general purpose to specialized. The algorithmic success of neuro-inspired and quantum computing models in dealing with large data sets have now opened up exciting new possibilities, but demands computing performance even beyond exa-scale [6]. Moving to specialized architecture and dedicated hardware for such compute-expensive applications is likely to provide substantial performance advantage leading to significant breakthroughs [7].

A lot of progress are being made in this front. Machine Learning (ML) a decades old concept has now set to become a ubiquitous part of life [8] as its computational load is accelerated by specialized hardware like the highly parallel graphic processor units (GPUs)[9], [10]. Quantum computing, also a decades old concept has become a billion dollar industry [11] as big companies like Google, IBM, Intel, and Microsoft are taking on the grand challenge of developing commercial quantum computers demonstrating quantum supremacy [12]. Although significant progress has been made, the difficulty of scaling quantum bits

for building large-scale, error-corrected quantum computer to carry out relevant calculations that a classical computer cannot is perhaps decades away from complete [13].

In recent years, probabilistic computing has emerged as an attractive alternative for bridging the computational gap between the classical computers of today and the quantum computers of tomorrow. Most real-problems that quantum computers are aiming to solve can be formulated as combinatorial optimization problems. A way to solve these computationally complex problems efficiently is to map them onto an Ising model [14], [15] and use its intrinsic convergence properties to search for the ground state of the system and reach the solution [16]. Companies like Hitachi, Fujitsu, Toshiba, NTT, D-Wave, and others have invested in building dedicated hardware accelerators based on the Ising Model broadly termed Ising Machines. The various approaches range from quantum computers based on quantum annealing (QA) or adiabatic quantum optimization (AQC) implemented with superconducting circuits [17], coherent Ising machines (CIMs) implemented with laser pulses [18], phase-change oscillators [19], or CMOS oscillators [20]–[23] to digital annealers based on simulated annealing (SA) [16] implemented with digital circuits [24]–[30]. Digital annealers are at the forefront of the race offering scalable, room-temperature mm sized chips. However, these deterministic digital hardware only emulates the probabilistic nature of the algorithms. Infact, the random number generators (RNGs) occupy a major portion of the annealing processors today [28]. A substantial performance advantage could be achieved if probabilistic algorithms ran directly on probabilistic hardware with naturally probabilistic constituents. The probabilistic spin logic (PSL) framework based on the concept of using probabilistic bits or 'p-bits', is an embodiment of this idea [31]–[33]. p-bits are classical quantities intermediate between the stable bits of digital electronics and the q-bits of quantum computing. **In this work, we analyze the physical implementations of p-bits using naturally stochastic elements to serve as building-blocks for realization of compact, scalable, energy-efficient PSL hardware.**

1.1 Probabilistic Spin Logic

The idea of a probabilistic computer to prelude quantum computers can be found in the seminal keynote address by Richard Feynman on simulating physics with computers [34]. Feynman articulated that the only efficient means to simulate a phenomenon was with a system governed by its same fundamental constituents. A wide range of practical problems of great interest today, like machine learning and combinatorial optimization, essentially involve probabilities. So, an efficient way to solve these problems would be by using a probabilistic computer whose fundamental constituents are probabilistic bits.

In 2016, Behin-Aein et. al. [33] proposed the idea of interconnecting transistor like three terminal stochastic devices, subsequently dubbed the 'p-bit' by Camsari et. al. [31] to serve as a building block for probabilistic networks. Probabilistic spin logic (PSL) is the name given to the study of these networks of p-bits. The 'spin' in PSL is originally motivated by its strong ties to magnetic-devices and efficient realizations of hardware p-bits using low-barrier magnets which we document in this thesis.

A wide variety of problems encompassing two active but disjoint fields of research, stochastic machine learning [35], [36], and quantum computing [37]–[41] can be mapped onto the p-computer through proper design of the interconnections between the p-bits. Basically, what we propose is a naturally stochastic hardware that can implement probabilistic alogrithms that are based on Markov chain Monte Carlo (MCMC) efficiently.

A p-circuit solves a problem by naturally converging to the ground state of the system described by its energy:

$$E = I_0 \left(-\frac{1}{2} \sum_{\mathrm{i,j}=1}^{N} J_{\mathrm{ij}} m_{\mathrm{i}} m_{\mathrm{j}} - \sum_{\mathrm{i}=1}^{N} h_{\mathrm{i}} m_{\mathrm{i}} \right) \tag{1.1}$$

where, m denotes the p-bit, J is the coupling co-efficient or interconnection strength between the p-bits, h is the external bias applied to a p-bit and I_0 is a dimensionless quantity representative of the system's temperature. For machine learning applications I_0 is typically kept constant, while for optimization problems involving simulated annealing I_0 is varied.

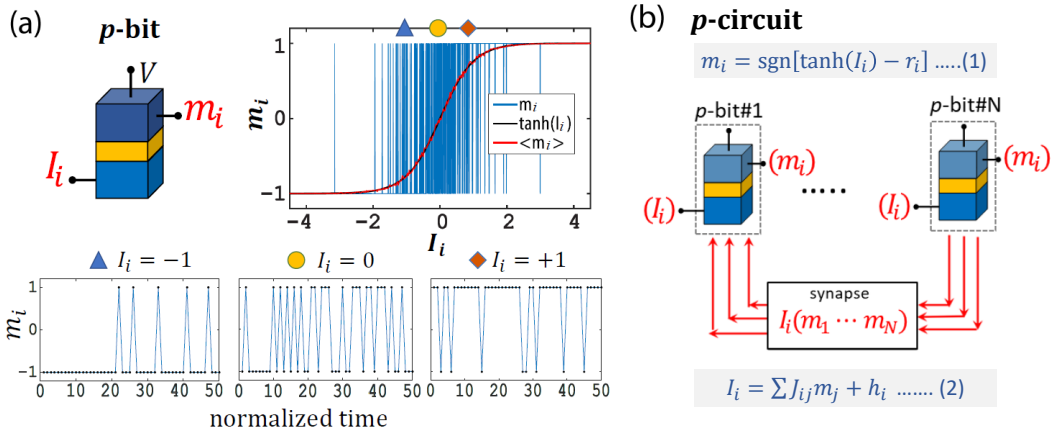


Figure 1.1. PSL framework: (a) ***p*-bit**: The p-bit is a classical quantity that fluctuate rapidly between $+1$ and -1 and its fluctuations can be tuned through an input bias I_{i}. (b) ***p*-circuit**: multiple p-bits can be connected through synaptic connections to form p-circuits to perform useful functions.

The p-bits are essentially tunable random number generators (RNGs), analogous to the binary stochastic neurons (BSNs) [42] of stochastic neural networks and can be described mathematically by

$$m_{\mathrm{i}} = \mathrm{sgn}(\tanh(I_{\mathrm{i}}) - r_{\mathrm{i}}) \tag{1.2}$$

where, r_{i} is a random number between ± 1 and I_{i} is the input to the p-bit. Here, we use bipolar variable $m_{\mathrm{i}} = \pm 1$ to represent the two states '1' and '0' of the system. The output fluctuation probability of the p-bits are controlled by their individual input I_{i} generated from the weighted sum of the states of other p-bits according to:

$$I_{\mathrm{i}} = I_0 \left(\frac{1}{2} \sum_{\mathrm{j}=1}^{N} J_{\mathrm{ij}} m_{\mathrm{j}} + h_{\mathrm{i}} \right) \tag{1.3}$$

Eq. 1.2 and 1.3 together describes the PSL framework. Problems can be mapped onto PSL through appropriate J and h [15]. The same framework accompanied by a learning rule can also be used to calculate the J and h themselves [36], [43]. Infact, eq. 1.2 and 1.3 are widely used in many modern algorithms, but they are commonly implemented in software whose performance could be accelerated by building dedicated hardware. So, *how do we build a scalable energy-efficient hardware for implementing PSL?*

Much work has gone into developing accelerators for performing the matrix multiplication and addition of eq. 1.3 which the PSL hardware can directly benefit from [44]–[47]. Our primary focus in this thesis is on *the design of a hardware accelerator for implementing eq. 1.2, the p-bit.*

1.2 Realization of Probabilistic Hardware

Any random signal generator whose randomness can be tuned with a third terminal could serve as a suitable physical realization of p-bit, but *what is the most efficient way to do it?* Completely digital implementations using conventional CMOS technology are possible [48]–[50], but getting true randomness from deterministic circuits require elaborate circuits with unfavorable size, power-consumption, and latency [19], [51]. Also, it beats our original motivation of leveraging from nature's innate stochasticity. In this thesis, we show that hardware p-bits can be efficiently and compactly realized using low-barrier magnets (LBMs) in structures similar to those in conventional magnetic random access memory (MRAM) technologies [52], [53].

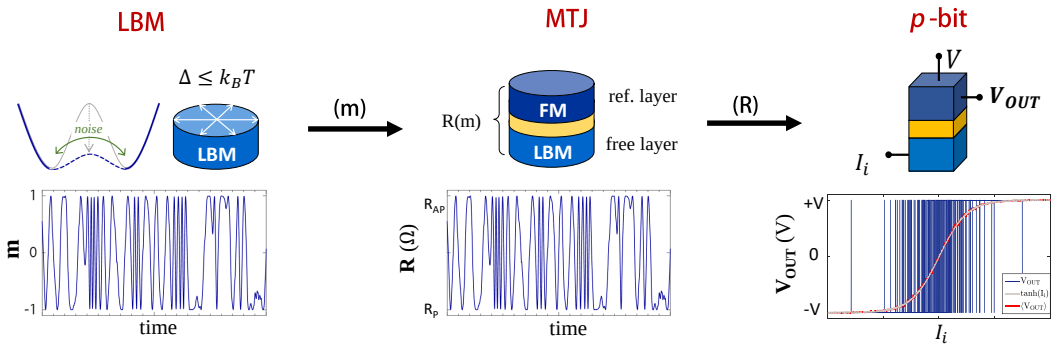


Figure 1.2. Basic Design Principle: The random thermal fluctuations in low barrier magnet's magnetization can be utilized to realize a stochastic resistor (SR) through the tunnel magnetoresistance (TMR) effect in an MTJ structure. The stochastic MTJ (s-MTJ) acts as the source of randomness (r_i) in hardware realizations of p-bits.

The key element in the designs is the stochastic magnetic tunnel junction (s-MTJ) which has been shown to be well-suited for the physical implementation of random number generators [54]–[56]. LBMs whose magnetization fluctuates randomly under thermal perturbations

can be built from perpendicular magnetic anisotropy (PMA) magnets designed to be close to the in-plane transition or from circular in-plane magnetic anisotropy (IMA) magnets [57]. MTJs utilizing these unstable LBMs as free layers present themselves as fluctuating stochastic resistances (SR). The basic design principle for p-bit hardware realization involves using the resulting fluctuating resistance (R) of such stochastic MTJ structures in conjunction with necessary electrical components to realize a tunable random number generator as shown in fig. 1.2.

We apply magnet and circuit physics to comprehensively evaluate and characterize low-barrier magnet (LBM) based p-bit implementations shown in fig. 1.3. The fundamental design principle of these devices is actually independent of the magnetic realizations, it applies to any stochastic resistor (SR). So, we also classifying necessary and sufficient conditions for designing p-bits from this general perspective and hope these design rules stimulate discussion in the realization of different stochastic resistors.

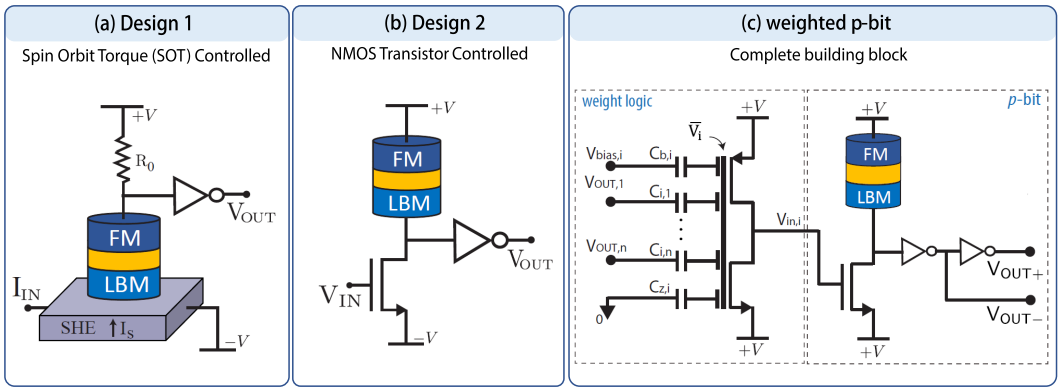


Figure 1.3. p-bit realizations with stochastic MTJ: In each design the LBM MTJs act like a stochastic fluctuating resistance. The fluctuations are tuned to behave like a p-bit. In (a) Design 1, the tunability is achieved through spin current manipulation of magnetic state. The structure and operation principle is similar to the spin-orbit-torque (SOT) controlled MRAM. In (b) Design 2, the structure looks like a spin-transfer torque (STT) MRAM, but it achieves tunability mostly through the NMOS transistor. (c) shows the realization of a compact building block - the weighted p-bit (wpbit) using design (b) coupled to a capacitive voltage adder to perform the weight logic. We demonstrate fully hardware realization of p-circuit operation using this building block through SPICE simulation.

We evaluate the performance of the proposed designs using SPICE compatible multi-physics modules, where benchmarked spintronic device models [58] are coupled with state-of-the art transistor models [59]. We project the overall performance of LBM based PSL hardware based on the individual p-bit performance characteristics. The LBM implementations of p-bits enable autonomous or clock-less operation of p-circuits and are thus not limited by clock-frequencies like digital circuits. We benchmark PSL hardware performance against digital implementation of Ising Machines by focusing on problem and substrate-independent performance metrics - flips per second (fps) and energy per flip. These metrics could serve as key figures of merit in the benchmarking test suits for the emerging class of specialized probabilistic hardware.

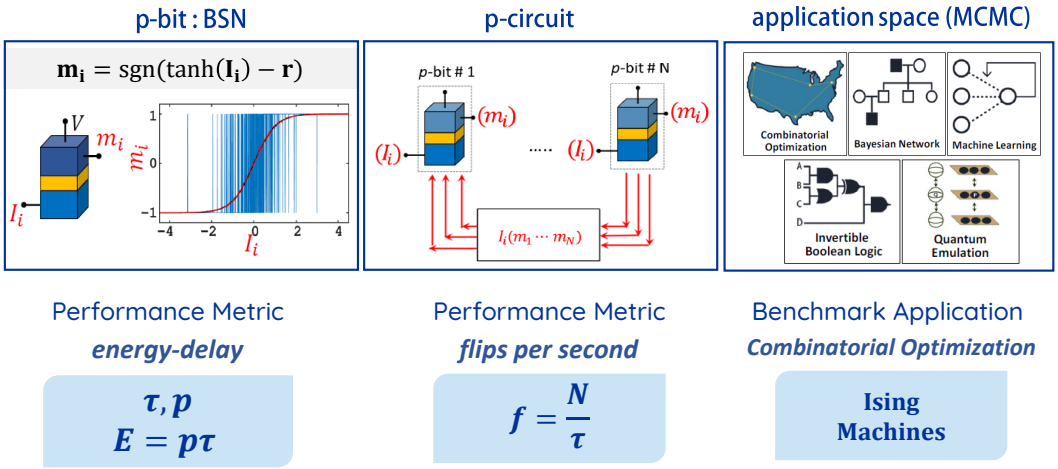


Figure 1.4. Evaluation of PSL framework: We define a set of performance metrics to benchmark the performance of PSL hardware. We evaluate the p-bit performance in terms of the average time and energy it takes to flip to a new random state. We emphasise the evaluation of hardware performance in terms of a problem independent metric - flips per second. It has been shown that PSL can be realized as a hardware accelerator for a wide spectrum of applications, in this thesis we benchmark our hardware performance against the digital implementations of Ising Machines.

1.3 Organization of Thesis

This thesis is aimed to present a comprehensive evaluation of probabilistic spin logic hardware for compact and efficient implementation of probabilistic algorithms. Our primary

focus is on sprintronic devices realized with low-barrier magnets (LBMs). We discuss the interesting physics of mono-domain low-barrier ($\Delta \leq k_BT$) magnets which is typically ignored and provide analytical expressions to characterize its fluctuation dynamics and predict current and magnetic-field response. The thesis is organized as follows:

Chapter 2 presents a detailed performance analysis of p-bit realizations using low-barrier magnet (LBM) based magnetic tunnel junctions (MTJ). We define the energy and delay for these class of devices and identify the magnet and transistor properties that contribute to them. Our analysis identifies the correlation time of magnetization be a crucial parameter that determines the response of such devices. We show that this correlation time can be in sub-ns timescales for circular disk magnets with in-plane magnetic anisotropy (IMA) having low energy barriers ($\leq k_BT$). These fast fluctuations and the compact realization leads to energy requirements of only $\sim$ a few fJ to evaluate the BSN function, orders of magnitude lower than the digital CMOS implementations.

A key result we highlight in this chapter is difference between the fluctuation dynamics of mono-domain LBM with in-plane magnetic anisotropy (IMA) and perpendicular magnetic anisotropy (PMA). The presence of large out-of-plane demagnetization fields enable an almost two-orders of magnitude faster precession-like fluctuation dynamics in the circular-IMA magnets compared to its PMA counterpart. The striking numerical observation is backed by physical understanding and analytical expressions in this chapter. Following the theoretical predictions, $\sim$ GHz fluctuations have been observed in circular IMA LBM MTJ structures recently.

Chapter 3 presents a structured design guideline for realization of p-bit hardware using stochastic MTJs for designers. We identify necessary conditions for successful realization of p-bits and define a systematic approach to matching the magnetic and circuit parameters for the embedded MTJ hardware.

We also evaluate the performance of autonomous probabilistic computing hardware realized using s-MTJs against the clocked digital implementations of Ising Machines in this chapter. We connect our device-level analysis to problem independent hardware figures-of-

merits: *flips per second* and dissipated *energy per flip* that can be used to benchmark such probabilistic hardware. The naturally stochastic hardware can overcome the technological difficulties of producing random numbers with deterministic hardware and also eliminate the need for a global clock and sequencers. The compact unit can drastically reduce the area footprint while promising massive scalability by leveraging the existing Magnetic RAM (MRAM) technology.

Chapter 4 presents the design of a complete building block for PSL by augmenting the p-bits with a floating-gate MOS-based capacitive adder to provided the weighted-sum input locally. p-bit interconnections can be implemented off-chip either in software or with a hardware matrix multiplier unit, but that requires data to be transferred back and forth. So instead we present a low-level compact hardware implementation where each p-bits come with its own local capacitive adder network to provide inputs is proposed. We call these building-blocks weighted p-bits. We demonstrate that such weighted p-bits can interconnected like gates and scale from 1-bit invertible full-adder to small instances of more complex problems like the subset-sum problem. This type of building blocks are suited for realizing p-circuits with sparse and discrete weights.

Chapter 5 departs a little from probabilistic hardware and proposes a new type of magnetoelectric memory device that stores information on magnetic easy-axis or pseudo-magnetization, in piezoelectric/ferromagnetic (PE/FM) heterostructures. We present an equivalent circuit model of the magnetoelectric (ME) phenomena and use SPICE simulations to benchmark this model against experimental data that demonstrate the read and write operation through the ME effect. We show how the magnetoelectric coupling between the PE/FM combination can lead to non-volatility in pseudo-magnetization even when the magnet is designed as a low-barrier nanomagnet.

Chapter 6 provides a summary of this work and a future outlook.

2. LOW BARRIER MAGNET DESIGN FOR HARDWARE PROBABILISTIC BITS

Most of the materials in this chapter have been extracted verbatim from the paper: "Low Barrier Magnet Design for Efficient Hardware Binary Stochastic Neurons", O. Hassan, R. Faria, K. Y. Camsari, J. Z. Sun, and S. Datta, published in IEEE Magnetic Letters, vol. 10, 2019 [60].

In this chapter we evaluate stochastic magnetic tunnel junction (s-MTJ) based realizations of the fundamental building block of probabilistic spin logic (PSL) - the probabilistic bit (p-bit). Low barrier magnets (LBMs) built either from perpendicular magnets designed to be close to the in-plane transition or from circular in-plane magnets can provide a natural physical source of randomness for the realization of p-bits [37], [57], [61]. MTJs utilizing such LBMs have been shown to be well-suited for the implementation of random number generators (RNGs) [55], [56], [62]. The p-bits, which are essentially three terminal tunable RNGs can be realized by combining s-MTJs with standard CMOS transistors, similar to spin-orbit torque (SOT) and spin-transfer torque (STT) magnetoresistive random access memory (MRAM) devices [31], [63]. We discuss the physics of low-barrier magnets and evaluate the performance of two such p-bit designs in this chapter.

Asp-bits are analogous to the binary stochastic neurons (BSNs) in stochastic neural networks, we use the word BSN and p-bit interchangeably throughout this thesis.

2.1 Hardware p-bit realizations

Many inference and machine learning algorithms are based on networks of binary stochastic neurons [42], [64]–[68] each of whose response m_{i} at time step (n+1) is determined by the input I_{i} at time n (r_{i}: random number between -1 and $+1$):

$$m_{\mathrm{i}}(n+1) = \mathrm{sgn}[\tanh\, I_{\mathrm{i}}(n) - r_{\mathrm{i}}] \tag{2.1}$$

In the absence of an input I_{i} the output m_{i} fluctuates randomly between two values -1 and $+1$. A positive $I_{\mathrm{i}}(n)$ makes $+1$ more likely, while a negative $I_{\mathrm{i}}(n)$ makes -1 more likely [69]. Each BSN described by eq. 2.1 receives its input from a weighted sum of other BSNs obtained

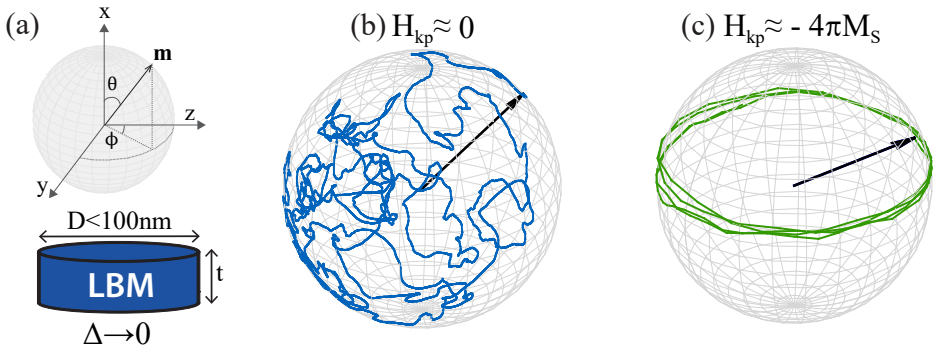


Figure 2.1. Fluctuation Dynamics of LBM: (a) Schematic illustration of circular LBM with saturation magnetization M_s and volume $\Omega = \pi(D/2)^2 t$ and the magnetization $\mathbf{m} = \mathbf{M}/M_s = (m_x, m_y, m_z) \equiv (\cos\theta, \sin\theta \sin\phi, \sin\theta \cos\phi)$. SPICE simulation shows $m(t)$ dynamics on Bloch sphere of a low barrier circular magnet with ($\Delta \approx 0$) for magnet with (b) $H_{kp} \approx 0$ and (c) $\mathrm{H_{kp}} \approx -4\pi\mathrm{M_s} \approx -13.8$ kOe, where $H_{kp} = 2K_s/t - 4\pi M_s$ is the perpendicular anisotropy along x-axis and the in-plane anisotropy $H_{ki} \approx 0$ due to circular shape.

from a "synapse" $I_{\mathrm{i}}(n) = \sum_{\mathrm{j}} W_{\mathrm{ij}}\, m_{\mathrm{j}}(n)$. A wide variety of functions can be implemented by properly designing or learning the weights W_{ij} [39], [70], [71].

The BSN function (eq. 2.1) is evaluated repeatedly in modern algorithms but they are typically implemented in software. Efforts have been put into developing a suitable hardware for accelerating evaluation of this function, many of which are based on magnetoresistive random access memory (MRAM) technology which is a major contender in the field of non-volatile memory using stable magnets to store information in the form of 0's and 1's. By contrast, BSNs can be built out of nanomagnets designed to have low energy barriers [37], [61], [62], [72]–[76]. The performance of such BSN designs are largely dependent on the magnetization fluctuation rates of the LBM's, making it important to design the low barrier magnet to have a high fluctuation rate.

The time scale of fluctuations can be very different for the two categories of low barrier magnets as shown in fig. 2.1b and c. In PMA with vanishing perpendicular anisotropy field making $\Delta \to 0$, the thermal noise makes the magnetization fluctuate randomly anywhere on the Bloch sphere, while in circular IMA with no preferred easy axis and a large effective

demagnetization field ($H_D = 4\pi M_s$) restricts the fluctuations to to a compressed region near the equator (i.e. in-plane moment), making more rapid fluctuations possible.

In this chapter, we present a distinction between fluctuation dynamics of low barrier PMA and IMA magnets providing analytical expressions for two very important parameters for performance evaluation of hardware BSNs: the correlation time τ_c and pinning current I_p for $\Delta \approx k_B T$ and below. Circular IMA magnets have a correlation time two orders of magnitude smaller compared to PMA and a pinning current that is much higher. We also present a device level performance evaluation on two previously proposed compact BSN designs [31], [77] using circular IMA magnet and show that the sub-ns operation results in only $\sim$ a few fJ of energy requirement for evaluating the BSN function which is orders of magnitude lower than its CMOS implementation [78], [79].

2.2 Low Barrier Magnet ($\Delta \leq \mathrm{k_B T}$) Dynamics

2.2.1 Correlation Time

A key parameter defining the BSNs performance would be the rate at which it produces the random numbers. For an LBM BSN, this rate is related to the magnetization fluctuation rate of the low barrier magnet. The time it takes for the magnet to lose its memory, the *correlation time* τ_c is defined by the full-width-half-maxima of the temporal auto-correlation function $C(t)$ of magnetization and could be used to characterize the relevant time-scale of operation of BSN.

In low barrier magnets where the energy barrier is well below the thermal energy ($\Delta \ll k_B T$) its magnetization becomes a continuous variable. The Arrhenius law which describes the thermal fluctuations of high barrier magnets ($\Delta \gg k_B T$) with two distinct magnetic states thus does not hold for LBM [61], [80]. Instead, thermal fluctuations in monodomain low barrier magnets could be characterized starting from Fokker-Planck equation (FPE)[81], [82] or the Landau-Lifshitz-Gilbert (LLG) equation including a Langevin term describing thermal fluctuation [80], [83].

Coffey et. al. [82] analyzes the magnetic fluctuations in a PMA magnet due to thermal noise in detail by using the Fokker-Planck equation (FPE) derived by W. F. Brown [81]. The

analysis presented in these references focused on high-barrier magnets but are not limited to it and thus can be evaluated for $\Delta \to 0$ to describe the low barrier magnet dynamics of PMA magnets which agree well with numerical results.

$$\text{PMA:}\quad C(t) = \exp\left(-2\alpha\gamma\frac{k_B T}{M_S\Omega}|t|\right) \qquad \tau_c = \frac{M_S\Omega}{\alpha\gamma k_B T}\ln(2) \tag{2.2}$$

In low barrier circular IMA magnets when thermal noise kicks the magnetization out-of-plane, due to absence of an easy axis and the presence of large orthogonal demagnetization field H_D the in-plane magnetization starts precessing. If we consider an ensemble of such magnets each with a different precession frequency due to thermal noise, the average magnetization vector would quickly dissipate. The auto-correlation function of the in-plane magnetization $m_z = \cos(\phi(t))$ could be expressed as:

$$C(t) = \int_{-1}^{1} dm_x \cos(\gamma H_D m_x t)\rho(m_x) \Big/ \int_{-1}^{1} dm_x \rho(m_x)$$

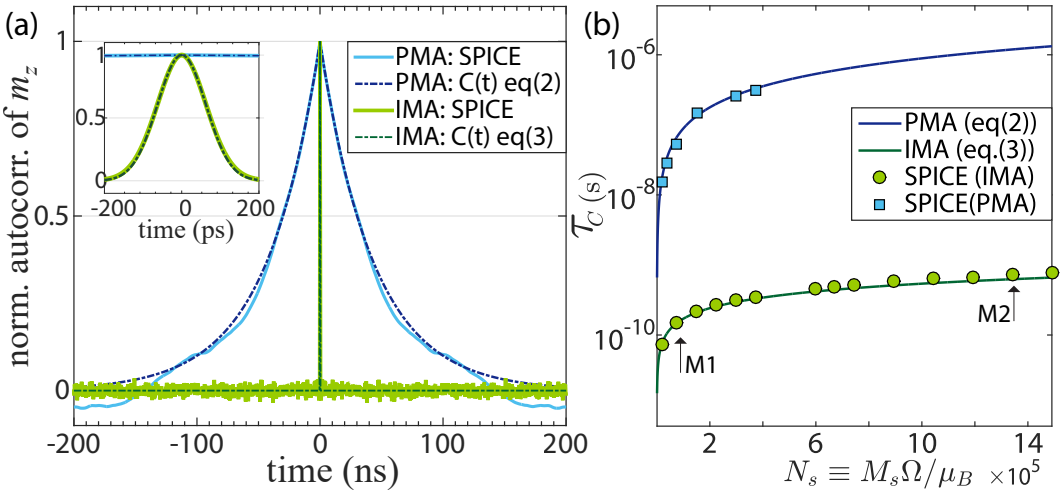


Figure 2.2. Correlation Time of PMA and IMA magnets (a) The normalized auto-correlation of magnetic fluctuations taken in the z direction, (b) Comparison of τ_c as a function of number of spins $N_s \equiv M_s\Omega/\mu_B$ where $M_s = 1100$ emu/cc and the volume Ω is varied. Damping coefficient α is assumed to be 0.01: Results from numerical simulations agree well with the equations cited in the text.

where the in-plane precession dynamics is described by $\phi(t) \approx \gamma H_D m_x t$ [83] for low damping α. The perpendicular magnetization m_x follows a Boltzmann distribution with $\rho(m_x) \approx \exp(-H_D M_S \Omega m_x^2 / 2k_B T)$. For large values of H_D the integral could be extended to $\pm\infty$ and evaluated to give an expression for the auto-correlation function and correlation time as follows:

$$\text{IMA:} \quad C(t) = \exp\left(-\gamma^2 \left(\frac{H_D k_B T}{M_S \Omega}\right) \frac{t^2}{2}\right)$$
$$\tau_c = \sqrt{8\ \ln(2)} \frac{1}{\gamma} \sqrt{\frac{M_S \Omega}{H_D k_B T}} \tag{2.3}$$

In numerical simulations, we observe essentially the same auto-correlation behavior, even when the correlation function is obtained from the time-dependent fluctuations of a single magnet fluctuating for long time periods as shown in fig. 2.2a. In PMA no such precessional fluctuation mechanism exists as the internal fields are compensated.

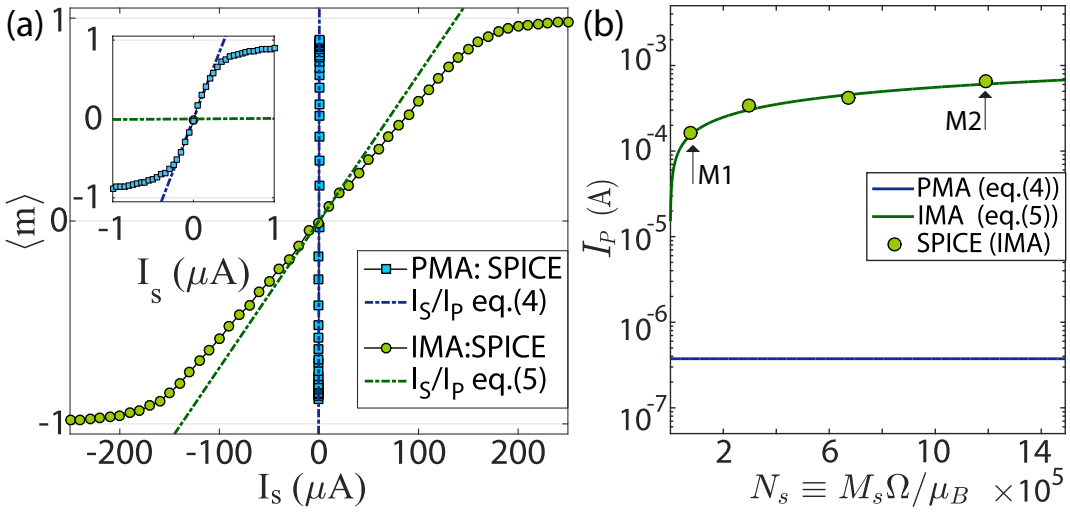


Figure 2.3. Pinning current of PMA and IMA magnets (a) PMA and IMA magnet's long time averaged magnetization $\langle m \rangle$ as a function of applied spin current I_S, (b) Comparison of PMA/IMA I_P as a function of number of spins $N_s \equiv M_s \Omega / \mu_B$ where $M_s = 1100$ emu/cc and the volume Ω is varied. Damping coefficient α is assumed to be 0.01: Results from numerical simulations agree well with the equations cited in the text.

2.2.2 Biasing Current

Another important parameter for evaluating an LBM based stochastic device performance is itds sensitivity to spin current. To maintain stochasticity in MRAM type devices, they should be immune to read current, and the amount of current required to bias BSN devices is also relevant for power considerations. In high barrier magnets the concept of switching current is presented [84], for low barrier magnets we refer to *pinning currents* as the relevant quantity which can be mathematically defined as: $I_P = (\langle m \rangle / I_S)^{-1}$ as shown in fig. 2.3. The pinning currents for PMA can be derived from steady-state Fokker-Planck equation as described in Ref. [85], while for IMA magnets with $\Delta \to 0$ and low damping, the pinning current can be approximated from the relation $I_P \equiv qN_S C(0) \Big/ \int_0^\infty dt C(t)$. fig. 2.3 shows that the numerical results are well described by the obtained expressions:

$$\text{PMA:} \quad I_P = \frac{6q}{\hbar}\, \alpha k_B T \tag{2.4}$$

$$\text{IMA: } I_P = \frac{2q}{\hbar}\sqrt{\frac{2}{\pi}}\sqrt{H_D M_S \Omega\, k_B T} \tag{2.5}$$

The derivation of eq. 2.4 and eq. 2.5 assume zero energy barriers, but numerically we observe that these equations are approximately valid for barriers up to $\Delta \approx k_B T$. In practice obtaining near-zero barrier circular magnets could be challenging due to process variation. For interconnected networks of p-bits, a distribution of correlation times for each p-bit needs to be considered as shown in Ref.[86].

Note that IMA-based designs can achieve sub-nanosecond correlation times even with fairly large volumes, provided that monodomain behavior can be preserved with a small enough diameter, while PMA-based designs tend to be much slower making IMA magnets more suitable for BSN applications. This is accompanied by fairly large pinning currents for IMA compared to PMA which minimizes read disturb effects.

In the following section we used circular IMA magnets M1 and M2 with volumes 800π and 20480π nm^3, respectively for evaluating the performance of two LBM based hardware BSN designs.

2.3 Performance Evaluation of p-bits

In this section we evaluate the steady-state and time response of two hardware BSN designs proposed in the past [31], [77] shown in fig. 2.4 and measure the energy and delay associated with each.

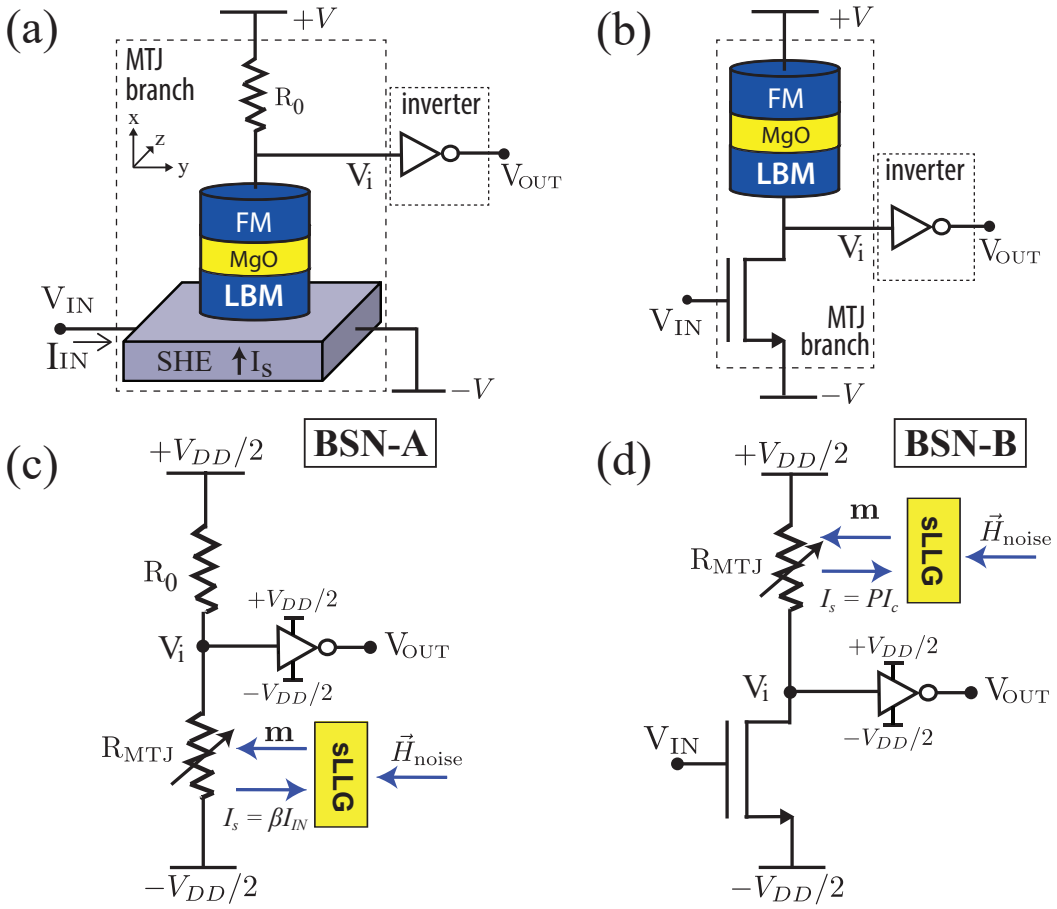


Figure 2.4. Two BSN designs using stochastic MTJ with fluctuating resistance: (a) **BSN-A** uses an input spin current to pin the fluctuating resistance [31]. Structurally it looks similar to spin-orbit torque magnetoresistive random access memory (SOT-MRAM). (b) **BSN-B** looks similar to spin transfer torque MRAM (STT-MRAM) but it makes no use of spin torque. The input voltage controls the resistance of a field effect transistor (FET) which is in series with the MTJ [77]. (c) and (d) show the circuit models used for SPICE simulations.

The designs makes use of a magnetic tunnel junction (MTJ) whose free layer is a low barrier magnet with a fluctuating magnetization $m_z(t)$, resulting in a fluctuating resistance, $R_{\rm MTJ}(t)^{-1} = G_0[1+m_{zi}(t){\rm TMR}/(2+{\rm TMR})]$ where G_0 is the average conductance and TMR

is the tunneling magnetoresistance. The fluctuating resistance $R_{MTJ}(t)$ is converted to a fluctuating voltage $V_{\rm i}$(t) by the potential divider:

$$\frac{V_{\rm i}(t)}{V_{DD}/2} = (\pm)\frac{R_{\rm MTJ}(t) - R_0}{R_{\rm MTJ}(t) + R_0} \tag{2.6}$$

The fluctuations are controlled by two different mechanisms in the two designs. BSN-A is a spin-orbit-torque controlled device [31] which uses the input spin current (in y direction) from the GSHE layer to pin the free layer magnetization (in z direction) of the MTJ thereby pinning R_{MTJ} and implements (+) configuration of eq. 2.6. BSN-B is a series resistance controlled device [77] which uses the input voltage to control the transistor resistance R_0 and implements the (−) configuration of eq. 2.6. Ideally R_{MTJ} remains unchanged, though in actual designs it may be important to consider unintended pinning effects of the current. Both designs use a minimum sized CMOS inverter to convert the fulctuating $V_{\rm i}$ into a rail-to-rail output V_{OUT}. In each case we will use SPICE simulations based on state-of-the-art stochastic Landau-Lifshitz-Gilbert (s-LLG) models for LBM's [87] free layer of the MTJ having $G_0 \simeq (25K\Omega)^{-1}$ and TMR $= 2P^2/(1-P^2) = 110\%$ with polarization $P \simeq 0.6$ coupled with 14 nm HP FinFET's [88] to show that the output voltage V_{OUT} from a specific BSN is approximately related to its input V_{IN} by an equation that mimics eq. 2.1 :

$$\frac{V_{OUT}(t+t_0)}{V_{OUT0}} \approx \text{sgn}\left[\tanh\ \frac{V_{IN}(t)}{V_{IN0}} - r(t)\right] \tag{2.7}$$

with scaling factors V_{OUT0}, V_{IN0}, t_0 characterizing the specific hardware design.

2.3.1 Steady-State Response

Fig. 2.5 shows the individual steady state response of design A,B using magnet M1 and M2, which can all collapse onto the same curve using appropriate scaling parameters. The output scaling quantity $\rm V_{OUT0} \simeq V_{DD}/2 = 0.4V$ is the same for all cases as this quantity is defined entirely by CMOS inverter output voltage swing. On the other hand, the input scaling parameters are very design dependent. For BSN-A $\rm I_{IN0}$ is determined by pinning currents of magnets M1 and M2. Indeed, the scaling parameters in fig. 2.5b were obtained

from eq. 2.5. For BSN-B $V_{IN0} \sim 50$mV for both magnets, determined by transistor characteristics. Note that the SPICE simulations include the read disturb current, but its effect is minimal due to the high pinning currents of low barrier IMA compared to PMA as can be seen from eq. 2.4 and eq. 2.5.

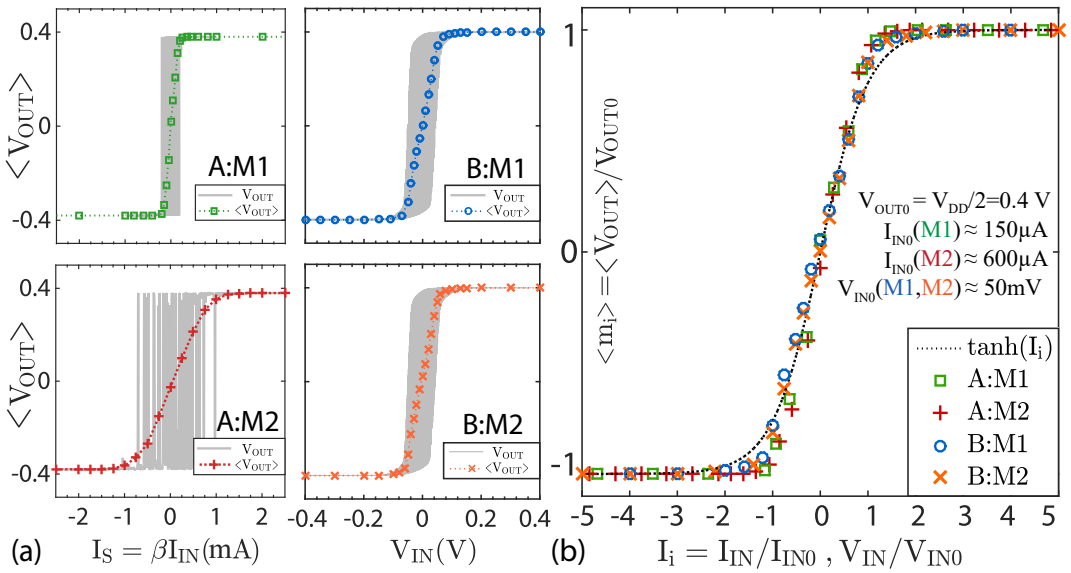


Figure 2.5. Steady-state Response: (a) Plot of $\langle V_{OUT} \rangle$ (averaged over a time window $\gg \tau_c$) vs V_{IN} for designs A, B using magnets M1, M2. The grey lines indicate V_{OUT} without time averaging. (b) All four plots in (a) collapse onto a single curve using appropriate scaling parameters $V_{OUT0}, I_{IN0}, V_{IN0}$. The resulting curve approximately follows the time averaged $\langle m_i \rangle$ of eq. 2.1.

2.3.2 Time Response

Fig. 2.6 shows the two relevant timescales associated with BSN operation. First is the correlation time of the output voltage which is determined by the magnet parameters. Indeed, the FWHM of the autocorrelation function corresponds well to eq. 2.3, which is expected since circuit related times are much shorter in this case. Second is the response time which is very design dependent. For BSN-A it is determined by magnet physics while for BSN-B it is determined by transistor physics [89]. Our analysis shows that the response time t_0 of a single BSN-B neuron is independent of magnet parameters. However, the response of an interconnected network of such neurons would also involve the magnet correlation time τ_c.

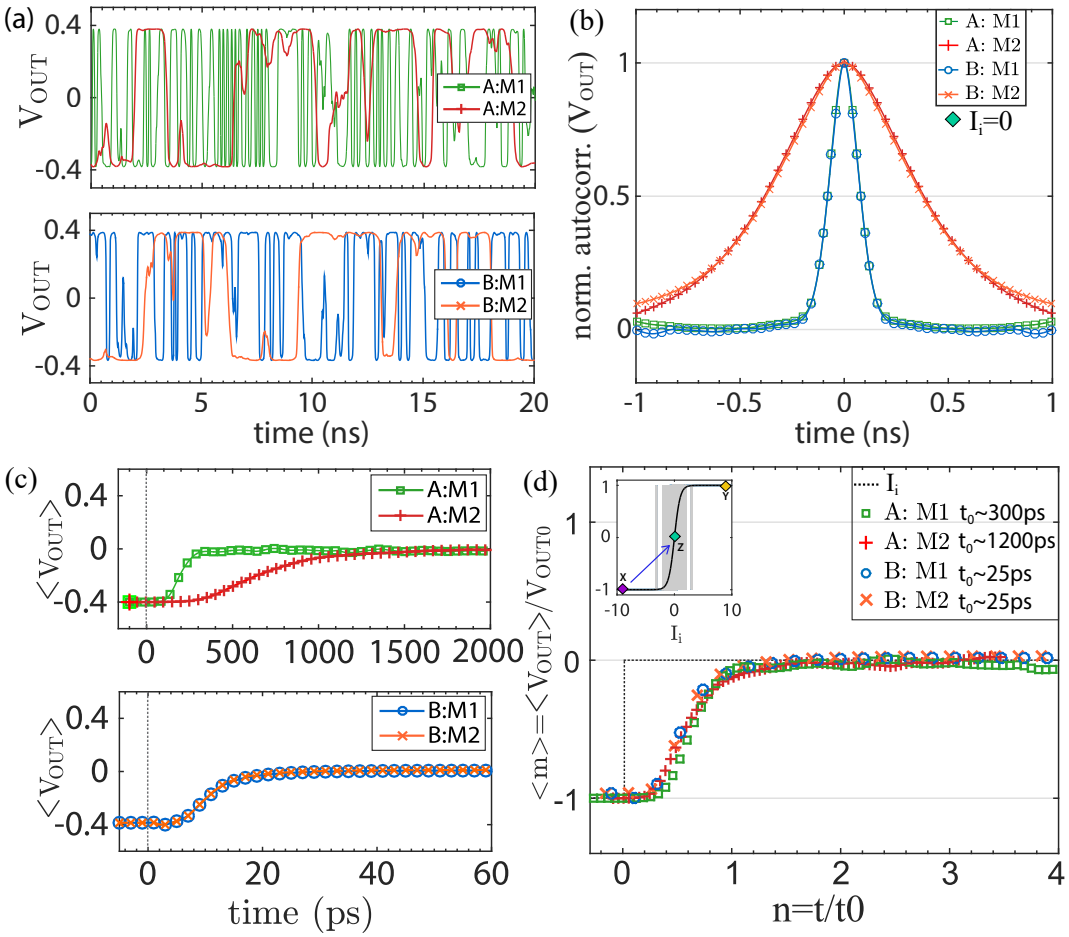


Figure 2.6. Two relevant time-scales for BSN Operation: (a), (b) show correlation time and (c),(d) show response time. (a) Output voltage fluctuations with $I_{\rm i} = 0$ for designs A, B using magnets M1, M2. (b) Corresponding normalized autocorrelation functions. (c) Response to a step function $I_{\rm i} : -10 \rightarrow 0$ at t=0 averaged over 1000 ensembles for all four cases.(d) All four curves in (c) collapse onto a single curve using appropriate scaling parameter t_0.

2.3.3 Power Consumption

Fig. 2.7 shows the power drawn from the sources $\pm V_{DD}/2$ individually by the MTJ branch and the inverter branch as V_{IN} is stepped at $t = 0$ from different initial to final values as indicated. The steady-state values of the power dissipated in both the MTJ and inverter branches agree quantitatively with the simple estimate (see dashed line in figures) from V_{DD}^2/R, where R is the appropriate resistance, namely $R_{MTJ} + R_0$ for the MTJ branch, and $R_{\rm NMOS} + R_{\rm PMOS}$ for the inverter branch. For the MTJ branch, the power dissipated is

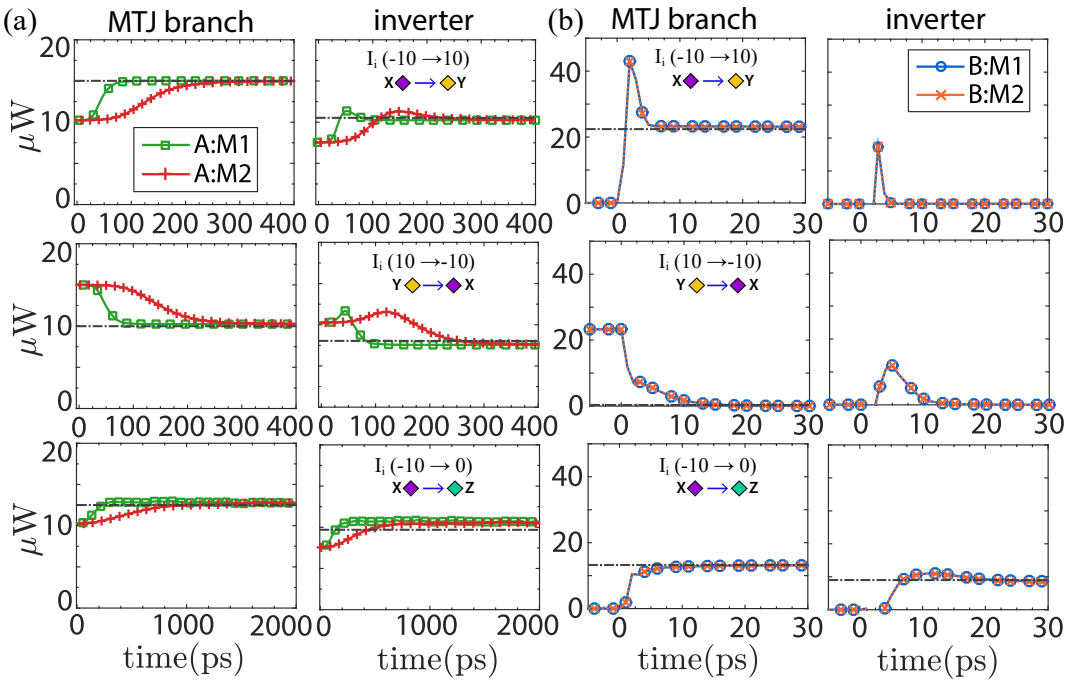


Figure 2.7. Power Consumption for (a) BSN-A and (b) BSN-B when the input is stepped at t=0 as indicated.

∼10-20 μW for all cases except in the middle panel for BSN-B. In this case the final state involves a large negative input voltage V_{IN} for which the series transistor is turned OFF, making the resistance R extremely large, so that $V_{DD}^2/R \to 0$. In all other cases, the total R is of the order of the MTJ resistance $\sim 25K\Omega$, so that $V_{DD}^2/R \sim 25\mu W$. For the inverter branch, BSN-A dissipates ∼10 μW since the voltage at the inverter input in all cases remains close to the threshold value making both NMOS and PMOS branches fairly conducting. On the other hand, for BSN-B, PMOS and NMOS get turned off for large positive and for large negative input V_{IN} respectively, making the effective R very large. Only for input voltages ~ 0, both PMOS and NMOS branches are conducting, giving rise to a steady-state power $\sim 10\mu W$ like BSN-A. This number could be lowered if we can engineer larger voltage fluctuations at the inverter input, $|\delta V_{\rm i}| \sim P^2 V_{DD}/(4 - P^4)$. Our assumed TMR of 110% corresponds to $P \sim 0.6$, giving a $|\delta V_{\rm i}| \sim 75\ mV$.

Note that in this analysis the power drawn from V_{IN} is not considered which is expected to be very different for a low input impedance design (BSN-A) compared to a high input impedance design (BSN-B) and will depend on the driving mechanism and circuitry. Overall,

both designs suffer from significant steady-state power losses and would need to be turned off when not in use. This can be done straightforwardly for BSN-B using a large negative input voltage V_{IN}. The key point to note is that the energy dissipated during the evaluation of the BSN function is $\sim 20\ \mu W \times 50\ ps$ =1 fJ which is orders of magnitude smaller than CMOS implementations of the same function [78], [79] as noted earlier from system level simulations in [90].

2.4 Summary

The device level analysis presented here elucidates the role of proper magnet design for achieving the subnanosecond response times that is crucial for fast and low energy operation. The analysis also suggests low barrier IMA magnets maybe a more suitable candidate for p-bit type applications due to its fast fluctuation dynamics, while modern non-volatile MRAM technology is largely based on PMA magnets [52].

3. EVALUATION OF PROBABILISTIC BITS FOR ACCELERATING ISING MACHINES

Most of the materials in this chapter have been extracted verbatim from the paper: "Quantitative Evaluation of Hardware Binary Stochastic Neurons", O. Hassan, S. Datta, and K. Y. Camsari. (to be submitted)

In the era of internet of things (IoT), combinatorial optimization problems are ubiquitous [25]. Infact, most of the real-problems that quantum computers are aiming to solve can be formulated as combinatorial optimization problems.From directing traffic flow [91], to routing interconnections in integrated circuit design [92], [93], to making financial decisions [94], drug discoveries [95], etc. - all involve solving a form of combinatorial optimization problems. The demand for solving these problems faster and more efficiently is ever-increasing. But such problems typically fall into the category of NP-hard or NP-complete class in computational complexity theory [14], with no known polynomial time solution, making them notoriously difficult to solve in digital computers using traditional computing methods. This has made the making way for a new paradigm in computing: Ising computing. Ising computing maps combinatorial optimization problems to an Ising model, and solves it by searching for the ground state of the system described by [15], [37]:

$$E = -\frac{1}{2}\sum_{\mathrm{i,j}=1}^{N} J_{\mathrm{ij}} m_{\mathrm{i}} m_{\mathrm{j}} - \sum_{\mathrm{i}=1}^{N} h_{\mathrm{i}} m_{\mathrm{i}} \tag{3.1}$$

where, m denotes the Ising spin, J is the coupling co-efficient and h is the external bias. In the machine learning field, the same underlying principle is used to for Boltzmann Machines. The binary stochastic neurons (BSNs) [65] of stochastic neural networks are well suited to function as a 'spin' is such systems, described mathematically by:

$$m_{\mathrm{i}} = \mathrm{sgn}[\tanh(I_{\mathrm{i}}) - r_{\mathrm{i}}] \tag{3.2}$$

where r_{i} is a random number between $+1$ and -1, and $I_{\mathrm{i}} = -\partial E/\partial m_{\mathrm{i}}$ is the input to the neuron.

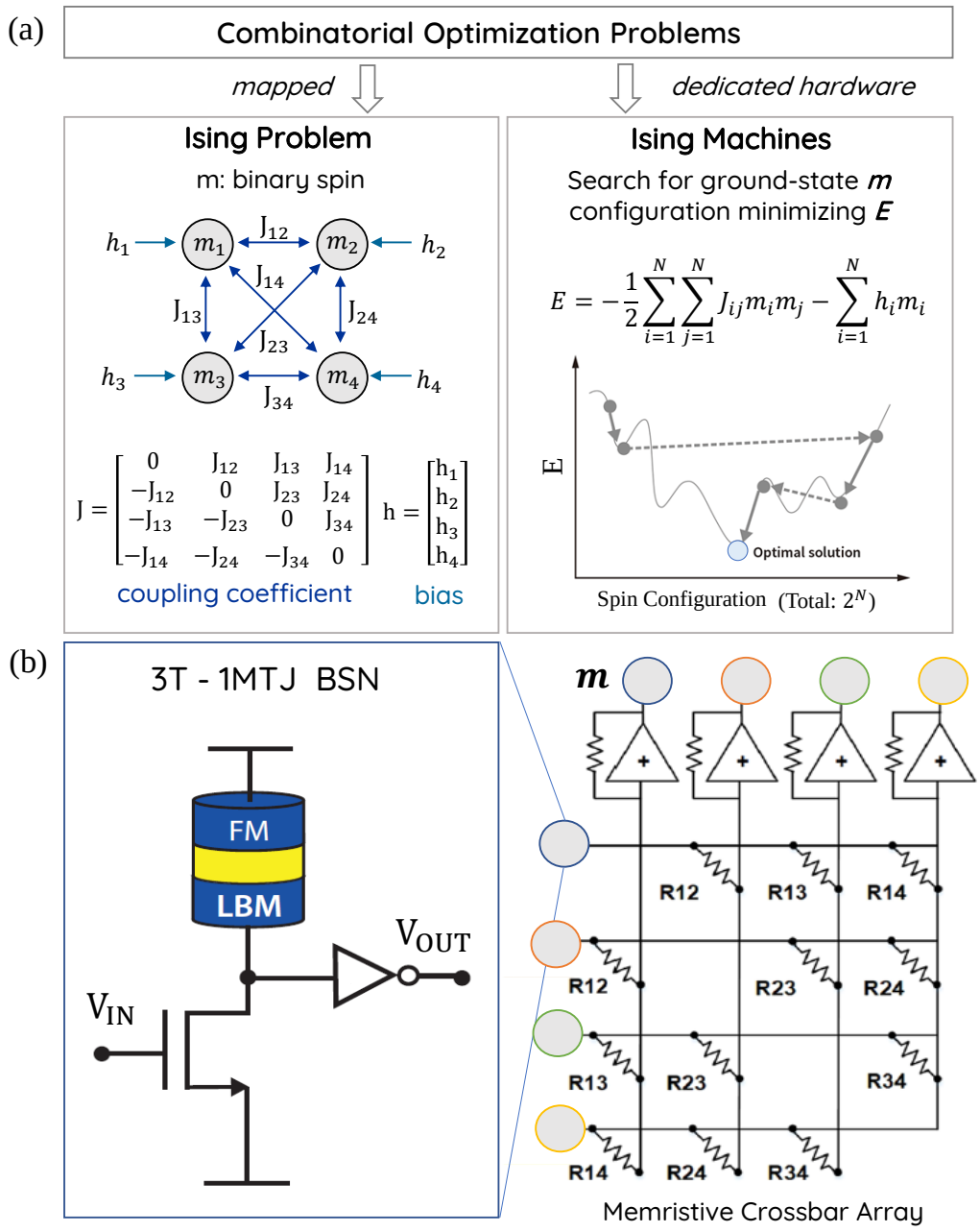


Figure 3.1. 1MTJ-3T compact BSN hardware which utilizes the natural physics of low-barrier nanomagnets holds the promise to accelerate the simulated annealing processors.

Given the importance of optimization problems, a lot of research has gone into developing algorithms and identifying appropriate hardware for Ising computing. Various approaches including quantum computers based on quantum annealing (QA) or adiabatic quantum

optimization (AQC) implemented with superconducting circuits [17], coherent Ising machines (CIMs) implemented with laser pulses [18], phase-change oscillators [19], or CMOS oscillators [20]–[23] and digital annealers based on simulated annealing (SA) [16] implemented with digital circuits [24]–[30] are being explored.

In this chapter, we comprehensively evaluate and characterize a stochastic magnetic tunnel junction (sMTJ) based realization of the Ising spin (eq. 3.2) where random numbers are generated using the natural physics of low barrier nanomagnets [77] in a compact design. A network of these BSN units can be coupled with a memristive crossbar array [96]–[98] to perform the synaptic operation as shown in fig. 3.1 can drastically improve the area requirements and accelerate computation speed of Ising Machines. We evaluate the performance of the BSN device in terms of its energy and delay metrics and connect these to the problem and substrate-independent metric of *flips per second* that the probabilistic system makes [48].

Our evaluation of 1MTJ-3T BSN design considers different types of low-barrier nanomagnet realizations of MTJs. As the MTJ essentially functions as a two-terminal stochastic resistor (SR), we first take a general 3T-1SR design approach, classifying necessary and sufficient conditions for achieving the BSN operation for different types of SRs in Section 3.1. We relate these conditions to the different sMTJ realizations in Section 3.2. We report the timescale of operation, power and energy for each case based on benchmarked SPICE simulations of the BSN hardware consisting of spintronic elements from a modular circuit framework [99] coupled to 14nm FinFET PTM models [59], and provide analytical results for relevant quantities in Section 3.3. Lastly, we use these device performance metrics to project onto hardware performance figures of merit such as flips per second that a probabilistic sampler makes. Our projections indicate orders of magnitude improvement potential over current digital implementations.

3.1 General Approach to Design of BSN

Binary stochastic neurons (BSNs) are well suited to function as a 'spin' in Ising machines for solving combinatorial optimization problems [60], [65]. A compact and efficient hardware

realization of the BSN leveraging the natual physics of stochastic nanomagnets can be made by using unstable magnetic tunnel junctions (MTJs) [62], [100]–[103] as shown in fig. 3.1.

The compact design of BSN based on low-barrier magnet (LBM) stochastic MTJs (sMTJs) was first proposed in 2017 [77]. Using magnet and circuit physics to analyze the performance, it was reported that using an LBM in a circular disk geometry with energy barriers below k_BT as the free layer of an MTJ results in sub-ns response times requiring only $\sim$ a few fJ of energy per random bit [60]. The proposed design and the performance analysis considers a very specific type of sMTJ which had circular in-plane magnetic anisotropy (IMA) whose fluctuations are undisturbed by the current in the circuit for typical current drive conditions. However, in 2019, a version of the BSN design that was implemented in hardware to solve an 8-bit factorization problem [38], consisted of an sMTJ with perpendicular anisotropy (PMA) and a barrier of a few k_BT as its free layer. Unlike the circular in-plane design, the PMA design relied on its resistance being tunable by the spin-transfer-torque effect in order to achieve the BSN operation. This has called for an extension of our initial analysis presented in [60] which we systematically perform in this chapter.

As the MTJs in the BSN circuit effectively act as a fluctuating resistor, R [104] and the design principle is independent of this realization, for establishing the fundamental design rules we approach it from a general perspective and we hope these design rules stimulate discussion in the realization of different stochastic resistors that use different mechanisms [105]–[109].

3.1.1 Types of fluctuating resistances

We categorize the fluctuating R into four types. First based on the fluctuating nature it can be continuous or bipolar (telegraphic). Second, it can be tunable or non-tunable depending on whether it is affected by the current that is flowing through it.

A continuous resistor can have its resistance being any value between $[R_P \rightarrow R_{AP}]$ while a bipolar resistor only assumes the two values R_P and R_{AP} as shown in fig. 3.2(a). The distribution of continuous resistances can be of different types as well. It can be uniform or follow slightly bimodal distribution in the case of an MTJ as shown in the figure. Differ-

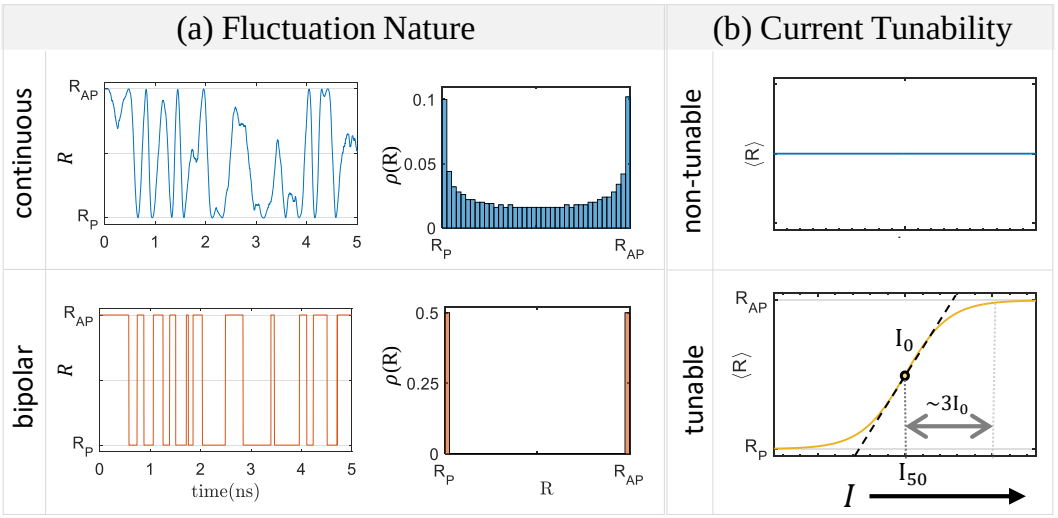


Figure 3.2. Categorizing Resistances: (a) Fluctuating nature: they can be continuous or bipolar. The time dynamics and distribution are shown for each category. (b) Current-Tunability: The fluctuations could be unaffected by I or it could be a function of I as indicated by their transfer characteristics. I_{50} is the current at the 50:50 point where the resistance spends equal time in R_P and R_{AP} states. I_0 is the biasing current defined as the slope of the (R vs I) curve at 50:50 point. The pinning current is typically $\sim 3-5\ I_0$.

ent distributions typically result in different average R values, slightly bimodal or uniform distributions are better suited than Gaussian distributions for BSN realizations.

The current I flowing in the circuit can tune the probability distribution of the resistance fluctuations, and we call such resistors tunable resistors. When designing a BSN with current tunable R, we need to know the current where fluctuations are equal between the two extreme states (I_{50}) [104] and the current required to pin the resistance to one of those states. An important parameter in this case is the bias current I_0, which is the slope of the R vs I curve at the 50-50 point. Typically, $\sim 3-5$ I_0 current is required to pin the fluctuating resistance to one of its states. We will later provide analytical expressions for I_0 for four cases of resistors that can be obtained by various MTJs (fig. 3.10).

Based on this analysis, we categorize the fluctuating resistance into four types: Non-tunable continuous (NTC), Non-tunable bipolar (NTB), tunable continuous (TC) and tunable bipolar (TB).

3.1.2 Performing the BSN function

We first take a look at the transfer characteristics of the device to see whether the four types of resistance can faithfully mimic BSN operation described by eq. 3.2. The fluctuating R is a physical realization of the random variable r_i, the NMOS acts as a constant current source that provides tunability, and the inverter performs the sgn operation in eq. 3.2.

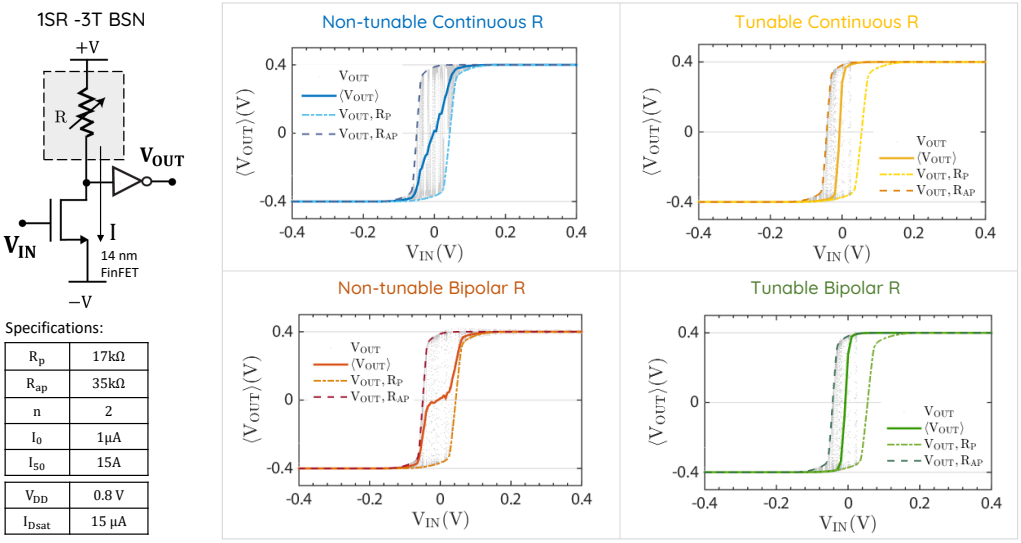


Figure 3.3. Transfer Characteristics : The BSN circuit is realized by coupling the fluctuating resistor which is the physical realization of the random variable r_i in the BSN equation to an NMOS which provides the tunability, and then to an inverter which thresholds the output. The four types of resistances are coupled to a 14nm FinFET and the resistance parameters (based on experimental demonstrations of MTJs [110]) are chosen to match the transistor characteristics. All resistance types except for the bipolar non-tunable were able to achieve BSN operation following eq. 3.2. To function as a BSN the bipolar resistances need some means of tuning their probability distribution.

Fig. 3.3, shows that while all other resistance types were able to reproduce the desired sigmoidal average curve $\langle m_i \rangle = \tanh(I_i)$, the non-tunable bipolar resistor gives a staircase-like function instead.This is because of the fixed delta function like resistance distribution at the two extreme states (see fig. 3.2(a)ii). As there is no continuity in the resistance distribution and no means of tuning the delta distribution itself, the BSN output fluctuations are equal until either of the threshold points are crossed, resulting in the stair-case like function.

Mathematically, when the resistance is bipolar, it means r_i is ± 1. So, for any input I_i where $|\tanh(I_i)| < 1$, the output $\langle m \rangle$ is equal to zero. In fig. 3.4(b), if we look at a simple invertible AND gate [31], [77] operation, it is evident that devices with stair-case like function cannot be used as BSN. This has been demonstrated experimentally in ref. [111], [112] where a stable MTJ was used as a bipolar resistor whose distribution was tuned by an external field.

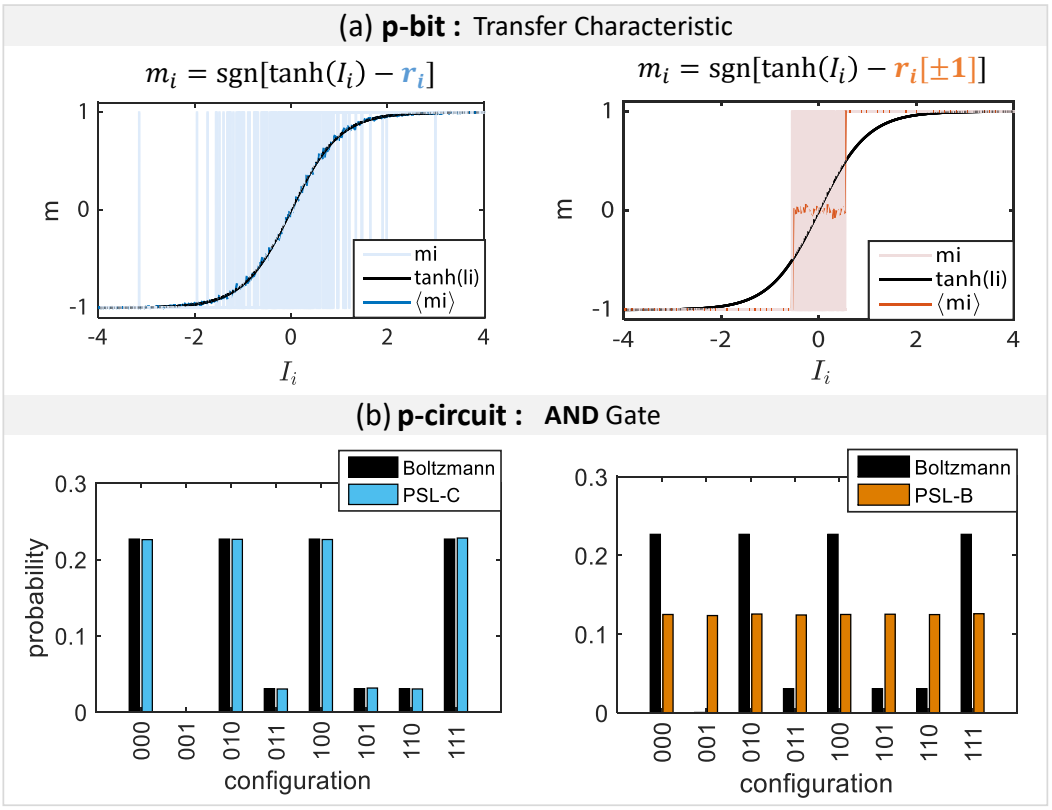


Figure 3.4. Non-tunable Continuous vs Bipolar Resistance: (a) Transfer Characteristics shows that while the continuous resistor results in a sigmoidal output, the bipolar gives a stair-case like function. (b) The bipolar R is unable to follow the Boltzmann distribution of the invertible AND gate (description in ref.[31]). All states remain equally probable.

3.1.3 Parameter Dependence and Design Choices

Fig. 3.3 is created with a fixed set of parameters for the resistor and coupled with a specific transistor technology, 14 nm FinFET models. In this section we explore how the transfer

characteristics are affected by different parameters of the resistors and FET characteristics and how to choose the right combination of R and FET to be coupled.

Stochastic Region: The stochastic region, which we define next, is a function of the resistance ratio n for non-tunable resistors and biasing current I_0 for tunable resistors as shown in fig. 3.5, that needs to be matched with the transistor characteristics.

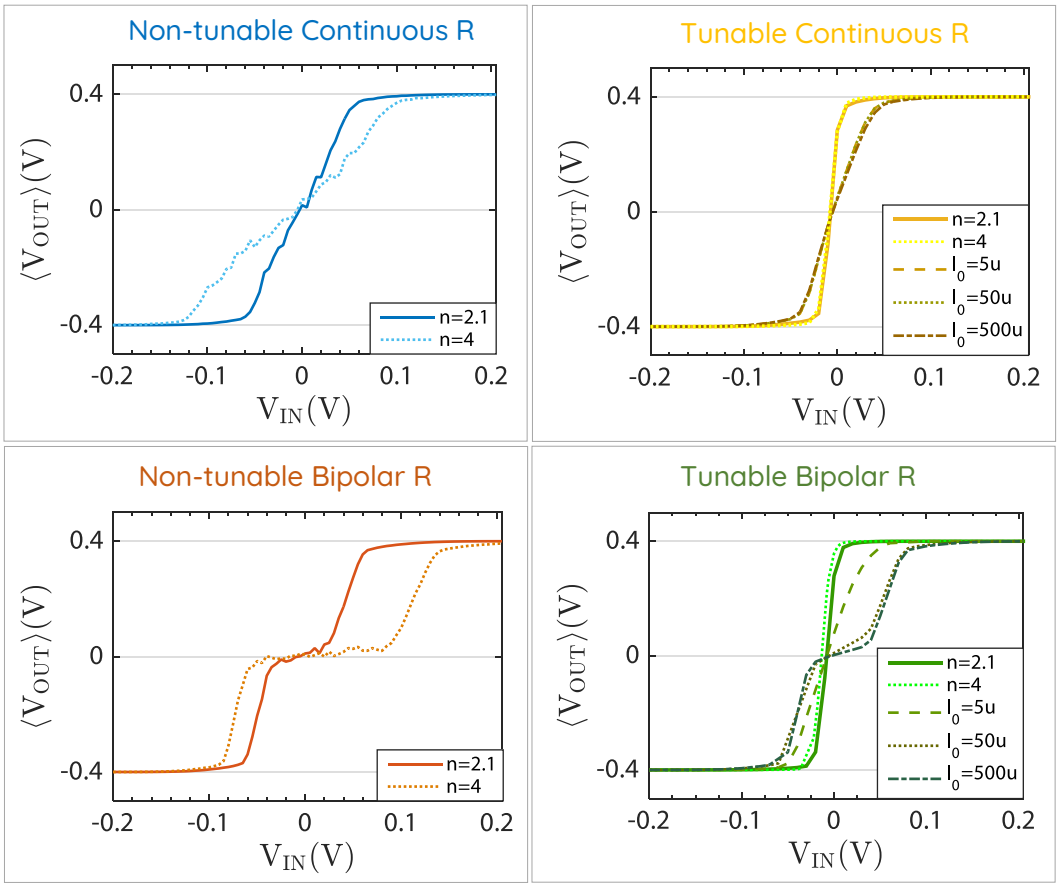


Figure 3.5. Effect of n and I_0 : The stochastic region of the non-tunable resistances are determined by the resistance ratio $n = R_P/R_{AP}$, while the biasing current I_0 of tunable resistances control the stochastic region. For large biasing currents, the tunable resistors behave effectively like non-tunable resistances.

Effect of n: The resistance ratio $n = R_P/R_{AP}$ is directly related to the stochastic region Δv through the NMOS characteristics in case of non-tunable resistor designs. The edge of the stochastic region $v^{\pm}$ is defined by when $V_i = V_{DD}/2 - [I^+R_P, I^-R_{AP}] \approx 0$ where the current $I^{\pm}$ is determined by the NMOS as shown in fig. 3.6(c). For a desired $\Delta v = v^+ - v^-$

(stochastic region) and NMOS transistor, the required n = R_{AP}/R_P should approximately equal I^+/I^-. Ideally, the minimum value of the resistance should be $R_P = (V_{DD}/2)/I^+$ and to get full pinning, Δv should be less than V_{DD}. For a 14nm FinFET, to get a stochastic region of $\Delta v = 50 - 200$mV, the resistance ratio n should be around $2 - 50$. The resistance ratio n is a measure for tunneling magneto-resistance, TMR (= (n − 1) × 100%) in case of MTJs. Typically MTJs have TMRs ranging from 100−300% [113] with a maximum reported TMR of 604% [114], so the resistance ratio of MTJs are well within the desired range, but the general requirements we outline should be applicable for other types of stochastic resistors as well.

Effect of I_0: In case of tunable resistances, the stochastic region is independent of the resistance ratio and depends on the pinning current and thus the bias current ($I_P^{\pm} \propto I_0$) instead as shown in fig. 3.6(d). For large bias currents ($I_0 \gg I$), the tunable resistances act essentially like non-tunable resistances. To get the full range of R,the NMOS needs to be able to supply the pinning current. If the pinning current is $(3 - 5)I_0$ as shown in fig. 3.2, then to get the full range of the resistance I^+_{Pmax} needs to be around $\sim (6 - 10)I_0$. In case of 14nm FinFETs, I^+_{max} is around $\sim 40\ \mu A$, restricting I_0 to values less than $7\ \mu A$.

Choice of I_{50}: Another parameter that is important for the operation of tunable resistors is the I_{50} which determines the midpoint of the sigmoid. I_{50} is the current at which the resistance on average spends equal time in R_P and R_{AP} states [104]. As the circuit can only support positive current values, it needs to be a positive quantity and preferably matched with the saturation point ($V_{DS} = V_{GS}$) current I_{Dsat} of the NMOS transistor. Changing I_{50} shifts the transfer characteristics laterally as shown in fig. 3.7(a).

R vs I: One last requirement is that, for current tunable resistance with increasing current I, the resistance needs to increase from $R_P \rightarrow R_{AP}$. This can be understood intuitively: Increasing I means the NMOS transistor is becoming more conductive. If the MTJ concomitantly becomes more conductive as I is increasing, the transfer characteristics can show non-monotonic behavior as shown in fig. 3.7(b). This requirement holds true irrespective of whether the circuit's R branch consists of a PMOS-1R or 1R-NMOS topology.

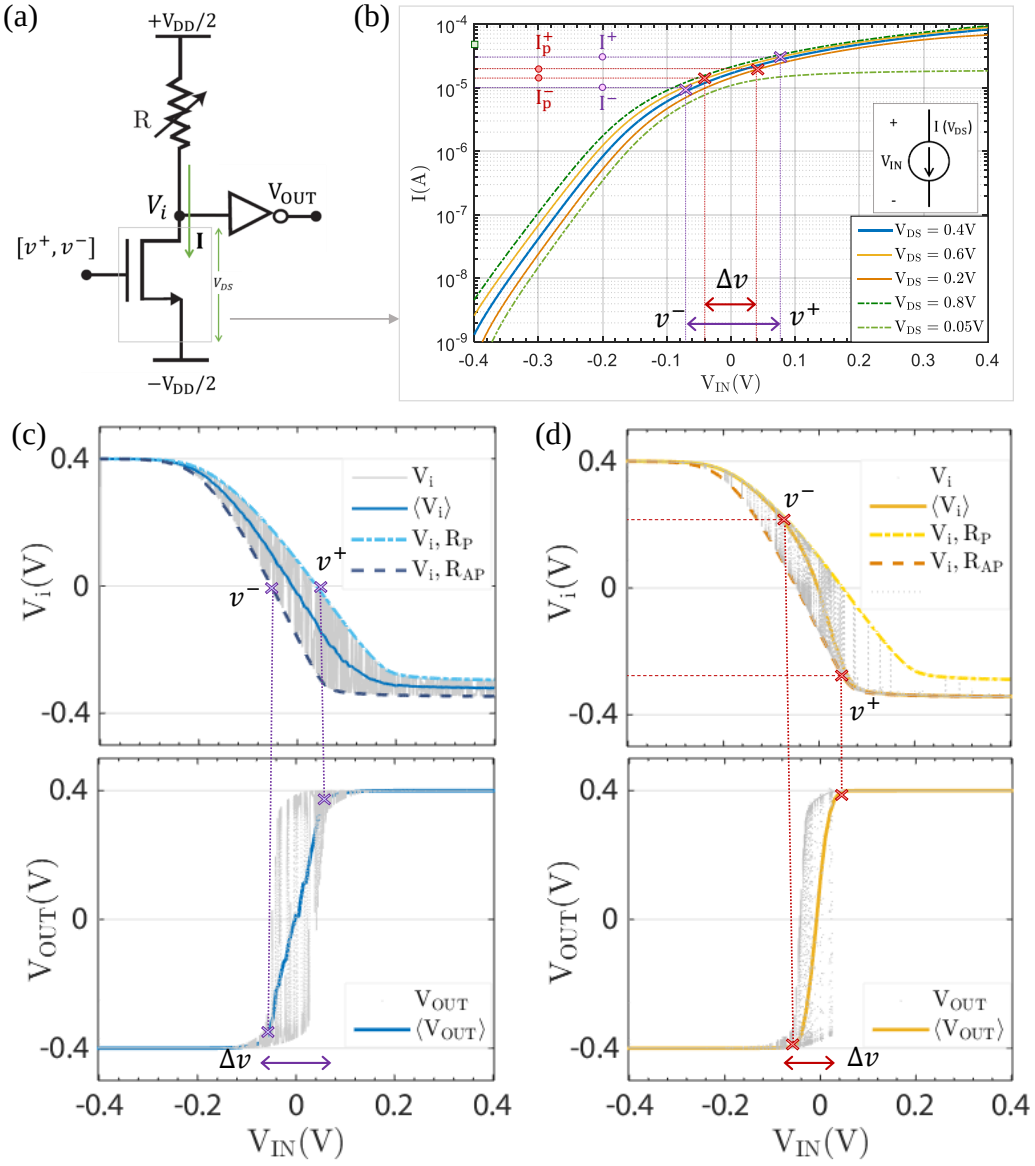


Figure 3.6. Stochastic Region boundaries : The stochastic region boundaries $[v^+, v^-]$ are set by different parameters for tunable and non-tunable resistors. (a) Shows the BSN circuit with (b) the current transfer characteristics of the 14nm FinFET NMOS when $V_i \sim 0$V. (c) Non-tunable R : In this case the boundaries are set by when $V_i \approx 0$ when resistance ratio $n = R_{AP}/R_P \approx I^+/I^-$. (d) Tunable R : The stochastic range is determined by pinning current I_P characteristics of the resistance. The transfer characteristics of each stage in (c) and (d) indicates the stochastic range v^+ and v^- and the relation to the NMOS characteristics in each case in (b).

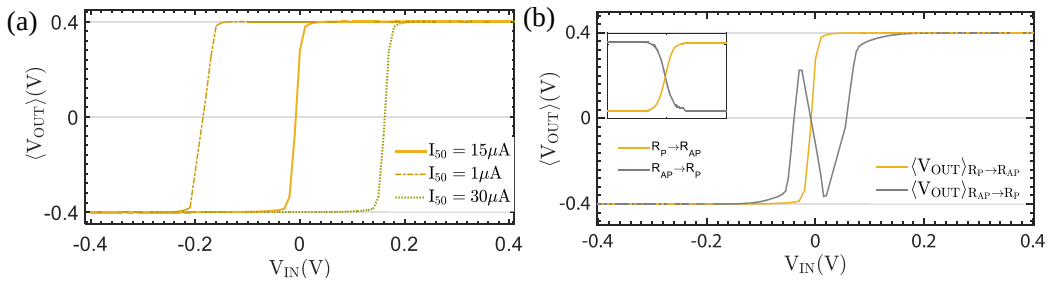


Figure 3.7. (a) **Choice of I_{50}:** I_{50} is ideally a positive quantity matched with the I_{Dsat} of the transistor, changing I_{50} results in a lateral shift of the sigmoid. (b) ***R* vs *I* relationship:** The output characteristics also depend on the nature of the resistance tunability with the circuit current I. If R decreases with I ($R_{AP} \rightarrow R_P$), the opposing characteristics of the transistor current and resistance change result in a non-monotonic output.

3.2 Realization of fluctuating resistances with sMTJs

A magnetic-tunnel-junction (MTJ) whose free layer is a low-barrier magnet (LBM) could serve as a physical realization of fluctuating resistors. Depending on the nature and characteristics of the LBM magnetization fluctuations, we can get different types of R. Our previous analysis [60] was restricted to one type of LBM, the circular IMA with barrier $< k_BT$, in this section we extend it to include all possible LBMs.

A general description of the energy associated with a magnet is given by [60]:

$$\begin{aligned} E = &\frac{1}{2}H_{kp}M_s\Omega(1-m_x^2) + \frac{1}{2}H_{ki}M_s\Omega(1-m_z^2) \\ &- \hat{H}_{ext}M_s\Omega \cdot \hat{m} \end{aligned} \tag{3.3}$$

where, $H_{kp} = 2K_s/t - 4\pi M_s$ is the perpendicular anisotropy field along the x-axis, K_s is the surface anisotropy density, H_{ki} is the in-plane anisotropy along z-axis, H_{ext} is the external field, M_s is the saturation magnetization and $\Omega = \pi(D/2)^2 t$ is the volume of the magnet. By adjusting the thickness or the shape of the magnet, the magnetic anisotropy of the magnet can be scaled to behave like a low-barrier magnet [57], [60]. We use the stochastic LLG module from our spintronics library [58] to simulate the LBM dynamics. This model has been carefully benchmarked against general Fokker-Planck based methods [99].

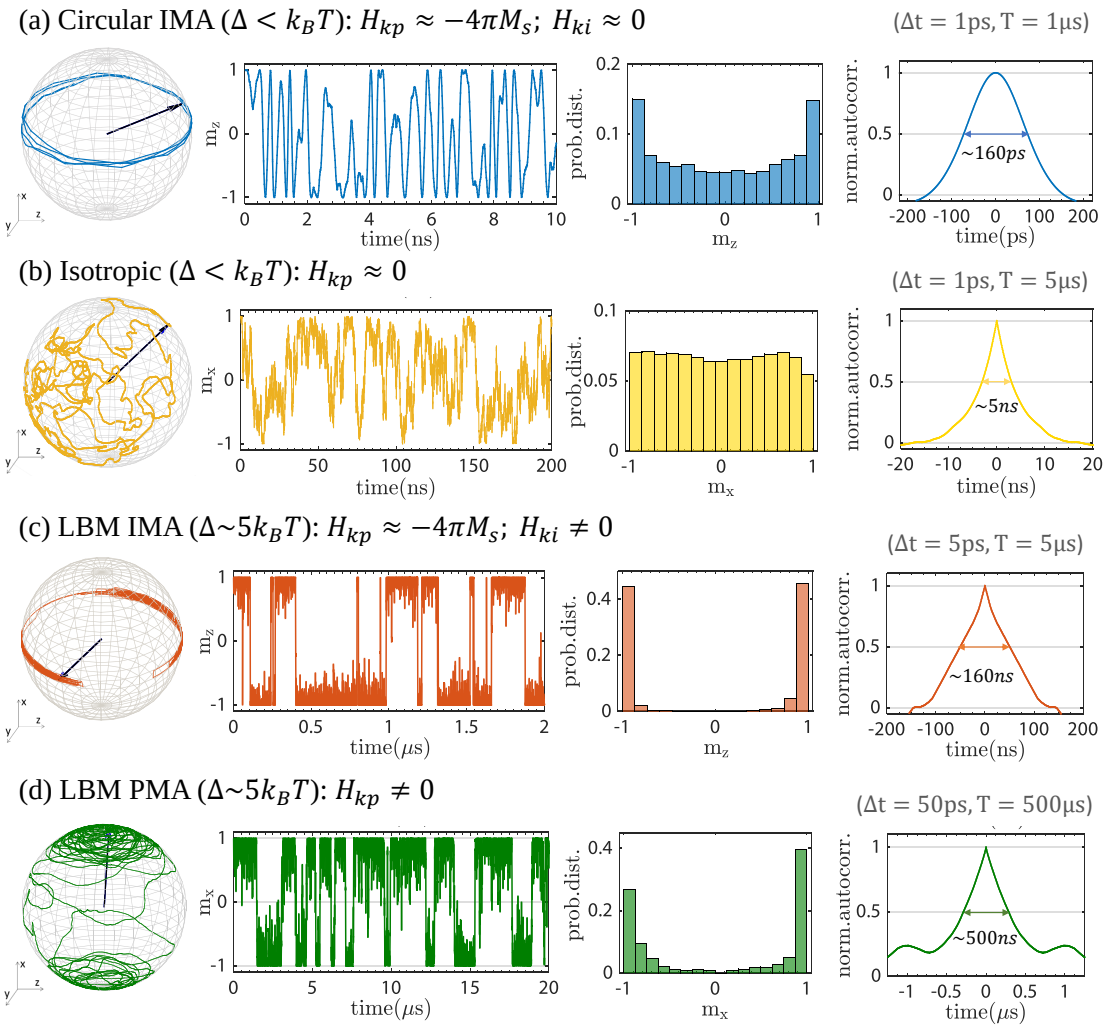


Figure 3.8. Low-barrier magnet fluctuation dynamics: We use the benchmarked stochastic LLG module to simulate LBM dynamics. Each simulation is carried out with a time-step at least ×100 smaller for a time-duration ×1000 than characteristic timescales to avoid any simulation time dependencies, the exact parameters are indicated. $\Delta < k_BT$ magnets have more continuous fluctuations with (b) having a more uniform distribution than (a) while slightly higher barrier magnets have a more telegraphic fluctuation. In both cases, the presence of high demagnetization fields cause faster fluctuations in IMA magnets.

LBM Fluctuation Dynamics: By low-barrier magnet we refer to magnets whose barrier is $< 10k_BT$ or so, whose magnetization fluctuates randomly in presence of thermal noise. Interestingly, the magnetization dynamics of low-barrier magnets with barrier $< k_BT$ are different from those with a slightly higher barrier [60], [83]. The simple exponential dependence of retention time of the magnetization state on the barrier height is not valid around or below k_BT [82].

Fig. 3.8 shows the fluctuation dynamics, the magnetization distribution, and the autocorrelation time (τ_{CORR}) for low barrier magnets. Magnetization fluctuations translate into resistance fluctuations in MTJ, and we see that magnets with barrier $< \mathrm{k_BT}$ act like continuous resistances, while slightly higher barrier magnets, which have a more defined two states, give telegraphic fluctuations, and in both cases IMA magnets fluctuate orders of magnitude faster than their PMA counterparts due to a new mechanism where the demagnetization field plays a central role [60], [61], [83], [115], [116].

Current Response of LBM: Magnetic fluctuations can be tuned by spin-current. For high barrier magnets, the minimum current required to switch the magnetization is called the critical current [84], in case of low-barrier magnets, we refer to it as a biasing current, defined by the inverse of the derivative taken at $\langle \mathrm{m} \rangle = 0$, mathematically expressed as: $\mathrm{I_0} = (\langle \mathrm{m} \rangle / \mathrm{I_S})^{-1}$ at low bias ($\mathrm{I_S}$). The current required to pin the magnetization, similar to switching current in high-barrier magnets is assumed to be $\sim 3-5\ \mathrm{I_0}$, as indicated in fig. 3.2. IMA magnets have a much larger pinning current than PMA magnets because of the large demagnetization field present due to their disk shape [35], [60], [84], meaning transistors with much larger current ranges would be required for IMA magnet MTJs than PMA for tunable resistors.

An important thing to note here is the current tunability in presence of an external field which can arise, for example, due to the fixed, stable layer that acts as a reference to the free layer in the MTJ. In the case of high-barrier magnets, the spin-current induced magnetic switching hysteresis loop just shifts in case of PMA magnets depending on the direction of field, but for IMA magnets the shape of the hysteresis and magnet dynamics is changed [84]. The large demagnetizing field present perpendicular to the magnetization plane in

IMA magnets causes the magnetization to precess around it when spin-current is applied in the opposite direction to the external field. The same is observed in low-barrier magnets as shown in fig. 3.9. The larger the external field the more pronounced the effect is. The uniform precessional motion kicks in at high-field, when the current is close to the biasing current or higher applied in the opposite direction to the field. Very recently, this has been observed experimentally for low fields [116]. While this is an undesired effect in case of our BSN operation, this can be useful in context to oscillator based networks [117].

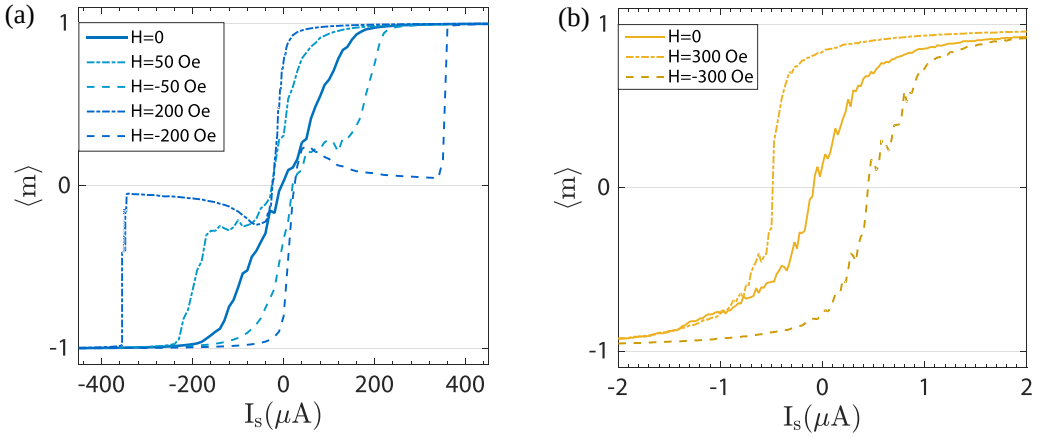


Figure 3.9. Current Response of LBM: LBM response to spin-current with and without external-fields for (a) circular IMA magnet ($H_{ki} \sim 0, H_{kp} \sim -H_D$) and (b) isotropic anisotropy magnet ($H_{kp} \sim 0$). Each point on the curve is a long-time ($T = 1\mu s$, $\Delta t = 1$ps) average magnetization from our benchmarked sLLG module. The critical field for IMA magnet was ~ 130Oe and for isotropic magnet ~ 200Oe.

This has important implications in terms of acting as a fluctuating resistance in a BSN circuit. IMA magnets with external fields (i.e. uncompensated dipolar fields in MTJ [118]) greater than its pinning field is not suited to function as a tunable or non-tunable resistor. IMA magnets with continuous magnetization coupled to a transistor with small saturation current (tens of μA) compared to the biasing current of IMA (hundreds of μA) can work as non-tunable resistors, and as experimental observations in ref. [116] suggest, it can withstand small (compared to its pinning field) stray fields.

PMA magnet MTJs with their small biasing current ($\sim$ few to few tens of μA) when coupled to typical transistors act as tunable resistors in BSN circuit. In this case the external bias field is actually preferred, since this enables positive I_{50} current [38].

So, if we coupled an MTJ with a 14nm FinFET ($V_{DD} = 0.8$ and $I_{Dsat} = 15\mu A$) [59], the table in fig. 3.10 summarizes the resistance mapping and the associated parameters.

R Type	MTJ Free Layer	τ_{CORR}	I_0	I_{50}	H_0
Non-tunable Continuous	$\Delta < k_BT$ *Circular IMA*	$\sqrt{8\ln(2)}\,\frac{1}{\gamma}\sqrt{\frac{M_s\Omega}{H_D k_B T}}$ ($\alpha < 0.1$) (sub-ns)	$\frac{2q}{\hbar}\sqrt{\frac{2}{\pi}}\sqrt{H_D M_S \Omega k_B T}$ ($\alpha < 0.1$) (0.1~1mA)	0 (n/a)	$\frac{2k_BT}{M_S\Omega}$
Tunable Continuous	$\Delta < k_BT$ *Isotropic 'PMA'*	$\ln(2)\,\frac{1}{\gamma}\frac{M_s\Omega}{\alpha k_B T}$ ~10 ns	$\frac{6q}{\hbar}\alpha k_B T$ (0.4~4 μA)	$\frac{4q\alpha}{\hbar}\left(\frac{1}{2}H_{ext}M_s\Omega\right)$	$\frac{3k_BT}{M_S\Omega}$
Non-tunable Bipolar	$2k_BT < \Delta < 10k_BT$ *IMA*	$\propto \frac{e^{\Delta/k_BT}}{(1+H_D/2H_k)}$ 1 ns ~ 1 μs	$\frac{4q\alpha}{\hbar}\Delta\left(1+\frac{H_D}{2H_K}\right)$ (0.05~25mA)	0 (n/a)	$\sim H_k$
Tunable Bipolar	$2k_BT < \Delta < 10k_BT$ *PMA*	$\propto e^{\Delta/k_BT}$ 0.1~ 100 μs	$\frac{4q\alpha}{\hbar}\Delta$ (0.5~25μA)	$\frac{4q\alpha}{\hbar}\left(\frac{1}{2}H_{ext}M_s\Omega\right)$	$\sim H_k$

Figure 3.10. Characterization Table: MTJ Free layer and its corresponding R type along with corresponding characteristic parameters and their analytical expression. The numbers in bracket indicates an approximate range of values for each parameter. The proportionality constant for correlation time of magnets with $\Delta > k_BT$ is $\tau_0 \sim 0.1 - 1$ ns, exact equation can be found in [82].

3.3 Performance Evaluation of sMTJ based BSN

In the final section we compare the physical performance of these different sMTJs in a BSN and project how Ising Machines built with such devices would perform in contrast to digital annealers of today.

3.3.1 Device-Level Performance Evaluation

Timescale of Operation: The two relevant timescales of operation for a BSN are, the correlation time τ_C which is the average time it takes to produce new output at given input and the response time τ_N which is defined as the average time it takes for the circuit to

give a random output with correct statistics as the input is changed [60]. fig. 3.11 shows the two timescales for the three types of fluctuating resistances for MTJs with two different timescales. For simplicity we assumed the correlation time to be same for all types of magnets, but in reality they would follow the τ_{CORR} relations indicated in fig. 3.10 [60], [83].

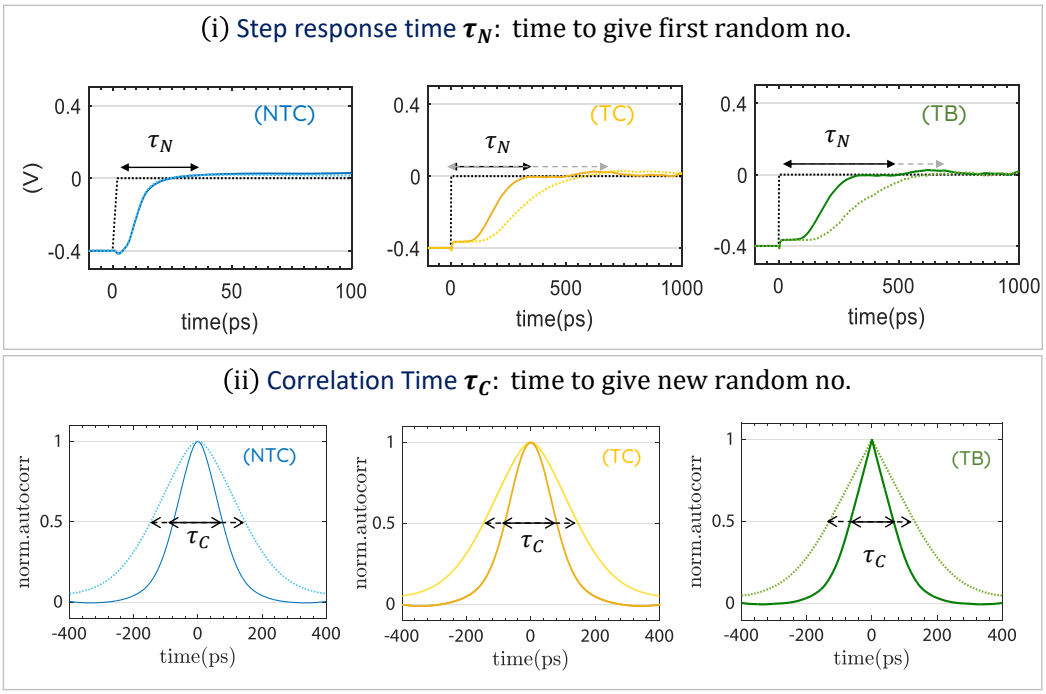


Figure 3.11. Timescale of Operation for each resistor type with two fluctuation rates $\tau_{\mathrm{C}} \sim [160\ \mathrm{ps}, 320\ \mathrm{ps}]$. The resistances are engineered to have similar characteristic timescales but different fluctuation behavior (tunable, non-tunable and continuous and bipolar fluctuation) for comparison purposes.

Fig. 3.11 shows that the response time, τ_N for non-tunable resistor is independent of the fluctuation time of the resistance, it is rather proportional to the RC delay of the circuit. While for the tunable cases, the response time is related to the characteristic timescales of the resistor. But the time to give new numbers or flip rate τ_C at $\mathrm{V_{IN}} = 0$ is entirely resistance fluctuation time dependent for all cases ($\tau_C \approx \tau_{CORR}$). So for the tunable case, the two said timescales of operation are likely to be similar as they are governed by the magnet fluctuation characteristics while for the non-tunable case, the response time which is RC dependent has the potential to be very short compared to the magnet dependent correlation time. For most applications this difference may not be of importance but for some applications where the

network is directed, like Bayesian inference having two different timescales seems to be a requisite [119].

Power: Our SPICE simulations indicate that the average power consumed by the BSN circuit is $\langle P \rangle \approx 2 \times V_{DD} I_{Dsat}$ [60]. The 2 is for the two branches, the MTJ branch and the inverter branch. This holds true for all types of resistors. For a 14nm FinFET with $V_{DD} = 0.8V$ and $I_{Dsat} \sim 15\mu A$, $\langle P \rangle \sim 20\mu W$. The MTJ branch power could be reduced by operating in subthreshold region $I_{Dsub} \sim 1\mu A$, but this reduces the total power by $\times 0.5$ while trading-off with an $\times 10$ increase in the RC response time. Given the flexibility, it is preferable to design the MTJ to operate in the saturation region of transistor. For tunable case this means matching $I_{50} \sim I_{Dsat}$, for non-tunable this means having $\langle R \rangle \approx (V_{DD}/2)/I_{Dsat}$.

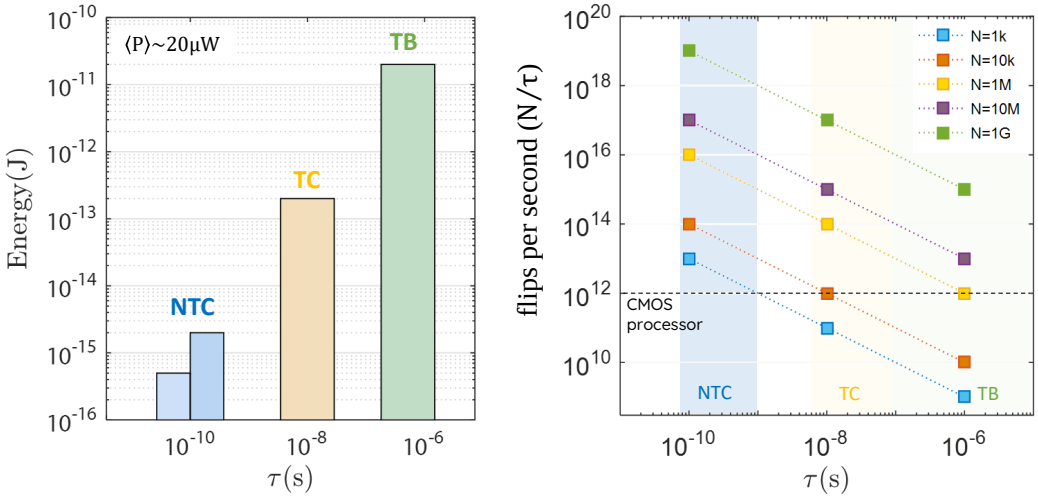


Figure 3.12. (a) Energy-Delay of each type of MTJ based BSN assuming an average power of 20 μW and timescales in fig. 3.9. (b) Plots the fps for different no. of neurons for each type of MTJs. For the projections only BSN performance numbers are used, synapse would add to the power and thus energy per flip number.

Energy: As there are two timescales associated with the BSN operation, we can define two energy as well, the energy to give first random number $E_N \sim \tau_N \langle P \rangle$ and the energy expanded between producing a new random number $E_C = \tau_C \langle P \rangle$. fig. 3.12(a) shows an energy delay plot indicating the ranges for each type of MTJs. When describing the energy-delay performance of BSN instead of quoting two numbers, we quote the larger number

which is the correlation time τ_C and the energy E_C. The individual energy-delay numbers can be used to project performance parameters for processors built with them.

3.3.2 Hardware Projections:

Typically the performance of an Ising hardware is measured in terms of time and energy it takes to solve a specific problem. Time to solution depends not only on the physical hardware performance but also on the algorithm that is being implemented. Here, we emphasize measuring the hardware performance in terms of a purely hardware metric *flips per second* (fps) [24], [48], [120], which refers to the maximum number of spin configurations the hardware can cycle through per second. It depends on the number of spins in the system (N) and the time it takes for a spin to flip (τ), $f = N/\tau$.

For the digital annealers the spin update time is usually determined by its clock period (τ_{clk}) which ranges typically in tens of ns range. To ensure fidelity simultaneous updates of connected spins needs to be avoided [121] forcing digital annealers that operate on clock edge to update spins sequentially. So in a network where all spins are connected effectively only one spin can update per clock cycle [27]. But it need not be if some spins are unconnected (i.e. nearest neighbor [24], [25], or king-graph [26] connection, or if spins are parallelized by implementing special algorithms [28]–[30]. Based on the reported total spin number and clock speeds of digital annealing hardware today which have about $\sim$ 10K neurons that can update per $\sim$ 10ns clock period, we derive an estimation of their performance at f $\sim 10^4/10^{-8} = 10^{12}$ flips per second [25], [48] as shown in fig. 3.13.

Compared to digital annealers the Ising spin hardware we presented in this work can work autonomously, i.e, without a synchronizing clock or a sequencer [36], [48], [119]. In this mode, the speeds are governed by neuron (τ_{neu}) and synapse (τ_{syn}) time only, and to ensure fidelity and avoid simultaneous updates of connected BSNs the synapse needs to update faster than the the neuron ($\tau_{\text{syn}} < \tau_{\text{neu}}$). Sutton et. al. [48] defines a metric s $= \tau_{\text{syn}}/\tau_{\text{neu}}$ and shows that with s < 1 ensures fidelity of operation, the exact requirements are problem and architecture dependent. Memristive crossbar arrays paired with a fast summing amplifier synapse could operate very efficiently at as low as few tens of ps speeds [46], [51], [96]–[98],

	Affiliates	BIFI	Hitachi	Fujitsu	Tokyo Tech.	UC Berkeley	Purdue
	Name	**Janus II**	**annealing machine**	**Digital Annealer**	**STATICA**	**RBM-based**	**Purdue-P (ApC)**
[reported]	Technology	FPGA	40nm CMOS + FPGA	65nm CMOS	65 nm CMOS	FPGA	FPGA
	Latest	2014	2019	2018	2020	2020	2020
	Connectivity	Local (5,N-N)	Local (8,King's Graph)	All-to-All	All-to-All	All-to-All	Local (5,N-N)
	Total Neurons, **N**	2,000	30,000	1024	512	150	8,100
	Parallel Neurons $\mathbf{N_p}$	N/2= 1,000	N/4= 7,500	1	N=512	N=150	N=8,100
	Clock Frequency, f	250 MHz	100 MHz	100 MHz	320 MHz	70 MHz	125 MHz
	Weight Precision	1 bit	3 bit	16 bit	5 bit	9 bit	16 bit
[derived]	*Neuron Time (MC step)* $\tau = 1/f$	4 ns	10 ns	10 ns	~3 ns	14 ns	32 ns
	flips per second (N_p/τ)	2.5×10^{11}	7.5×10^{11}	10^8	$\sim 2 \times 10^{11}$	10^{10}	$\sim 2.5 \times 10^{11}$

Figure 3.13. ***flips per second (fps)*** is a substrate and algorithm independent performance metric for simulated annealing processors much like the flops per second metric used for general purpose computers. It is a measure of how many flips, and hence spin configurations the system can cycle through in a second. fps can be derived from the reported performance metrics of the processors following ref. [48]. The reported and derived quantities as indicated. Current CMOS based annealing processors perform at $\sim 10^{12}$ fps. We project that MTJ based hardware can increase by a few orders of magnitude.

[122]. The digital annealers mimic the Ising spin using a combination of random-number generators (LFSR, Xoshiro, etc.), look-up-tables (LUT) and comparators. The random number generator (RNG) unit is one of the most are expensive elements in the design [123]. Even in the most optimized design, the RNG unit take up $\sim 11\%$ of the total logic gate area [28]. The 3T-1MTJ design offers drastic reduction in the area footprint, promising massive scalability leveraging existing 1T-1MTJ Magnetic RAM technology that already has 1Gbit integrated cells [53], [124].

Fig. 3.12(b) projects *fps* number considering $\tau \equiv \tau_{\text{neu}} \approx \tau_{\text{CORR}}$ for different no of spins, N. An MTJ realization with circular IMA, with $\sim$ ns timescale can offer almost two orders of magnitude speedup with $<$ 10k neurons. If spins are implemented in Gbit densities all stochastic implementations seem to outperform the CMOS implementations. For such systems the upper bound for N is ultimately determined either by area or by power budget of the chip. Note that the fps number does not reflect the connectivity of the spins or the algorithm implemented by the hardware. It also does not indicate he solution accuracy obtainable for specific problems [125]. What we highlight here is that using the natural physics of the MTJ we can design a very compact realization of eq. 3.2 compared to current

state of the art CMOS implementations, and despite being a magnetic circuit, low barrier magnet implementations even offer an overall speed up due to their fast fluctuation rates.

3.4 Summary

In this chapter, we presented a comprehensive evaluation of naturally stochastic magnetic building blocks for implementing probabilistic algorithms compactly and efficiently. We generalized the proposed 1MTJ-3T design to a 1SR-3T design and presented necessary design rules for BSN operation that we hope will stimulate further interest in finding stochastic resistance (1SR) with suitable properties. We extended the physical performance analysis of the 1MTJ-3T BSN design to include unstable MTJ's with different low-barrier-magnets as free layers. They are evaluated as physical realizations of the general stochastic resistor (SR) with respect to 14nm FinFET transistors. IMA magnets with barrier $\leq k_BT$ proved to be the best option, low-barrier PMA can function as current-tunable resistors as well. While careful optimization of the fixed layer to cancel the stray fields in IMA MTJ is preferred, PMA can benefit from the presence of stray fields (can be a source of the I_{50}). The most challenging set of working conditions are set for telegraphic IMA magnets, even if they are highly optimized and no stray fields are present in the circuit, they need to be coupled with high current transistors due to their high pinning currents, because if paired with low current transistors like 14 nm FinFET results in a staircase-like functional behavior which does not work as a p-bit as we discussed.

These BSNs are an integral part of Ising machines which are often referred to as annealing processors. Using 1MTJ-3T BSN could speed up the operation of these processors by orders of magnitude. Another important application space for these BSN is stochastic neural networks [36], [43], [126], [127]. Infact, binary stochastic neurons are desired for deep learning networks, but are typically avoided because it is harder to generate random bits in CMOS hardware [128]. Use of this compact neuron that relies on MTJs natural physics to provide stochastic binarization could accelerate computation in custom hardware [129], [130] by faster evaluation of BSN function [60] and also encourage algorithmic advancement using BSN.

4. REALIZATION OF WEIGHTED p-BIT

Most of the materials in this chapter have been extracted verbatim from the paper: " Voltage-driven Building Block for Hardware Belief Networks", O. Hassan, K. Y. Camsari and S. Datta, published in IEEE Design & Test, vol. 36, 2019 [39].

There are two equations (eq. 1.2 and eq. 1.3) that constitute the behavioral model of probabilistic spin logic (PSL) framework. So far we have focused mainly on the hardware realization of eq. 1.2. Eq. 1.3 could be implemented on software or hardware to enable p-circuit operation. In this chapter we present a complete hardware building-block (weighted p-bit) that combines the functionality of both the equations in a single composite unit. We propose augmenting the embedded s-MTJ MRAM structures presented in Chapter. 2,3 with floating-gate MOS (FGMOS) based capacitive adder [131] with the embedded s-MTJ MRAM structures [77]. We show a hardware mapping and demonstrate how the results of a fully interconnected wp-bit circuit closely approximate the ideal PSL equations using an example of an "invertible" full-adder (FA) that can perform 1-bit addition and subtraction. We also show how such invertible FAs can be interconnected to solve a simple instance of the NP-complete subset sum problem (SSP). The examples in this chapter has been obtained using full-SPICE models that simply uses transistors, capacitors, and resistors without any additional complex circuitry or processing.

4.1 Weighted p-bit Building Block

The PSL model is defined by two equations:

$$m_{\mathrm{i}}(t+\Delta t) = \mathrm{sgn}\{\mathrm{rand}(-1,1) + \tanh(I_{\mathrm{i}}(t))\} \tag{4.1a}$$

where rand(−1,+1) is a random number uniformly distributed between −1 and +1, and t is the normalized time unit. The synapse generates the input I_{i} from a weighted sum of the states of other p-bits according to the relation

$$I_{\mathrm{i}}(t) = I_0\left(h_{\mathrm{i}}(t) + \sum_{\mathrm{j}} J_{\mathrm{ij}} m_{\mathrm{j}}\right) \tag{4.1b}$$

where, h_i is the on-site bias and J_{ij} is the weight of the coupling from j^{th} p-bit to i^{th} p-bit and I_0 is a dimensionless constant. These two equations constitute the behavioral model of PSL. The objective of this chapter is to present a voltage-driven hardware building block using present day device technologies such as embedded MRAM [110] and Floating-Gate MOS transistors, such that identical copies of the same block can be interconnected with wires to implement Eqs. 4.1.

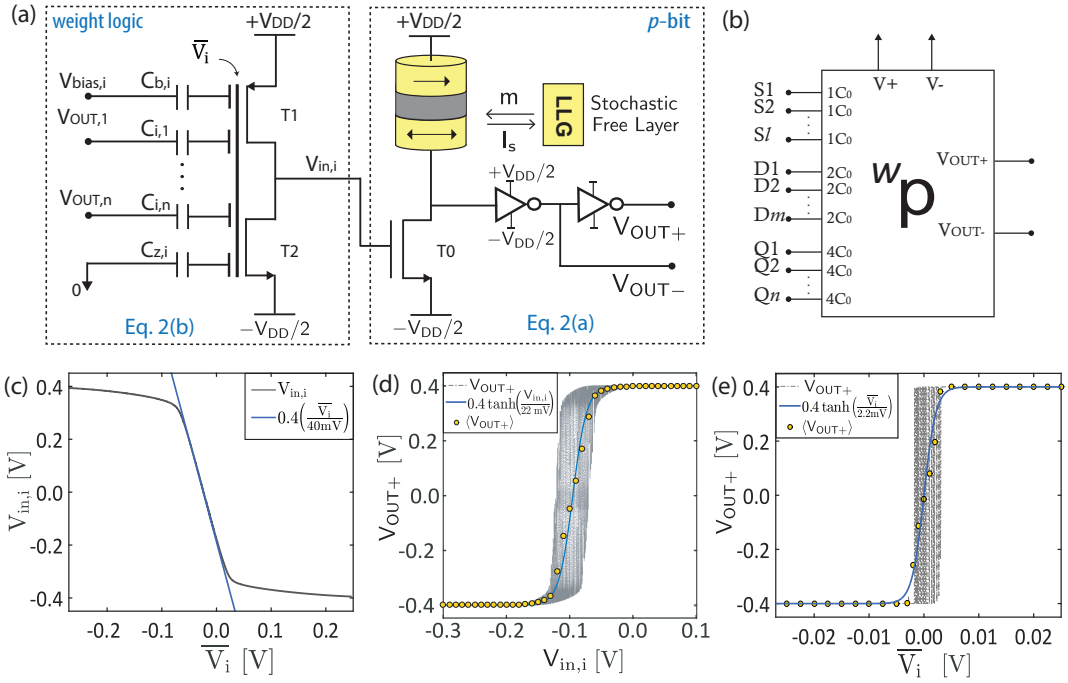


Figure 4.1. (a) **Weighted p-bit ($^W p$-bit)** has two components. The first is the p-bit implemented through an embedded s-MTJ with two inverters added to give positive and negative outputs. The second is the capacitive voltage adder with an inverter structure on the left similar to floating gate MOS transistors. (b)Shows the the block diagram of $^W p$-bit. (c) Shows how an inverter helps amplify the input (V_i) of the capacitive network to give $V_{in,i}$ at the gate of the p-bit's NMOS transistor T0. (d) Shows the relation of the input gate voltage of the NMOS ($V_{in,i}$) to output (V^+_{OUT}). (e) Shows the transfer characteristics of the $^W p$-bit as a whole. The inputs in each case is swept from -0.4V to $+0.4$V in 1 μs. The yellow dots are time averaged values at each point over 300 ns and the solid blue lines are numerical fits.

Our building block has two components corresponding to the two eq. 4.1a,b.Eq. 4.1a is implemented by the p-bit in fig.4.1a which consists of an embedded low-barrier unstable MTJ

coupled to two CMOS inverters which provides a stochastic output whose average value is controlled by the input voltage:

$$V_{out,\mathrm{i}} = \frac{V_{DD}}{2}\mathrm{sgn}\left(\mathrm{rand}(-1,+1) + \tanh\frac{V_{\mathrm{in,i}}}{V_0}\right) \tag{4.2a}$$

where $\pm V_{DD}/2$ are the supply voltages, and V_0 is a parameter ($\sim$ 22 mV) describing the width of the sigmoidal response.

The value of V_0 depends on the details of the 1T/1MTJ in the embedded MRAM structure [77] and the transistor characteristics. The conductance, G_0 of the MTJ is chosen to match the MTJ switching characteristics to the transistors in the ^{W}p-bit so that the overall transfer characteristics is centered at zero as shown in fig. 4.1e. To do that, an input voltage of $\overline{\mathrm{V_i}}$ =0V is applied at the input of T1 and T2 transistors turning both of them ON ($|\mathrm{V_{GS}}| = 0.4\mathrm{V}$) and G_0 is swept to observe the outputs. The G_0 value for which $\mathrm{V^+_{OUT}}$=$\mathrm{V^-_{OUT}}$ = 0V is the value chosen to be the MTJ conductance. For minimum sized 14nm HP-FinFET transistors models with $\mathrm{V_{DD}} = 0.8\mathrm{V}$, $1/G_0 \approx$62 kΩ and it seems reasonable considering the RA-products of modern MTJs [74].

Eqs. 4.1b is implemented by the weighted synapse portion of fig. 4.1a , which is a capacitive voltage adder just like those used in neuMOS devices [131], [132]. We can write

$$\overline{V}_{\mathrm{i}} = \frac{V_{bias,\mathrm{i}}C_{b,\mathrm{i}} + \sum_{\mathrm{j}} V_{out,\mathrm{j}}C_{\mathrm{ij}}}{C_g + C_{z,\mathrm{i}} + C_{b,\mathrm{i}} + \sum_{\mathrm{j}} C_{\mathrm{ij}}} \tag{4.2b}$$

Note that the capacitive voltage divider typically attenuates the voltage $\overline{V}_{\mathrm{i}}$ at its output, and the inverter scales it up to $V_{in,\mathrm{i}}$ as shown in fig. 4.1c, the two being related approximately by

$$\begin{aligned} V_{in,\mathrm{i}} &\approx \frac{V_{DD}}{2}\tanh\frac{\overline{V}_{\mathrm{i}}}{\nu_0} \\ &\approx \frac{V_{DD}}{2\nu_0}\,\overline{V}_{\mathrm{i}} \quad \text{if} \quad \overline{V}_{\mathrm{i}} \ll \nu_0 \end{aligned} \tag{4.2c}$$

where ν_0 is a parameter characteristic of the inverter. Eqs. 4.2a,b can be mapped onto the PSL Eqs. 4.1a,b by defining

$$m_{\rm i} = \frac{V_{out,{\rm i}}}{V_{DD}/2}, \quad I_{\rm i} = \frac{V_{in,{\rm i}}}{V_0} \tag{4.3a}$$

$$C_{b,{\rm i}} = b_{\rm i} C_0 \quad C_{z,{\rm i}} = z_{\rm i} C_0 \tag{4.3b}$$

$$h_{\rm i} = b_{\rm i} \frac{V_{bias,{\rm i}}}{V_{DD}/2}, \quad J_{\rm ij} = \frac{C_{\rm ij}}{C_0} \tag{4.3c}$$

$$I_0 = \frac{(V_{DD}/2\nu_0)(V_{DD}/2V_0)}{(C_g/C_0) + z_{\rm i} + b_{\rm i} + \sum_{\rm j} J_{\rm ij}} \tag{4.3d}$$

C_g is the intrinsic gate capacitance of the neuMOS inverter. The significance of C_0 is that we assume the input is composed of many identical capacitors C_0, and that the weights $J_{\rm ij}$ have been designed to have *integer* values such that $C_{\rm ij}$ can be implemented by connecting $J_{\rm ij}$ elementary capacitors in parallel. The other coefficients $z_{\rm i}$, $b_{\rm i}$ are also integers. We adjust the number $b_{\rm i}$ of bias capacitors to facilitate external biasing and the number $z_{\rm i}$ of grounded capacitors to make $z_{\rm i} + b_{\rm i} + \sum_{\rm j} J_{\rm ij} = K$ a constant, so that I_0 is independent of index i:

$$I_0 = \frac{(V_{DD}/2\nu_0)(V_{DD}/2V_0)}{(C_g/C_0) + K} \tag{4.4}$$

Note that K is usually a fairly large number equal to the sum of all the weights, and to implement an $I_0 \sim 1$ it is important to keep the factor $(V_{DD}/2\nu_0)(V_{DD}/2V_0)$ to be much grater than 1. This is the reason for using an inverter between the capacitive voltage adder and the p-bit. Our model neglects any leakage resistances associated with the capacitive weights. Modern transistors with thin oxides can have gate leakage currents ~1nA, with RC $\sim \mu$s-ms. This should not affect the weighting, since the examples presented here operate at sub-ns time scales. For slower neurons, it may be advisable to use thicker oxides for the capacitive weights to ensure lower leakage.

Fig. 4.1b shows the icon we use to represent our building block which we call a weighted p-bit. The input consists of three types of inputs designated S, D and Q having capacitances C_0, 2 C_0 and 4 C_0. Combinations of these are used to implement different weights $J_{\rm ij}$ and different bias $h_{\rm i}$. Each block has two outputs V^+_{OUT}, V^-_{OUT}. The choice of output depends on

the sign of the corresponding J_{ij}. Similarly different signs of h_i are implemented by choosing $V_{bias,i}$ to be $+V_{DD}/2$ or $-V_{DD}/2$.

4.2 Invertible full adder

In PSL, any given truth table can be implemented using eq. 4.1 by choosing an appropriate $[J]$ and $[h]$ matrices [31]. Here we show how those $[J]$ and $[h]$ are mapped onto physical hardware using our proposed building block using only transistors, resistors and capacitances.

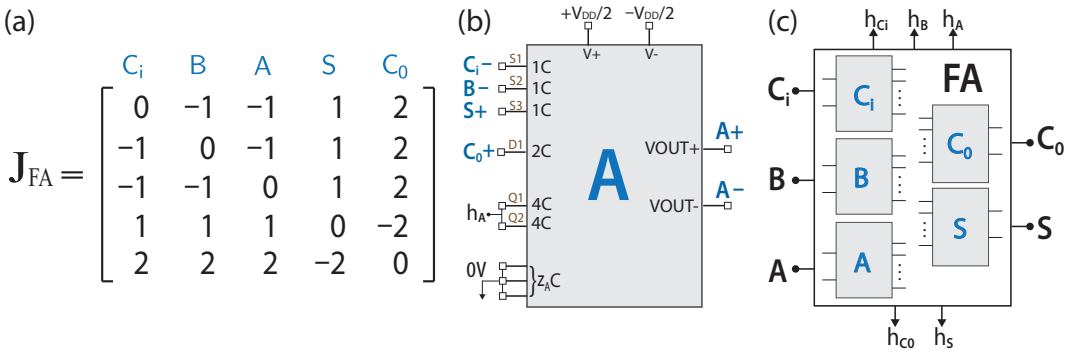


Figure 4.2. Invertible Full Adder with $^W p$-bit: (a)$[J]$ matrix for implementing a Full Adder. (b) Explicitly shows the hardware connections made to one of the inputs (A) from the other p-bits where $1C$, $2C$, and $4C$ represent capacitors in units of $C = C_0 = 100aF$. (c) Shows the subcircuit representation of the Full Adder with its input/output terminals. C_i, B, A input and S, C_o output read terminals and separate corresponding clamping terminals $h_{C_i}, h_B, h_A, h_S, h_{C_0}$. We used $8C$ for the clamping terminals to ensure input / outputs follow what is dictated by the external signals.

A Full Adder can be implemented in PSL using the $[J]$ matrix shown in fig. 4.2. In this chapter, we improve the 14 p-bit implementation of the invertible Full Adder (FA) in Ref.[31] and implement the same functionality using 5 p-bits. This is achieved by first noting that the first half of the FA truth table is complementary to the second half for the FA (fig. 4.3a inset)The first 4 lines in the truth table is turned into an orthonormal set by a Gram-Schmidt process and a [J] matrix is obtained using eq. 12 in Ref.[31] which is finally rounded to integer values, with diagonal entries replaced by zeros. This $[J]$ defines the interconnection between

the 5 ^{W}p-bits of the Full Adder in hardware. Each row of the $[J]$ matrix are realized in terms of capacitive coupling to the gate of the associated terminal.

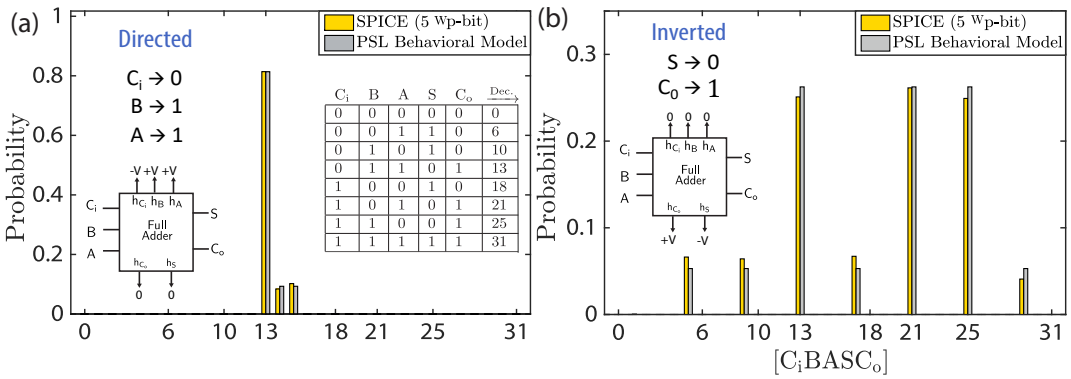


Figure 4.3. Full SPICE implementation of an Invertible Full Adder(5 ^{W}p-bit): The 5 ^{W}p-bit invertible Full Adder circuit is simulated in (a) Directed and (b) Inverted modes. The clamping values are indicated. All biasing terminals that are not clamped to 1 or 0 are grounded. The histogram of $[C_iBASC_0]$ is obtained after thresholding voltages $((V < 0) \equiv -1, (V > 0) \equiv +1)$. The SPICE model is run for $1\mu s$ and compared with the PSL equations where each p-bit is updated in random but sequential order [31]. In this example $I_0 \simeq 1$ is chosen to emphasize how the models are in good agreement even in the magnitudes of the minor peaks of the histogram.

To ensure a uniform I_0 is applied to each p-bit (eq. 4.4), the same weighting factor K needs to be used for all ^{W}p-bits. To apply a given I_0, we first find $\max(b_i + \sum J_{ij})$ for any given $[J]$, and then ground $z_i = M - b_i + \sum J_{ij}$ $(z_i \geq 0, z_i \in N)$ unit capacitances for all terminals where M is a number that can be used to control I_0, a larger M causing a smaller I_0. Fig. 4.2b shows explicit connections made to one of the inputs "A" and fig. 4.2c shows the subcircuit of the Full Adder with C_i, B, A as inputs, S, C_0 as the outputs, and $h_{Ci}, h_B, h_A, h_S, h_{Co}$ as the clamping pins.

Fig. 4.4 shows the operation of a Full Adder in the usual forward mode with C_i, B, A clamped to values (0,1,1) which forces the S and C_0 to (0,1) according to the truth table. In the invertible mode S and C_0 are clamped to (0,1) and the circuit stochastically searches *consistent* combinations of C_i, B, A to satisfy the truth table: $\{C_i, B, A\} = \{\{0,1,1\}, \{1,0,1\}, \{1,1,0\}\}$. Fig. 4.4 shows steady state (t = 1 μs) histogram plots of the

Full Adder operation in direct and inverted mode side by side with results from the PSL behavioral model.

The good agreement between the ideal PSL behavioral model and the coupled SPICE simulation that solves PTM-based transistors models with stochastic LLGs validates the hardware mapping of the ideal p-bit equations with the weighted p-bits.

4.3 3SUM Problem

3SUM is a decision problem in complexity theory that asks whether three elements of a given set can sum up to zero. A variant of the problem is when the set of three numbers have to add up to a given constant number. This problem has a polynomial time solution and is not in NP. In this section, we show how the invertibility feature of the Full Adders can be utilized to design a hardware 3SUM solver. In the next section, we show how the 3SUM hardware can be modified to design a general solver for the NP-complete Subset Sum Problem.

The invertibility property of the Full Adders ensure that given the sum, it can provide the possible input combinations for that sum as shown in fig. 4.4a. So an n-bit 3 number adder circuit implemented in PSL can essentially provide solution sets for the 3SUM problem when the sum is clamped to a given value.

Fig. 4.4a shows the circuit constructed out of Full Adders to solve a 4-bit 3SUM problem. Each of the Full Adders in the circuit are the 5 p-bit invertible adders that were shown in fig. 4.3. The first row of adders adds the two 4-bit numbers A and B, and feeds its output X, to the next row of adders which adds X and C to give the sum $S = C + X = C + B + A$. Because p-circuits are invertible, if we clamp the sum S, the circuit naturally explores through all possible sets and multisets of the set of all integers from 0 to $2^4 - 1$ that add up to S. The given set for the problem could be implemented through clamping certain bits of A,B and C or externally circuitry could be used to detect only the results that belong to the given set. Fig. 4.4b shows the how A,B,C is fluctuating between values that satisfy the clamped sum 15.

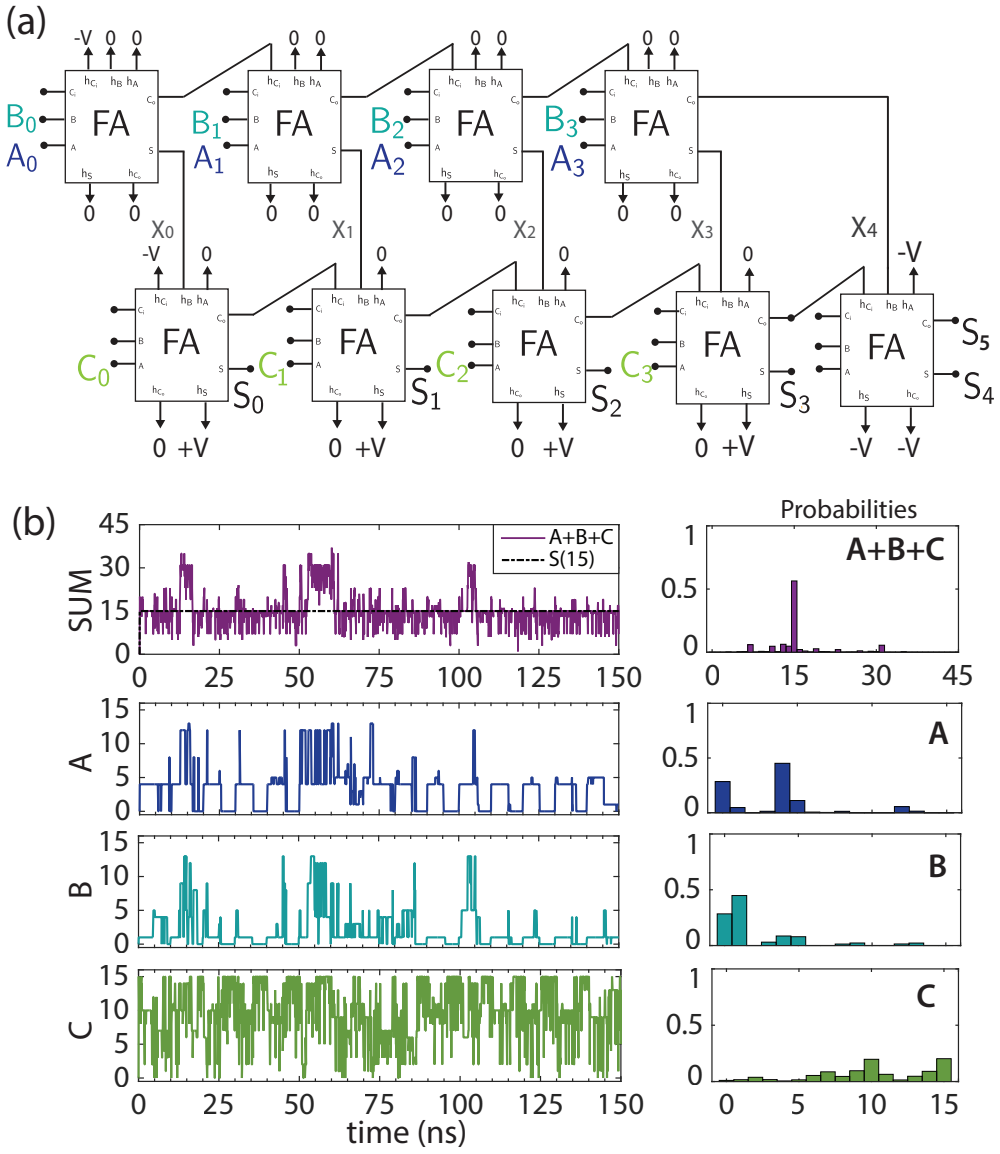


Figure 4.4. SPICE simulation of a 4bit 3-SUM Problem ($9 \times 5 = 45$ ^{W}p-bit network): (a) The circuit is constructed by interconnecting two rows of invertible Full-Adders (FA) to construct a 3 number, 4-bit adder. The sum S is clamped to the desired value and A, B, C resolves themselves to create all the possible 3 number subsets out of all positive numbers 0 to $2^4 - 1$ that satisfy $A + B + C = S$. (b) Shows the results when S is clamped to 15. A, B and C get correlated to satisfy the sum with different combinations. In this example, the inputs A, B, C are unconstrained and can take on any value between 0-15.

4.4 Subset-sum Problem (SSP)

In this section, we show how the hardware circuit that was designed for 3SUM problem could be modified to solve a small instance of subset-sum problem (SSP) [133] which is believed to be a fundamentally difficult problem in computer science (NP-complete). In the SSP, a set G with a finite number of positive numbers is defined. And then the decision problem is to ask whether there is a subset S' such that S' $\subseteq$ G whose elements sum to a specified target. For example, fig. 4.5 shows a circuit that is programmed to choose a set, G=$\{1, 2, 4\}$ and a target that is defined by 4-bits. In the 3SUM circuit the input bits (A, B, C) were left "floating", here, the inputs are constrained to a given number (1,2,4) by clamping the remaining bits of an input. For example, the inputs A_1 and A_0 are clamped to zero to make A either 4 or 0. Under these conditions, clamping the output to a specified target makes the circuit search for a *consistent* input combination to find a subset that satisfies the clamped target. Fig. 4.5c shows three example targets where the inputs get correlated to satisfy the clamped sum. The invertibility feature that is utilized to solve the SSP in this hardware is similar to those discussed in the context of memcomputing [134], however the physical mechanisms are completely different.

One striking difference in the design of the SSP we considered, compared to the 3SUM hardware is the *direction* of information. In 3SUM the connections were from the first layer of Full Adders to the second, as in normal addition (fig. 4.4a). In the SSP, we observed that reversing these connections from the second layer of adder to the first layer drastically improves the accuracy of the solution (fig. 4.5a). A similar observation regarding the directional flow of information for another inverse problem using p-circuits (integer factorization) was made in [31]. Here we have limited the discussion to a small instance of the SSP which would in general require more layers of Full Adders in both vertical and horizontal directions to account for more numbers of elements in S and their size. The purpose of this example is to illustrate how invertibility can be combined with standard digital VLSI design to construct any general "cost function" for hard problems of computer science in an asynchronously running *hardware* platform without any external clocking.

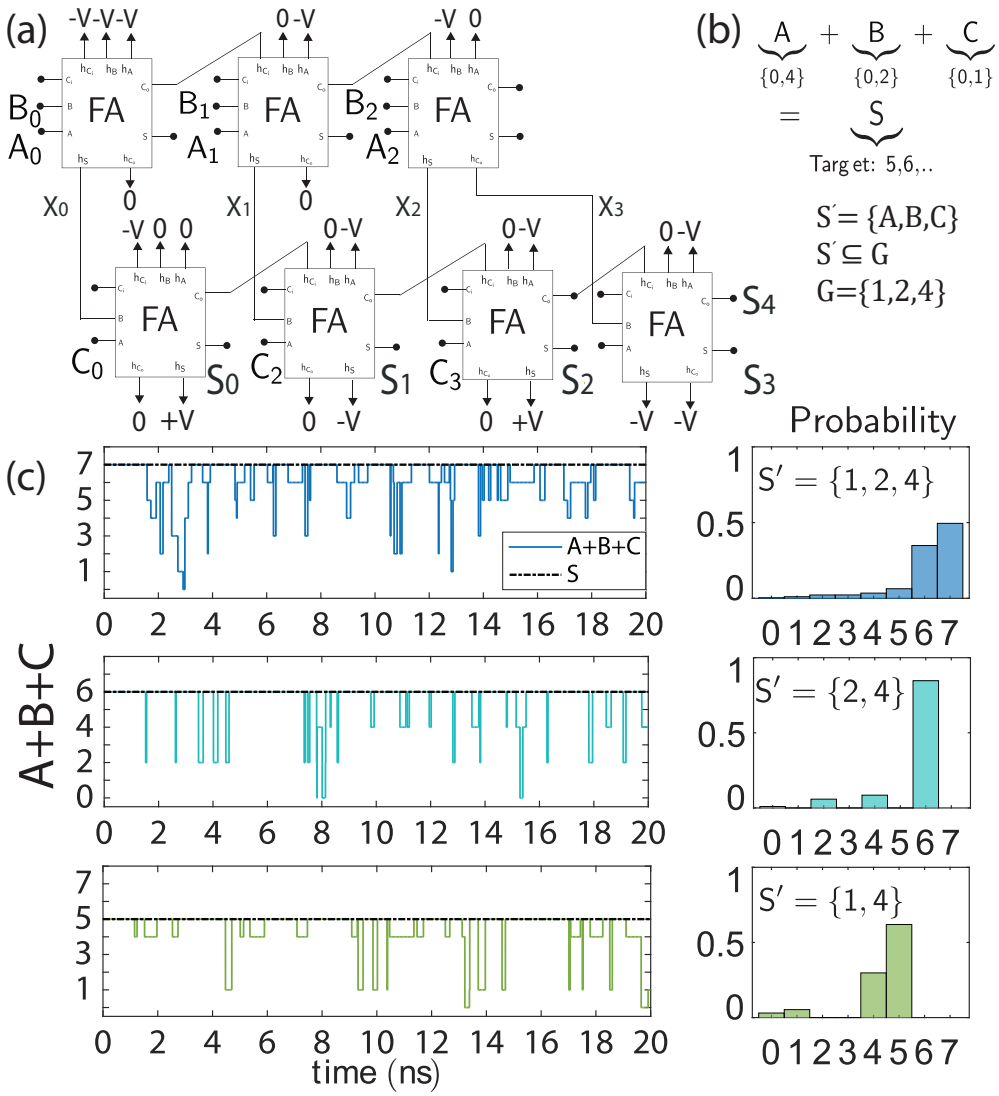


Figure 4.5. SPICE simulation of a 3 input, 3-bit Subset Sum Problem ($7 \times 5 = 35$ ^{W}p-bit network): (a) A 3-input 3-bit binary adder that adds three numbers A,B,C. Unlike the 3SUM, in this case inputs are constrained to a given value specified by the set G $=\{1, 2, 4\}$ in this example. A target S is selected and the output of the adders are clamped to the target value as shown in (b). (c) Shows three different instances of a target where the inputs find a consistent combination (the correct subset of G) to satisfy the target. Histograms show that the highest probable state is the correct subset. An important difference from the 3SUM circuit is that the information flow is *directed* from the target (second layer of adders) to the first layer of adders.

4.5 Summary

In this chapter we have proposed a compact building-block for Probabilistic Spin Logic (PSL) combining a recently proposed Embedded MRAM-based p-bit, with an integrated capacitive network that can be implemented using Floating Gate MOS (FGMOS) transistors similar to the neuMOS concept. We have shown by extensive SPICE simulations that the results of the hardware model for the weighted p-bit agree well with the behavioral equations of PSL. Having dedicated MTJ based hardware stochastic neurons could help minimize the footprint and consume lower power for applications as also indicated by ref.[74], [135]. Even though an FGMOS-based capacitive network for performing the voltage addition seems like a natural option for, we note that the device equations for any capacitance $[\mathsf{C}_{ij}]$ or conductance network $[\mathsf{G}_{ij}]$ would have been essentially the same. Moreover, our discussion was only about static weights, but an FPGA-like re-configurable weighting scheme can also be employed either by using transistor-based gates or by additional multiplexing circuitry to perform online learning or redesign p-circuit connectivity. Finally, using the basic building block we have shown how a small instance of the NP-complete Subset Sum Problem hardware solver can be designed using the unique invertibility feature of p-circuits.

5. MAGNETOELECTRIC MEMORY DEVICE BASED ON PSEUDO-MAGNETIZATION

Parts of the material presented in this chapter have been extracted verbatim from the paper: "Equivalent Circuit for Magnetoelectric Read and Write Operations", K. Y. Camsari, R. Faria, O. Hassan, B. M. Sutton, and S. Datta, published in Phys. Rev. Applied, 2018 [136], along with unpublished results.

In this chapter, we propose a new type of magnetoelectric memory device that stores magnetic easy-axis information or pseudomagnetization, rather than a definite magnetization direction, in piezoelectric/ferromagnetic (PE/FM) heterostructures. We show how a PE/FM combination can lead to non-volatility in pseudo-magnetization exhibiting ferroelectric-like behavior. The pseudo-magnetization can be manipulated by extremely low voltages especially when the FM is designed as a low-barrier nanomagnet. Using a circuit model that is benchmarked against experiments, we determine the switching energy, delay, probability and retention time of our memory device in-terms of magnetic and circuit parameters and discuss its thermal stability. READ and WRITE operations of a 1T/1C memory architecture are shown. The proposed memory device combines the advantages of ferroelectric memory devices, such as energy-efficiency and high speed with those of magnetic memory such as non-volatility and high density.

5.1 Equivalent Circuit Model for Magnetoelectric Effect

In this section, we describe an equivalent circuit model applicable to a wide variety of magnetoelectric phenomena and use SPICE simulations to benchmark this model against experimental data. There is increasing interest in magnetic random access memory (MRAM) technology to develop voltage-driven units based on different types of magnetoelectric phenomena [137]–[154] for low power operation. We present an equivalent circuit model (Fig. 5.1) applicable to a range of magnetoelectric (ME) phenomena including both write and read op-

erations. It consists of a capacitor circuit which incorporates the back voltage from the magnetoelectric coupling described by (5.1):

$$\mathrm{V_{IN}} = \frac{Q}{C_L} + \frac{Q}{C} + \frac{\partial E_m}{\partial Q} \tag{5.1}$$

where E_m is the magnetic energy including the part controlled by the charge Q on an adjacent capacitor C, through the ME effect. Equation (5.1) is solved self-consistently with the stochastic Landau-Lifshitz-Gilbert (s-LLG) equation which feels an effective field ($\vec{H}_{me} = -\nabla_{\mathbf{m}}\ E_m/\{M_s \mathrm{Vol.}\}$), $\nabla_{\mathbf{m}}$ represents the gradient operator with respect to magnetization directions $\hat{m}_\mathrm{i}$, M_s is the saturation magnetization and Vol. is the volume of the magnet.

We first benchmark this equivalent circuit against the recently demonstrated Magneto-ELectric Random Access Memory (MELRAM) device [155], [156] which uses the magneto-electric effect (ME) and its inverse (IME) for write and read operations, using a structure whose energy E_m is given by Eq. 5.2. We then argue that, unlike MELRAM, the "1" and the "0" states need not be represented by states with a net magnetization. For example, using a structure whose energy E is given by eq. 5.4, one could instead switch the easy axis with a write voltage, and this change in the easy axis can be read as a change in the voltage across a series capacitor through the inverse effect, allowing a "field-free" operation without any symmetry breaking magnetic field.

Experimental Benchmark

We start with the MELRAM device (Fig. 5.2b) reported recently in [156] where the magnetic energy has the form

$$E_m = -E_A m_x m_y + E_H/\sqrt{2}(m_x - m_y) + v_M Q\ (m_x^2 - m_y^2) \tag{5.2}$$

We note that this energy expression is essentially the same as what was reported in Ref. [156] expressed using magnetization components, m_x, m_y, m_z. For example, the anisotropy energy is written in [156] as $-E_A \sin^2\phi$, with ϕ measured from the magnetic field $\vec{H}_{ext}$ such that $m_x = \cos(3\boldsymbol{\pi}/4 - \phi), m_y = \sin(\boldsymbol{\pi}/4 - \phi)$ and $m_x m_y = \sin^2\phi$, ignoring an unimpor-

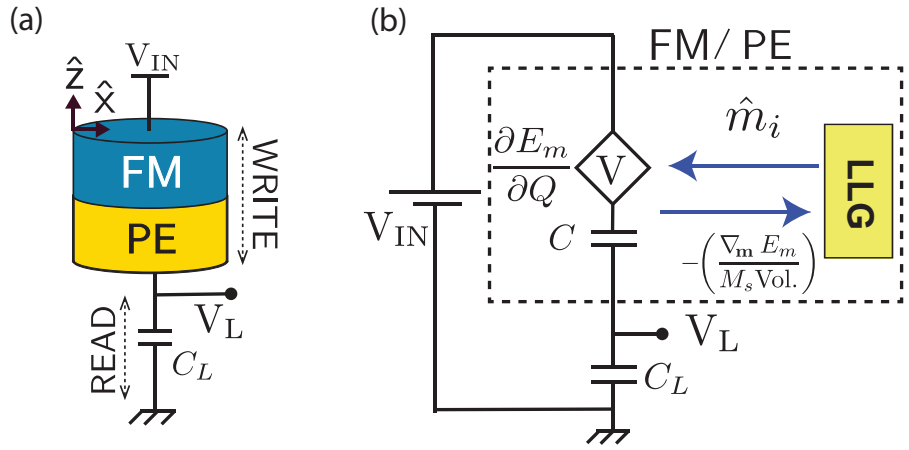


Figure 5.1. Equivalent circuit for magnetoelectric (ME) read and write operations (a) The charge on the piezoelectric (PE) capacitor changes the easy-axis of the ferromagnet (FM) and this causes a change in the output voltage V_L through the inverse effect. (b) Equivalent circuit model obtained from (5.1). Write operation is through the effective field $\vec{H}_{me} = -\nabla_{\mathbf{m}} E_m/(M_s \text{Vol.})$ that enters the stochastic Landau-Lifshitz-Gilbert (s-LLG) equation. Read operation is through the dependent voltage source V that is proportional to $\partial E_m/\partial Q$, where E_m is the magnetic energy.

tant constant. Similarly the Zeeman term is written in [156] as $-E_H \cos\phi$ which equals $E_H(m_x - m_y)/\sqrt{2}$. In [156], the uniaxial anisotropy energy term and the external magnetic field were ingeniously balanced (by choosing $E_H = E_A\sqrt{2}$) to provide two unique low energy states that represent "0" and "1" at $\phi = \pi/2$ and $\phi = \pi$.

Finally, the last term represents the ME effect where an applied voltage generates a charge Q, controlled by the input voltage V_{IN}, which changes the anisotropy energy such that a positive (or negative) Q causes the magnetic energy to favor the y-axis (or the x-axis) for a positive v_M. This is due to the anisotropic piezoelectric coefficients d_{31} and d_{32} having different signs, a special property of the $\langle 011 \rangle$-cut (PMN-PT) that was chosen in the experiment.

The equivalent circuit incorporates the back voltage from the ME coupling using (5.1), with the load capacitor C_L replaced by a resistor R:

$$V_{IN} = R\frac{dQ}{dt} + \frac{Q}{C} + \frac{\partial E_m}{\partial Q} = \quad R\frac{dQ}{dt} + \frac{Q}{C} + v_M(m_x^2 - m_y^2) \tag{5.3}$$

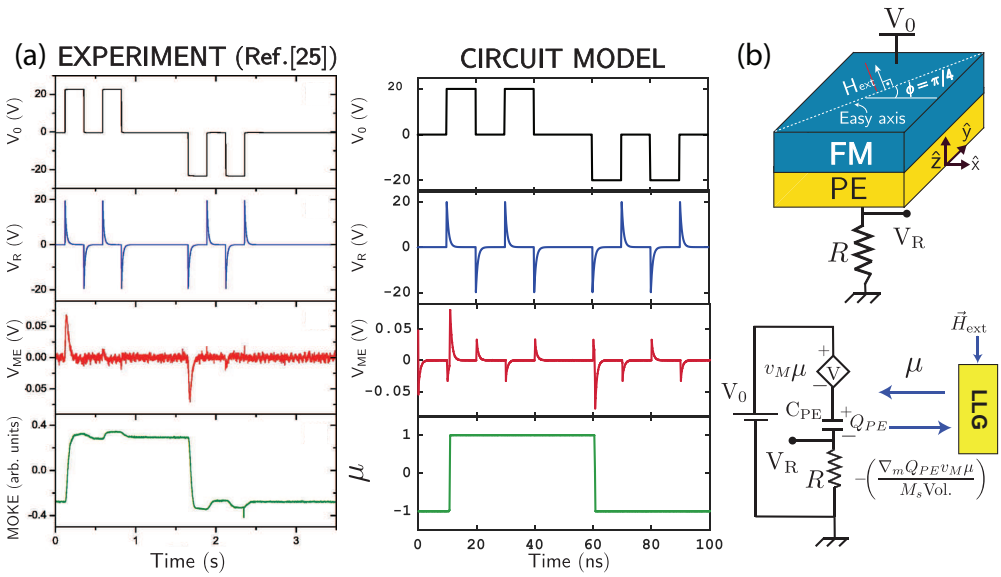


Figure 5.2. Experiment vs circuit model: (a) The results of the self-consistent circuit model for the structure in (b) are in good agreement with the experimental results in [156]. $V_{\rm ME}$ is the mathematical difference of two measurements of V_R with and without the external magnetic field, $V_{\rm ME} = V_R(H \neq 0) - V_R(H = 0)$. (b) Experimental structure reported in [156] where the piezoelectric (PE) is ⟨011⟩-cut PMN-PT and the ferromagnet (FM) is N layers of $TbCo_2$/FeCo. The back-voltage is V=$v_M\mu$ where $\mu = m_x^2 - m_y^2$ and the magnetic energy is $E_m = Q_{PE} v_M \mu$ where Q_{PE} is the charge on the capacitor $C_{\rm PE}$. The following parameters are used: Coercivity for FM (H_K=200 Oe), saturation magnetization M_s=1100 emu/cc, FM thickness, $t_{\rm FM}$=200 nm, PE thickness $t_{\rm PE}$=30 μm, Area=520 × 520 nm^2, Magnetoelastic constant $B = -7$ MPa, a net PE constant, $d = d_{31} - d_{32} = 2500$ pC/N, permittivity $\epsilon = 4033\ \epsilon_0$, resistance $R = 2$ MΩ, back voltage $v_M = B d t_{\rm FM}/2\epsilon$. In the experiment, magneto-optic Kerr effect (M.O.K.E) is used to show the variation of magnetization, which is compared to the pseudo-magnetization in our simulation. Experimental panel is reproduced with permission of AIP Publishing LLC, from Reference [156].

It is possible to write the ME energy as $q_M V$ in terms of an applied voltage V rather than charge Q, but this choice would lead to a back charge $\partial E_m/\partial V$ instead of a back voltage $\partial E_m/\partial Q$, giving a different but equivalent looking circuit model.

Fig. 5.2a shows the write and read signals for the experimental structure in Fig. 5.2b calculated using a SPICE model, that are in good agreement with the experimental results presented in [156]. The reason for the very different time scales of the experiment and the circuit model is that the circuit model solves the real-time dynamics of the nanomagnet with time steps of the order of a fraction of the inverse FMR frequency of the nanomagnet ($1/f \sim 2\pi/\gamma/\sqrt{[H_K(H_K + 4\pi M_s)]} \sim 0.2$ ns for the chosen parameters) to avoid large numerical integrations while the experimental measurement is performed with quasi-static pulses. Therefore the RC time constants in both cases are very different, however the maxima and minima of each signal closely match based on the chosen parameters.

We use this model to suggest a different mode of operation where the "1" and "0" states are not represented by states with net magnetization (like m_x, m_y or m_z) but by different easy axes, quantitatively described by $(m_x^2 - m_y^2)$ which switches from "0" to "1" through the write voltage. This change is directly detected as a read signal through the inverse effect. The use of $(m_x^2 - m_y^2)$ to represent a bit is a radical departure from the standard convention of using the magnetization (m) to represent information.

5.2 Pseudomagnetization - New Order Parameter

In recent years, voltage control of magnetism (VCM) has emerged as a promising alternative to current control of magnetism due to its potential for energy efficiency [157]. Apart from a special class of VCM phenomena that allows a deterministic 180 degree switching of magnetism [143], VCM typically results in 90 degree switching of magnetization or a change in the easy-axis of the magnetization, necessitating additional assist mechanisms or complex pulsing schemes [158], [159].

We show that the easy-axis information (or pseudo-magnetization) itself can be a state variable that can be switched between two deterministic states (WRITE) and that can be read out through the inverse effect (READ). The principle of pseudo-magnetism is general and could find use in voltage-control of magnetic anisotropy devices [158], but we focus

our theoretical and experimental discussion to piezoelectric/ ferromagnetic (PE/FM) heterostructures. We show that the interaction of magnetism and piezoelectricity can lead to non-volatility in pseudo-magnetization that can reach to years of retention time for experimentally demonstrated magnetic and circuit parameters. We describe a prototypical 1T/1C memory cell that encodes pseudo-magnetization and show its READ and WRITE operation through the equivalent circuit model [136] that is benchmarked against experiments [156], [160].

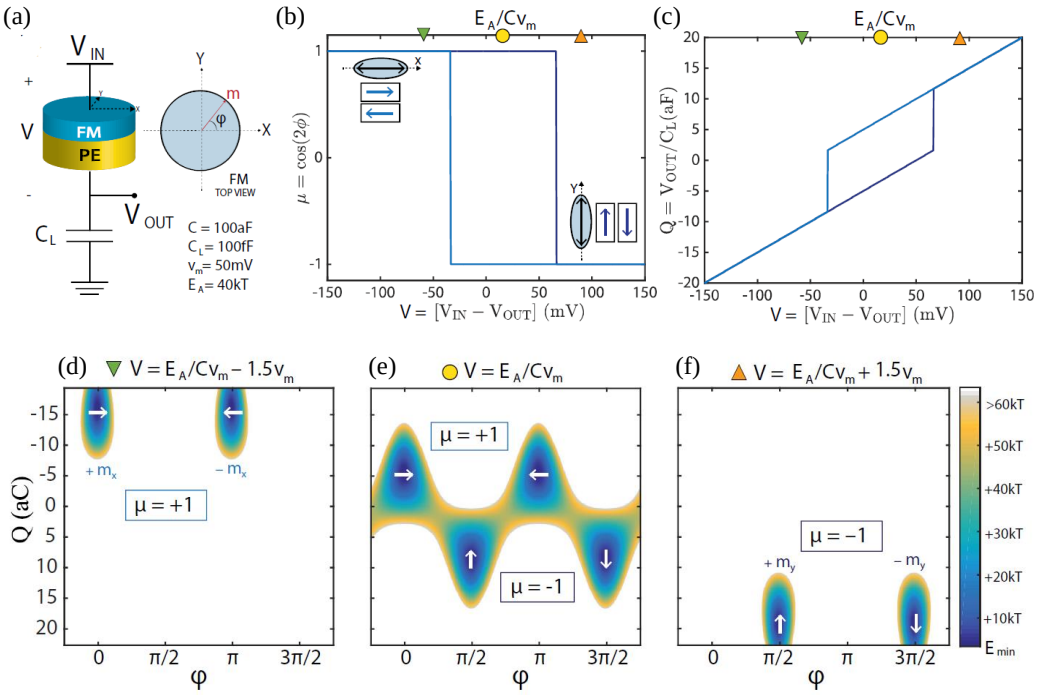


Figure 5.3. Pseudomagnetization (a) Basic electrical circuit for characterization of PE/FM structure. Information on the device is stored in the magnetic easy axis direction (± x or ± y) which we term pseudomagnetization, μ. (b) Shows the change of μ due to the applied voltage, V across the PE/FM structure and (c) shows the resulting charge versus voltage characteristics in the circuit which is similar to standard ferroelectrics. (d)-(f) shows the stable states at different voltages across the structure on a heatmap. Unlike conventional magnetic memory there are multiple states associated with each voltage indicating preferred easy axis. The states are separated by a large barrier, so which allows for non-volatile memory application.

We start from an energy expression associated with the PE/FM heterostructure in fig. 5.3 a:

$$E = \frac{Q^2}{2C} + Qv_m\mu - QV_{IN} - (\frac{E_A}{2})\mu \tag{5.4}$$

where μ is the pseudo-magnetization that defines the easy-axis for the magnet $\mu = \mathrm{m_x^2} - \mathrm{m_y^2}$, $\mathrm{E_A} = \mathrm{H_K M_S \Omega/2}$ is the magnetic anisotropy that defines an easy-axis for the magnet, $\mathrm{V_{IN}}$ is the applied voltage and v_m is the magnetoelectric (ME) back voltage that couples the charge Q on the PE capacitor (C) with the pseudo-magnetization μ of the FM through the internal strain. In the PE/FM heterostucture, μ is given by a combination of the material parameters of the PE and FM, $\mathrm{v_m} = \mathrm{Bdt_{FM}}/2\epsilon$ where, B is the magnetoelastic constant of the magnet, d is the net piezoelectric coefficient of the PE layer, ϵ is its the dielectric permittivity, and t_{FM} is the thickness of the magnet. The circuit model we use to generate results in fig. 5.4-5.5 is derived from this energy model, and the fundamental operation can be understood from the energy equation.

For a given $\mathrm{V_{IN}}$, charge is formed on the capacitor which creates an effective anisotropy like energy $\mathrm{Qv_m}$ in the magnet. If this energy is large enough ($\ll k_BT$), a preferred easy axis will be induced in the magnet and the magnetization will lie in that axis without a preference for a direction. Consider the case when $V_{IN} = 0$ and a low barrier magnet ($E_A \sim 0$). In this case the energy, $\mathrm{E} = (1/2\mathrm{C})(\mathrm{Q} + \mathrm{Cv_m}\mu)^2 - (1/2)\mathrm{Cv_m^2}\mu^2$ is minimized when $\mu = \pm 1$. Therefore, even when $\mathrm{V_{IN}} = 0$, as long as $Cv_m^2 \ll k_BT$, μ can get spontaneously polarized and induces an internal charge $Q = Cv_m\mu$, much like a standard ferroelectric.

A self-consistent solution of the energy equation for this minimum energy condition shows this phenomenon (fig. 5.3 b,c). The width of the pseudo-magnetization vs. voltage hysteretic loop is independent of the capacitance of the structure and depends only on the magnetoelectric voltage v_m, but the actual switching voltage depends on the capacitance, C through the anisotropy energy associated with the magnet, E_A. The loop is symmetric about the point E_A/Cv_m. It is important to note that although the charge on the load capacitor C_L may leak out after writing, the pseudo-magnetization information is still preserved in the cell much like ferroelectric random-access-memories (FeRAM).

Fig. 5.3(d)-(f) shows the heatmap of the charge and magnetization state associated with three voltage conditions using the energy expression. At the symmetry point E_A/Cv_m all four states of magnetization (two of pseudo-magnetization) are equally probable. They are separated by a barrier of the order of $Cv_m^2/2$ that can be designed to be much larger than k_BT for typical parameters. This means that, once the system is in one of the four states, it remains there. Applying $\pm\, v_m$ from the symmetry point switches the pseudo-magnetization to ∓ 1 states.

5.3 Magnetoelectric Memory Cell

A magnetoelectric memory cell that uses pseudo-magnetization can be constructed like a standard 1T-1C circuit where one end of the PE/FM capacitor is connected to the bit-line (BL) through a pass transistor and the other end is connected to a plate line (PL). The cell access is provided by the word-line (WL) as shown in fig. 5.4a. From SPICE simulations of the circuit model [159] we show the WRITE and READ process of the cell in fig. 5.4b,c.

To write a '0' the BL is charged to '$2v_m$', PL is kept grounded and then the transistor is turned on through the WL to complete the writing process. To write a '1' a similar procedure is employed where the BL is charged to '$-2v_m$' instead. Fig. 5.4b shows the writing process. Even after the charge on the load capacitor leaks, the internal state of the cell can be retained for a long time as long as $Cv_m^2/2k_BT \gg 1$. Therefore, for reading the state a read pulse needs to be applied. The BL is first pre-charged to '0V' then the access transistor is turned on which creates a capacitive divider circuit between PL and ground. When a positive read pulse is applied to PL, the voltage is divided between the ME capacitor and the bit-line capacitance (CBL) depending on their relative values and the state of the ME device. A sense amplifier can then be used to detect these voltages. As fig. 5.4b indicates, this reading process is destructive so data must be rewritten once read, similar to FeRAM [161]. In the simulations in fig. 5.4 we chose parameters for the ME circuit, which are based on experimentally reported material parameters [156], [160] and reasonable device dimensions. We also chose the bit-line parasitic capacitance to be equal to the PE capacitance ($C_{BL} = C$) for simplicity.

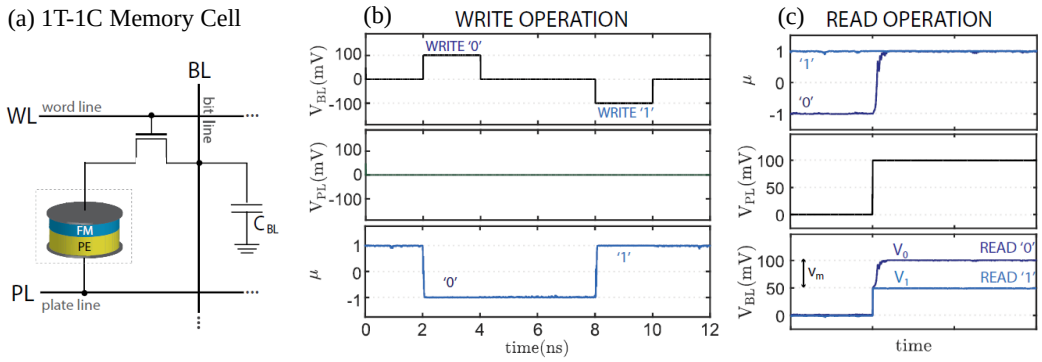


Figure 5.4. (a) Magnetoelectric 1T-1C memory cell.The READ/WRITE Operation of the cell mimics the scheme of FeRAM operation. (b) WRITE pulse is applied to the bit-line keeping plate line grounded. (c) READ pulse is applied to the PL and voltage at BL is detected. The read process is destructive as in FeRAM, but unlike DRAM is μ non-voltaile so does not require periodic refresh.

The energy barrier that determines the stability of pseudo-magnetization can be related to its equilibrium fluctuations (Fig. 5.5a). The RMS value of equilibrium fluctuations is related to the energy barrier of the magnet by: $\Delta = k_B T/2(1-\mu_{RMS}^2)$. Fig. 5.5b shows the extracted thermal barrier from 1000 samples, for different magnetoelectric voltages for a constant C. The results agree well with an analytically derived value of $Cv_m^2/2$ which can be $> 40k_BT$ for experimentally demonstrated parameters.

Additionally, we estimate switching energies and time associated with the write operation. The switching time of pseudo-magnetization is related to the magnet dynamics. The voltage generated stress can be expressed as an effective magnetic field $H_s \equiv (Qv_m/M_sVol.) \equiv (CV_{IN}v_m/M_sVol)$. This effective magnetic field can be used to estimate the typical switching time of magnetization where $\tau \sim 1/\alpha\gamma H_s$ that can result in sub-ns switching speeds for typical parameters. As the PE/FM heterostructure is a fully capacitive system the write energy approximated by $CV_{IN}^2/2$ can also be very low. Ignoring parasitics and other non-idealities, this number can be optimistically in the $\sim$ aJ range for our experimentally guided parameters.

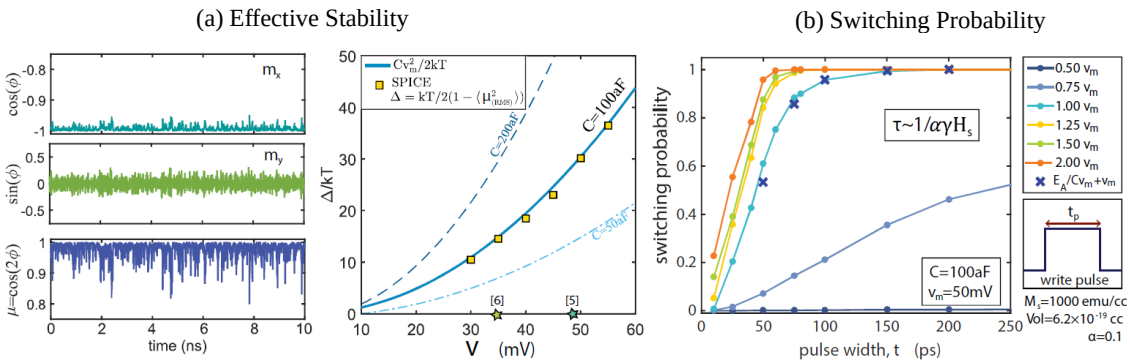


Figure 5.5. (a)The stability of pseudomagnetization states can be measured from equilibrium fluctuations. The effective stability (Δ) of μ can be attributed to an effective stress anisotropy field (H_s) it feels which depends on the back-voltage v_m and the capacitance value C. (b) Switching probability of pseudomagnetization is calculated from 1500 samples for different amplitudes and pulse widths. Sub-ns switching speeds (τ) can be attained due high stress fields ($H_s = CV_{IN}v_m/M_sVol$) in nanomagnets.

5.4 Extraction of v_m from FMR Results

The magneto-electric back voltage v_m in an FM/PE heterostructure can be extracted from ferromagnetic resonace (FMR) measurements. We characterize two sets of ferromagnetic resonance (FMR) measurements performed on a (011) cut PMN-PT/CoFeB film and array of nanodots. The experimental details can be found in [162]. The peak resonance frequency of FMR is typically described by the Kittel formula. Here, we derive a modified Kittel formula that includes the voltage induced stress term from the Landau-Lifshitz-Gilbert (LLG) equation. The free-energy associated with an in-plane magnet whose easy axis is along the x-direction and the field applied along the easy-axis is: $\mathrm{E_m} = [\mathrm{H_k}(1-\mathrm{m_x^2}) + \mathrm{H_D m_Z^2} + \mathrm{H_S}(\mathrm{m_x^2}-\mathrm{m_y^2}) - \mathrm{H_{res} m_x}]\mathrm{M_s}\Omega$, where $\mathrm{H_D}$ denoted the demagnetization field perpendicular to the plane. The resonance frequency of such a magnet can be derived from its free-energy expression which results in the modified Kittle equations as follows:

$$f = \frac{\gamma}{2\pi}\sqrt{(H_k + H_{res} - 2H_S)(H_k + H_{res} + H_D - H_S)}, \quad H_S > (H_{res} + H_k)/2 \tag{5.5}$$

$$f = \frac{\gamma}{2\pi}\sqrt{\frac{(H_{res}^2 - (H_k - 2Hs)^2))(H_D + H_{res} + H_K - H_S)}{H_k - 2H_s}}, \quad H_S \leq (H_{res} + H_k)/2 \qquad (5.6)$$

The key parameter is once again, the stress-induced magnetic field $H_S \approx Bd(V_{IN}/t_{PE})/M_s \approx 2CV_{IN}v_m/(M_s\Omega)$ that modifies the easy-axis anisotropy. We use the experimental FMR data to extract the back-volatge v_m for the two experimental cases as shown in fig. 5.6.

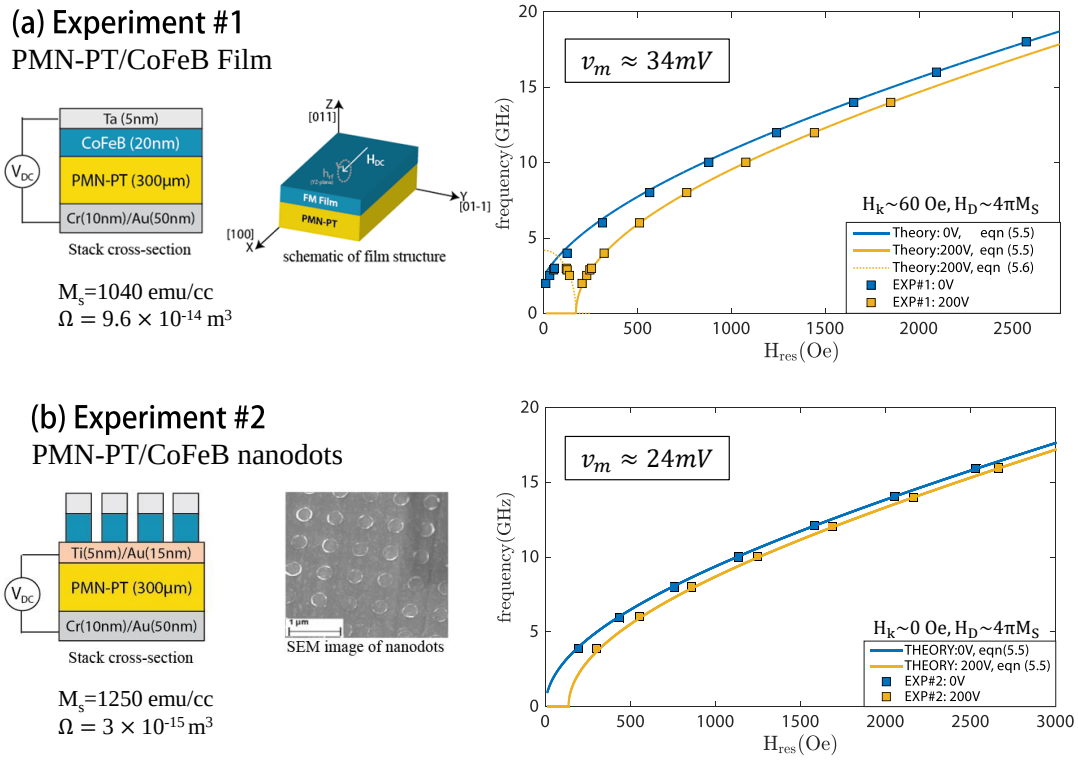


Figure 5.6. Characterizing FMR Measurements Ferromagnetic resonance (FMR) measurements performed on two samples (a) Film and (b) nanodot array show modification of magnetic anisotropy of CoFeB by applying voltage across the PMN-PT layer. The modified Kittle equations (eq. 5.5 and 5.6) including the voltage-induced stress term H_s are used to fit the measurements.The reported experimental parameters for the piezoelectric are relative permittivity $\epsilon_r = 600$, piezoelectric co-efficient d = 4500 pC/N, and for the magnet the magnetoelastic constant B = 4 MPa. For the film the theoretically expected ME back-voltage ($v_m = Bdt_{FM}/2\epsilon$) of 34 mV fits the data while a slightly lower value of 34 mV fits the nanodots which has a Ti/Au layer inbetween the PE and FM layer.

For a magnetic film, the data fits well with the theoretical expectation of $\sim$ 34mV, while a slightly lower value of $\sim$ 24mV is extracted for the nanodot arrays. These results show

that the easy-axis or pseudo-magnetization can be manipulated by voltages and can result in values of v_m in tens of millivolts range for both film and patterned magnets demonstrating WRITE operation. Electrical detection of the change in easy-axis and the resulting pseudo-magnetization from first and second harmonic measurements on a similar sample have been reported in [160], and characterized using the same theory. The measurements indicated a $v_m \sim$ few mV, demonstrating the feasibility of magnetoelectric READ operation of the pseudo-magnetization.

5.5 Summary

In this chapter, we have presented an equivalent circuit for magneotoelectric read and write and showed that it describes recent experiments on the MELRAM device quite accurately. We analyzed the feasibility of a new magnetoelectric memory device that uses a new order parameter, pseudo-magnetization. When the magnet is designed as a low-barrier nanomagnet this device can potentially operate with an energy-delay of hundreds of (aJ-ps) while combining attractive features of magnetic and ferroelectric memory technologies such as high-density and non-volatility.

6. SUMMARY

Probabilistic computing has emerged as an effective and more immediate means of handling search and recognition problems posed by the ever-increasing amounts and demands of big data. A probabilistic computer can bridge the gap between genuine quantum computers and the standard classical computers. In this thesis, we have presented a complete evaluation of naturally stochastic hardware based on low-barrier magnets for scalable and energy-efficient realization of p-computers. We demonstrated the potential benefits of leveraging natural stochasticity for efficient simulation of probability.Such probabilistic hardware can implement probabilistic algorithms that use Markov chain Monte Carlo process compactly and efficiently. The final chapter serves to highlight the key findings and contributions of this work and look ahead towards additional areas of exploration.

6.1 Realization of Naturally Stochastic Hardware

We evaluated the compact mixed-signal unit based on a low-barrier nanomagnet that uses a single magnetic tunnel junction (MTJ). Such a compact unit can drastically reduce the area footprint of p-bit hardware while promising massive scalability by leveraging the existing Magnetic RAM (MRAM) technology that has integrated 1T-1MTJ cells in $\sim$ Gbit densities. By employing circular in-plane LBM coupled the designs can respond in sub-ns timescales requiring only a few fJs of energy.

Use of naturally stochastic bits enable autonomous or clockless operation so that PSL is not limited by clock frequencies like digital circuits. This enables to go beyond the tera flips per second flips per second that current CMOS processors are stuck at and venture into peta to even beyond-exa scale operation.

To serve as bridge, probabilistic computing needs it to be a near-term technology unlike quantum computer. Our design evaluation demonstrates a path to realization of p-bits with what we believe are only slight modifications in the already established magnetic memory technology. Infact, since our proposals academic and industry efforts that were set-forth has seen success in realization of functional low-barrier magnet MTJs. But we would still like to

point out that the realization need not be an LBM MTJ based realization. Any two terminal stochastic resistor realizations can be adapted into the design.

Explore Novel Stochastic Mechanisms: We established the design rules for the p-bit device from a very general two terminal stochastic resistor to encourage further research into novel mechanisms that can harness natural stochasticity in circuits.

6.2 Physics of Low-barrier Magnets

Magnetic memory is a mature technology where stable magnets are typically used to store information which can be retained for many many years. Thus most of the theoretical predictions and characterizations involved looking into stable magnets which has a high barrier separating its two states. The design explorations and material research is focused on achieving high-barrier at nano-scale. Low-barrier magnets whose magnetization fluctuates instead of retaining its state is naturally considered a nuisance in this respect. LBMs have thus largely been ignored all this time, creating a gap in the theoretical understanding and predictions due to their irrelevance. However, p-bit designs that we propose aim to leverage from these "bad" bits of the memory world giving them relevance. In such applications it is desired to have a very small barrier between the magnetic states to enable thermal noise to cause rapid fluctuations of the states. In chapter 2 and 3 we aimed to fill this gap by analyzing the behavior of LBM and providing relevant expressions.

We discovered that in-plane LBM has surprisingly fast fluctuation rates as low as sub-ns timescales and high threshold currents compared to its isotropic or uniaxial magnetization counterparts. Our numerical results supported by the theoretical understanding which have recently received some experimental confirmations, suggest investing in fabricating circular IMA magnets with fast fluctuation rates for realization of efficient p-bits.

Novel Spintronic Devices: The theory of LBM does not only lend itself to p-bits, it can open up new ways to building new energy-efficient spintronics devices like oscillators (as hinted in Chapter. 3, multi-terminal rectifiers [85], and even memory devices as we showed

in Chapter. 6. The use of LBM can lead to extremely low voltage manipulation of data in coupled piezoelectric and ferromagnetic (PE/FM) heterostructures.

6.3 Benchmarking Metrics for Probabilistic Computing

A clear and definite set of metrics are needed to asses the performance of probabilistic hardware. Although probabilistic computing is not a new area of study, but as it has been mostly limited to software implementations until recently, it still lacks a standard industry standard for benchmarking.

In chapter 2 we define the performance metrics for an individual p-bit as time and energy per flip similar to the switching time and energy associated with digital electronics. Our LBM based design shows that it can flip at sub-ns timescales requiring only a few fJ energy orders of magnitude better than its CMOS counterparts.

In chapter 3 we emphasise the use of a problem and substrate independent performance metrics - flips per second and energy per flip to benchmark specialized probabilistic hardware like Ising Machines. Flips per second refers to the number of samples the system can produce per second. The basic operation of a probabilistic computer like Ising machines is to go through samples and reach a solution. The number of samples required to reach a solution is problem dependent, so a general way to benchmark just the hardware's performance could be through the more fundamental flips per second metric, similar to the FLOPS of digital computers, that can be continually improved in later technology generations of probabilistic hardware.

Compared to digital Ising machines today with tera fps performance our naturally stochastic hardware is projected to out-perform them by orders of magnitude (a million p-bit network is projected to achieve more than peta fps). As the realization depends on slight modifications of the existing MRAM industry, we hope the results, discussion, and design guidelines presented will serve to encourage industries to consider building this network in larger scale and also expand into different application areas.

Accelerating Machine Learning: Probabilistic bits are analogous to the binary stochastic neurons of stochastic neural networks. Stochastic neural networks like Boltzmann Machines (BM) and Restricted Boltzmann Machines (RBM) are an integral component to deep belief networks, which have become more common with increased interest in deep learning [127] today. Stochastic binarization is desired by typically avoided due to hardware complexity of realizing true random numbers with deterministic circuitry, to quote the expert Yoshua Bengio, "The stochastic binarization is more appealing than the sign function, but harder to implement as it requires the hardware to generate random bits when quantizing. As a result, we mostly use the deterministic binarization function..."[128]. Our s-MTJ based p-bits can evaluate the BSN function fast and efficiently, and thus could accelerate computation in custom neural-network hardware [129], [130].

REFERENCES

[1] J. Shalf, "The future of computing beyond moore's law," *Philosophical Transactions of the Royal Society A*, vol. 378, no. 2166, p. 20 190 061, 2020.

[2] G. E. Moore *et al.*, *Cramming more components onto integrated circuits*, 1965.

[3] M. Kanellos, *Moore says nanoelectronics face tough challenges*, Mar. 2005. [Online]. Available: https://www.cnet.com/news/moore-says-nanoelectronics-face-tough-challenges/.

[4] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted mosfet's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, 1974.

[5] *It's time to shift to data centric computing*. [Online]. Available: https://research.ibm.com/articles/datacentricdesign/.

[6] T. M. Conte, E. P. DeBenedictis, P. A. Gargini, and E. Track, "Rebooting computing: The road ahead," *Computer*, vol. 50, no. 1, pp. 20–29, 2017.

[7] N. Thompson and S. Spanuth, "The decline of computers as a general purpose technology: Why deep learning and the end of moore's law are fragmenting computing," *Available at SSRN 3287769*, 2018.

[8] W. G. Hatcher and W. Yu, "A survey of deep learning: Platforms, applications and emerging research trends," *IEEE Access*, vol. 6, pp. 24 411–24 432, 2018.

[9] S. Shi, Q. Wang, P. Xu, and X. Chu, "Benchmarking state-of-the-art deep learning software tools," in *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, IEEE, 2016, pp. 99–104.

[10] V. J. Reddi, C. Cheng, D. Kanter, P. Mattson, G. Schmuelling, C.-J. Wu, B. Anderson, M. Breughe, M. Charlebois, W. Chou, *et al.*, "Mlperf inference benchmark," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, IEEE, 2020, pp. 446–459.

[11] *Quantum Computing Market Research Report: By Offering, Deployment Type, Application, Technology, Industry - Industry Share, Growth, Drivers, Trends and Demand Forecast to 2030*. Apr. 2020. [Online]. Available: https://www.researchandmarkets.com/reports/5010716/quantum-computing-market-research-report-by?utm_source=dynamic&utm_medium=GNOM&utm_code=4m3fxs&utm_campaign=1375670%20-%20Worldwide%20Quantum%20Computing%20Market%20(2019%20to%202030)%20-%20Drivers,%20Restraints%20and%20Opportunities&utm_exec=jamu273gnomd.

[12] M. Kjaergaard, M. E. Schwartz, J. Braumüller, P. Krantz, J. I.-J. Wang, S. Gustavsson, and W. D. Oliver, "Superconducting qubits: Current state of play," *Annual Review of Condensed Matter Physics*, vol. 11, pp. 369–395, 2020.

[13] S. GAMBLE, "Quantum computing: What it is, why we want it, and how we're trying to get it," in *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2018 Symposium*, National Academies Press, 2019.

[14] F. Barahona, "On the computational complexity of ising spin glass models," *Journal of Physics A: Mathematical and General*, vol. 15, no. 10, p. 3241, 1982.

[15] A. Lucas, "Ising formulations of many np problems," *Frontiers in Physics*, vol. 2, p. 5, 2014.

[16] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[17] M. W. Johnson, M. H. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P. Bunyk, *et al.*, "Quantum annealing with manufactured spins," *Nature*, vol. 473, no. 7346, pp. 194–198, 2011.

[18] P. L. McMahon, A. Marandi, Y. Haribara, R. Hamerly, C. Langrock, S. Tamate, T. Inagaki, H. Takesue, S. Utsunomiya, K. Aihara, *et al.*, "A fully programmable 100-spin coherent ising machine with all-to-all connections," *Science*, vol. 354, no. 6312, pp. 614–617, 2016.

[19] S. Dutta, A. Khanna, H. Paik, D. Schlom, A. Raychowdhury, Z. Toroczkai, and S. Datta, "Ising hamiltonian solver using stochastic phase-transition nano-oscillators," *arXiv preprint arXiv:2007.12331*, 2020.

[20] H. Goto, K. Tatsumura, and A. R. Dixon, "Combinatorial optimization by simulating adiabatic bifurcations in nonlinear hamiltonian systems," *Science advances*, vol. 5, no. 4, eaav2372, 2019.

[21] T. Wang and J. Roychowdhury, "Oim: Oscillator-based ising machines for solving combinatorial optimisation problems," in *International Conference on Unconventional Computation and Natural Computation*, Springer, 2019, pp. 232–256.

[22] I. Ahmed, P.-W. Chiu, and C. H. Kim, "A probabilistic self-annealing compute fabric based on 560 hexagonally coupled ring oscillators for solving combinatorial optimization problems," in *2020 IEEE Symposium on VLSI Circuits*, IEEE, 2020, pp. 1–2.

[23] J. Chou, S. Bramhavar, S. Ghosh, and W. Herzog, "Analog coupled oscillator based weighted ising machine," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[24] M. Baity-Jesi, R. A. Baños, A. Cruz, L. A. Fernandez, J. M. Gil-Narvión, A. Gordillo-Guerrero, D. Iñiguez, A. Maiorano, F. Mantovani, E. Marinari, *et al.*, "Janus ii: A new generation application-driven computer for spin-system simulations," *Computer Physics Communications*, vol. 185, no. 2, pp. 550–559, 2014.

[25] M. Yamaoka, C. Yoshimura, M. Hayashi, T. Okuyama, H. Aoki, and H. Mizuno, "24.3 20k-spin ising chip for combinational optimization problem with cmos annealing," in *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*, IEEE, 2015, pp. 1–3.

[26] T. Takemoto, M. Hayashi, C. Yoshimura, and M. Yamaoka, "2.6 a 2× 30k-spin multichip scalable annealing processor based on a processing-in-memory approach for solving large-scale combinatorial optimization problems," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, IEEE, 2019, pp. 52–54.

[27] M. Aramon, G. Rosenberg, E. Valiante, T. Miyazawa, H. Tamura, and H. G. Katzgraber, "Physics-inspired optimization for quadratic unconstrained problems using a digital annealer," *Frontiers in Physics*, vol. 7, p. 48, 2019.

[28] K. Yamamoto, K. Ando, N. Mertig, T. Takemoto, M. Yamaoka, H. Teramoto, A. Sakai, S. Takamaeda-Yamazaki, and M. Motomura, "7.3 statica: A 512-spin 0.25 m-weight full-digital annealing processor with a near-memory all-spin-updates-at-once architecture for combinatorial optimization with complete spin-spin interactions," in *2020 IEEE International Solid-State Circuits Conference-(ISSCC)*, IEEE, 2020, pp. 138–140.

[29] S. Patel, L. Chen, P. Canoza, and S. Salahuddin, "Ising model optimization problems on a fpga accelerated restricted boltzmann machine," *arXiv preprint arXiv:2008.04436*, 2020.

[30] S. Patel, P. Canoza, and S. Salahuddin, "Logically synthesized, hardware-accelerated, restricted boltzmann machines for combinatorial optimization and integer factorization," *arXiv preprint arXiv:2007.13489*, 2020.

[31] K. Y. Camsari, R. Faria, B. M. Sutton, and S. Datta, "Stochastic p-bits for invertible logic," *Physical Review X*, vol. 7, no. 3, p. 031 014, 2017.

[32] K. Y. Camsari, B. M. Sutton, and S. Datta, "P-bits for probabilistic spin logic," *arXiv preprint arXiv:1809.04028*, 2018.

[33] B. Behin-Aein, V. Diep, and S. Datta, "A building block for hardware belief networks," *Scientific reports*, vol. 6, p. 29 893, 2016.

[34] R. P. Feynman, "Simulating physics with computers," *International journal of theoretical physics*, vol. 21, no. 6-7, pp. 467–488, 1982.

[35] R. Faria, K. Y. Camsari, and S. Datta, "Implementing bayesian networks with embedded stochastic mram," *AIP Advances*, vol. 8, no. 4, p. 045 101, 2018.

[36] J. Kaiser, R. Faria, K. Y. Camsari, and S. Datta, "Probabilistic circuits for autonomous learning: A simulation study," *Frontiers in Computational Neuroscience*, vol. 14, 2020.

[37] B. Sutton, K. Y. Camsari, B. Behin-Aein, and S. Datta, "Intrinsic optimization using stochastic nanomagnets," *Scientific Reports*, vol. 7, p. 44 370, 2017.

[38] W. A. Borders, A. Z. Pervaiz, S. Fukami, K. Y. Camsari, H. Ohno, and S. Datta, "Integer factorization using stochastic magnetic tunnel junctions," *Nature*, vol. 573, no. 7774, pp. 390–393, 2019.

[39] O. Hassan, K. Y. Camsari, and S. Datta, "Voltage-driven building block for hardware belief networks," *IEEE Design & Test*, vol. 36, no. 3, pp. 15–21, 2019.

[40] K. Y. Camsari, S. Chowdhury, and S. Datta, "Scaled quantum circuits emulated with room temperature p-bits," *arXiv preprint arXiv:1810.07144*, 2018.

[41] S. Chowdhury, K. Y. Camsari, and S. Datta, "Emulating quantum interference with generalized ising machines," *arXiv preprint arXiv:2007.07379*, 2020.

[42] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines," *Cognitive science*, vol. 9, no. 1, pp. 147–169, 1985.

[43] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[44] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves, S. Lam, N. Ge, J. J. Yang, and R. S. Williams, "Dot-product engine for neuromorphic computing: Programming 1t1m crossbar to accelerate matrix-vector multiplication," in *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, IEEE, 2016, pp. 1–6.

[45] M. Prezioso, F. Merrikh-Bayat, B. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, no. 7550, pp. 61–64, 2015.

[46] H. Huang, J. Heilmeyer, M. Grözing, M. Berroth, J. Leibrich, and W. Rosenkranz, "An 8-bit 100-gs/s distributed dac in 28-nm cmos for optical communications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 63, no. 4, pp. 1211–1218, 2015.

[47] V. Ostwal, R. Zand, R. DeMara, and J. Appenzeller, "A novel compound synapse using probabilistic spin–orbit-torque switching for mtj-based deep neural networks," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 5, no. 2, pp. 182–187, 2019.

[48] B. Sutton, R. Faria, L. A. Ghantasala, K. Y. Camsari, and S. Datta, "Autonomous probabilistic coprocessing with petaflips per second," *arXiv preprint arXiv:1907.09664*, 2019.

[49] A. Z. Pervaiz, B. M. Sutton, L. A. Ghantasala, and K. Y. Camsari, "Weighted p-bits for fpga implementation of probabilistic circuits," *IEEE transactions on neural networks and learning systems*, 2018.

[50] A. Z. Pervaiz, S. Datta, and K. Y. Camsari, "Probabilistic computing with binary stochastic neurons," in *2019 IEEE BiCMOS and Compound semiconductor Integrated Circuits and Technology Symposium (BCICTS)*, IEEE, 2019, pp. 1–6.

[51] F. Cai, S. Kumar, T. Van Vaerenbergh, R. Liu, C. Li, S. Yu, Q. Xia, J. J. Yang, R. Beausoleil, W. Lu, *et al.*, "Harnessing intrinsic noise in memristor hopfield neural networks for combinatorial optimization," *arXiv preprint arXiv:1903.11194*, 2019.

[52] S. Bhatti, R. Sbiaa, A. Hirohata, H. Ohno, S. Fukami, and S. Piramanayagam, "Spintronics based random access memory: A review," *Materials Today*, vol. 20, no. 9, pp. 530–548, 2017.

[53] "Everspin enters pilot production phase for the world's first 28 nm 1 gb stt-mram component," *Everspin Technology*, Jul. 2019. [Online]. Available: https://investor.everspin.com/news-releases/news-release-details/everspin-enters-pilot-production-phase-worlds-first-28-nm-1-gb.

[54] A. Fukushima, T. Seki, K. Yakushiji, H. Kubota, H. Imamura, S. Yuasa, and K. Ando, "Spin dice: A scalable truly random number generator based on spintronics," *Applied Physics Express*, vol. 7, no. 8, p. 083 001, 2014.

[55] W. H. Choi, Y. Lv, J. Kim, A. Deshpande, G. Kang, J.-P. Wang, and C. H. Kim, "A magnetic tunnel junction based true random number generator with conditional perturb and real-time output probability tracking," in *2014 IEEE International Electron Devices Meeting*, IEEE, 2014, pp. 12–5.

[56] H. Lee, F. Ebrahimi, P. K. Amiri, and K. L. Wang, "Design of high-throughput and low-power true random number generator utilizing perpendicularly magnetized voltage-controlled magnetic tunnel junction," *AIP Advances*, vol. 7, no. 5, p. 055 934, 2017.

[57] P. Debashis, R. Faria, K. Y. Camsari, J. Appenzeller, S. Datta, and Z. Chen, "Experimental demonstration of nanomagnet networks as hardware for ising computing," in *Electron Devices Meeting (IEDM), 2016 IEEE International*, IEEE, 2016, pp. 34–3.

[58] nanohub.org, *Modular approach to spintronics*, https://nanohub.org/groups/spintronics.

[59] *Predictive Technology Model (PTM) (http://ptm.asu.edu/).*

[60] O. Hassan, R. Faria, K. Y. Camsari, J. Z. Sun, and S. Datta, "Low-barrier magnet design for efficient hardware binary stochastic neurons," *IEEE Magnetics Letters*, vol. 10, pp. 1–5, 2019.

[61] R. Faria, K. Y. Camsari, and S. Datta, "Low-barrier nanomagnets as p-bits for spin logic," *IEEE Magnetics Letters*, vol. 8, pp. 1–5, 2017.

[62] B. Parks, M. Bapna, J. Igbokwe, H. Almasi, W. Wang, and S. A. Majetich, "Superparamagnetic perpendicular magnetic tunnel junctions for true random number generators," *AIP Advances*, vol. 8, no. 5, p. 055 903, 2018.

[63] K. Y. Camsari, R. Faria, O. Hassan, A. Z. Pervaiz, B. M. Sutton, and S. Datta, "P-transistors and p-circuits for boolean and non-boolean logic," in *Spintronics X*, International Society for Optics and Photonics, vol. 10357, 2017, 103572K.

[64] D. J. Amit and D. J. Amit, *Modeling brain function: The world of attractor neural networks*. Cambridge university press, 1992. [Online]. Available: https://doi.org/10.1016/0166-2236(90)90155-4.

[65] *Binary stochastic neurons in tensorflow (https://r2rt.com/binary-stochastic-neurons-in-tensorflow.html).*

[66] A. Alaghi and J. P. Hayes, "Survey of stochastic computing," *ACM Transactions on Embedded computing systems (TECS)*, vol. 12, no. 2s, p. 92, 2013. [Online]. Available: https://doi.org/10.1145/2465787.2465794.

[67] S. K. Esser, A. Andreopoulos, R. Appuswamy, P. Datta, D. Barch, A. Amir, J. Arthur, A. Cassidy, M. Flickner, P. Merolla, *et al.*, "Cognitive computing systems: Algorithms and applications for networks of neurosynaptic cores," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*, IEEE, 2013, pp. 1–10. [Online]. Available: https://doi.org/10.1109/IJCNN.2013.6706746.

[68] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014. [Online]. Available: https://doi.org/10.1126/science.1254642.

[69] Note that we are using a bipolar representation ±1 instead of the binary representation (0,1). This is reflected in the use of the *tanh* function in Eq. 2.1 instead of the usual logistic function.

[70] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano letters*, vol. 10, no. 4, pp. 1297–1301, 2010. [Online]. Available: http://doi.org/10.1021/nl904092h.

[71] U. Çilingiroglu, "A purely capacitive synaptic matrix for fixed-weight neural networks," *IEEE Transactions on Circuits and Systems*, vol. 38, no. 2, pp. 210–217, 1991. [Online]. Available: https://doi.org/10.1109/31.68299.

[72] B. R. Zink, Y. Lv, and J.-P. Wang, "Telegraphic switching signals by magnet tunnel junctions for neural spiking signals with high information capacity," *Journal of Applied Physics*, vol. 124, no. 15, p. 152 121, 2018. [Online]. Available: https://doi.org/10.1063/1.5042444.

[73] D. Vodenicarevic, N. Locatelli, A. Mizrahi, J. S. Friedman, A. F. Vincent, M. Romera, A. Fukushima, K. Yakushiji, H. Kubota, S. Yuasa, *et al.*, "Low-energy truly random number generation with superparamagnetic tunnel junctions for unconventional computing," *Physical Review Applied*, vol. 8, no. 5, p. 054 045, 2017.

[74] A. Mizrahi, T. Hirtzlin, A. Fukushima, H. Kubota, S. Yuasa, J. Grollier, and D. Querlioz, "Neural-like computing with populations of superparamagnetic basis functions," *Nature communications*, vol. 9, no. 1, p. 1533, 2018.

[75] D. Vodenicarevic, N. Locatelli, A. Mizrahi, T. Hirtzlin, J. S. Friedman, J. Grollier, and D. Querlioz, "Circuit-level evaluation of the generation of truly random bits with superparamagnetic tunnel junctions," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2018, pp. 1–4.

[76] C. M. Liyanagedera, A. Sengupta, A. Jaiswal, and K. Roy, "Stochastic spiking neural networks enabled by magnetic tunnel junctions: From nontelegraphic to telegraphic switching regimes," *Physical Review Applied*, vol. 8, no. 6, p. 064 017, 2017. [Online]. Available: https://doi.org/10.1103/PhysRevApplied.8.064017.

[77] K. Y. Camsari, S. Salahuddin, and S. Datta, "Implementing p-bits with embedded mtj," *IEEE Electron Device Letters*, vol. 38, no. 12, pp. 1767–1770, 2017.

[78] A. Ardakani, F. Leduc-Primeau, N. Onizawa, T. Hanyu, and W. J. Gross, "Vlsi implementation of deep neural network using integral stochastic computing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 10, pp. 2688–2699, 2017. [Online]. Available: https://doi.org/10.1109/TVLSI.2017.2654298.

[79] B. Yuan and K. K. Parhi, "Vlsi architectures for the restricted boltzmann machine," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 13, no. 3, p. 35, 2017. [Online]. Available: https://doi.org/10.1145/3007193.

[80] L. Lopez-Diaz, L. Torres, and E. Moro, "Transition from ferromagnetism to superparamagnetism on the nanosecond time scale," *Physical Review B*, vol. 65, no. 22, p. 224 406, 2002.

[81] W. F. Brown Jr, "Thermal fluctuations of a single-domain particle," *Physical Review*, vol. 130, no. 5, p. 1677, 1963.

[82] W. T. Coffey and Y. P. Kalmykov, "Thermal fluctuations of magnetic nanoparticles: Fifty years after brown," *Journal of Applied Physics*, vol. 112, no. 12, p. 121 301, 2012.

[83] J. Kaiser, A. Rustagi, K. Y. Camsari, J. Z. Sun, S. Datta, and P. Upadhyaya, "Subnanosecond fluctuations in low-barrier nanomagnets," *Physical Review Applied*, vol. 12, no. 5, p. 054 056, 2019.

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

[85] S. Sayed, K. Y. Camsari, R. Faria, and S. Datta, "Rectification in spin-orbit materials using low-energy-barrier magnets," *Physical Review Applied*, vol. 11, no. 5, p. 054 063, 2019.

[86] A. Z. Pervaiz, L. A. Ghantasala, K. Y. Camsari, and S. Datta, "Hardware emulation of stochastic p-bits for invertible logic," *Scientific reports*, vol. 7, no. 1, p. 10 994, 2017.

[87] K. Y. Camsari, S. Ganguly, and S. Datta, "Modular approach to spintronics," *Scientific reports*, vol. 5, p. 10 571, 2015.

[88] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "Predictive technology model," *Internet: http://ptm. asu. edu*, 2002.

[89] D. E. Nikonov and I. A. Young, "Benchmarking of beyond-cmos exploratory devices for logic integrated circuits," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 1, pp. 3–11, 2015.

[90] R. Zand, K. Y. Camsari, S. Datta, and R. F. DeMara, "Composable probabilistic inference networks using mram-based stochastic neurons," *arXiv preprint arXiv:1811.11390*, 2018.

[91] F. Neukart, G. Compostella, C. Seidel, D. Von Dollen, S. Yarkoni, and B. Parney, "Traffic flow optimization using a quantum annealer," *Frontiers in ICT*, vol. 4, p. 29, 2017.

[92] F. Barahona, M. Grötschel, M. Jünger, and G. Reinelt, "An application of combinatorial optimization to statistical physics and circuit layout design," *Operations Research*, vol. 36, no. 3, pp. 493–513, 1988.

[93] C. Cook, H. Zhao, T. Sato, M. Hiromoto, and S. X.-D. Tan, "Gpu based parallel ising computing for combinatorial optimization problems in vlsi physical design," *arXiv preprint arXiv:1807.10750*, 2018.

[94] G. Rosenberg, P. Haghnegahdar, P. Goddard, P. Carr, K. Wu, and M. L. De Prado, "Solving the optimal trading trajectory problem using a quantum annealer," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1053–1060, 2016.

[95] H. Sakaguchi, K. Ogata, T. Isomura, S. Utsunomiya, Y. Yamamoto, and K. Aihara, "Boltzmann sampling by degenerate optical parametric oscillator network for structure-based virtual screening," *Entropy*, vol. 18, no. 10, p. 365, 2016.

[96] L. Xia, P. Gu, B. Li, T. Tang, X. Yin, W. Huangfu, S. Yu, Y. Cao, Y. Wang, and H. Yang, "Technological exploration of rram crossbar array for matrix-vector multiplication," *Journal of Computer Science and Technology*, vol. 31, no. 1, pp. 3–19, 2016.

[97] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, and W. D. Lu, "A fully integrated reprogrammable memristor–cmos system for efficient multiply–accumulate operations," *Nature Electronics*, vol. 2, no. 7, pp. 290–299, 2019.

[98] F. M. Bayat, M. Prezioso, B. Chakrabarti, H. Nili, I. Kataeva, and D. Strukov, "Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits," *Nature communications*, vol. 9, no. 1, pp. 1–7, 2018.

[99] M. M. Torunbalci, P. Upadhyaya, S. A. Bhave, and K. Y. Camsari, "Modular compact modeling of mtj devices," *IEEE Transactions on Electron Devices*, vol. 65, no. 10, pp. 4628–4634, 2018.

[100] M. W. Daniels, A. Madhavan, P. Talatchian, A. Mizrahi, and M. D. Stiles, "Energy-efficient stochastic computing with superparamagnetic tunnel junctions," *Physical Review Applied*, vol. 13, no. 3, p. 034 016, 2020.

[101] J. Grollier, D. Querlioz, K. Camsari, K. Everschor-Sitte, S. Fukami, and M. D. Stiles, "Neuromorphic spintronics," *Nature Electronics*, pp. 1–11, 2020.

[102] M. A. Abeed and S. Bandyopadhyay, "Low energy barrier nanomagnet design for binary stochastic neurons: Design challenges for real nanomagnets with fabrication defects," *IEEE Magnetics Letters*, vol. 10, pp. 1–5, 2019.

[103] J. L. Drobitch and S. Bandyopadhyay, "Reliability and scalability of p-bits implemented with low energy barrier nanomagnets," *IEEE Magnetics Letters*, vol. 10, pp. 1–4, 2019.

[104] B. Parks, A. Abdelgawad, T. Wong, R. F. Evans, and S. A. Majetich, "Magnetoresistance dynamics in superparamagnetic co- fe- b nanodots," *Physical Review Applied*, vol. 13, no. 1, p. 014 063, 2020.

[105] S. Cheemalavagu, P. Korkmaz, K. V. Palem, B. E. Akgul, and L. N. Chakrapani, "A probabilistic cmos switch and its realization by exploiting noise," in *IFIP International Conference on VLSI*, 2005, pp. 535–541.

[106] N. Shukla, A. Parihar, E. Freeman, H. Paik, G. Stone, V. Narayanan, H. Wen, Z. Cai, V. Gopalan, R. Engel-Herbert, *et al.*, "Synchronized charge oscillations in correlated electron systems," *Scientific reports*, vol. 4, p. 4964, 2014.

[107] S. Kumar, J. P. Strachan, and R. S. Williams, "Chaotic dynamics in nanoscale nbo 2 mott memristors for analogue computing," *Nature*, vol. 548, no. 7667, pp. 318–321, 2017.

[108] B. Stampfer, F. Zhang, Y. Y. Illarionov, T. Knobloch, P. Wu, M. Waltl, A. Grill, J. Appenzeller, and T. Grasser, "Characterization of single defects in ultrascaled mos 2 field-effect transistors," *ACS nano*, vol. 12, no. 6, pp. 5368–5375, 2018.

[109] J. Cai, B. Fang, L. Zhang, W. Lv, B. Zhang, T. Zhou, G. Finocchio, and Z. Zeng, "Voltage-controlled spintronic stochastic neuron based on a magnetic tunnel junction," *Physical Review Applied*, vol. 11, no. 3, p. 034 015, 2019.

[110] C. Lin, S. Kang, Y. Wang, K. Lee, X. Zhu, W. Chen, X. Li, W. Hsu, Y. Kao, M. Liu, *et al.*, "45nm low power cmos logic compatible embedded stt mram utilizing a reverse-connection 1t/1mtj cell," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, IEEE, 2009, pp. 1–4.

[111] Y. Lv, R. P. Bloom, and J.-P. Wang, "Experimental demonstration of probabilistic spin logic by magnetic tunnel junctions," *IEEE Magnetics Letters*, vol. 10, pp. 1–5, 2019.

[112] B. R. Zink, Y. Lv, and J.-P. Wang, "Independent control of antiparallel-and parallel-state thermal stability factors in magnetic tunnel junctions for telegraphic signals with two degrees of tunability," *IEEE Transactions on Electron Devices*, vol. 66, no. 12, pp. 5353–5359, 2019.

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

[114] S. Ikeda, J. Hayakawa, Y. Ashizawa, Y. Lee, K. Miura, H. Hasegawa, M. Tsunoda, F. Matsukura, and H. Ohno, "Tunnel magnetoresistance of 604% at 300 k by suppression of ta diffusion in co fe b/ mg o/ co fe b pseudo-spin-valves annealed at high temperature," *Applied Physics Letters*, vol. 93, no. 8, p. 082 508, 2008.

[115] M. R. Pufall, W. H. Rippard, S. Kaka, S. E. Russek, T. J. Silva, J. Katine, and M. Carey, "Large-angle, gigahertz-rate random telegraph switching induced by spin-momentum transfer," *Physical Review B*, vol. 69, no. 21, p. 214 409, 2004.

[116] C. Safranski, J. Kaiser, P. Trouilloud, P. Hashemi, G. Hu, and J. Z. Sun, "Demonstration of nanosecond operation in stochastic magnetic tunnel junctions," *arXiv preprint arXiv:2010.14393*, 2020.

[117] M. Romera, P. Talatchian, S. Tsunegi, F. A. Araujo, V. Cros, P. Bortolotti, J. Trastoy, K. Yakushiji, A. Fukushima, H. Kubota, *et al.*, "Vowel recognition with four coupled spin-torque nano-oscillators," *Nature*, vol. 563, no. 7730, pp. 230–234, 2018.

[118] S. Jenkins, A. Meo, L. E. Elliott, S. K. Piotrowski, M. Bapna, R. W. Chantrell, S. A. Majetich, and R. F. Evans, "Magnetic stray fields in nanoscale magnetic tunnel junctions," *Journal of Physics D: Applied Physics*, vol. 53, no. 4, p. 044 001, 2019.

[119] R. Faria, J. Kaiser, K. Y. Camsari, and S. Datta, "Hardware design for autonomous bayesian networks," *arXiv preprint arXiv:2003.01767*, 2020.

[120] S. V. Isakov, I. N. Zintchenko, T. F. Rønnow, and M. Troyer, "Optimised simulated annealing for ising spin glasses," *Computer Physics Communications*, vol. 192, pp. 265–271, 2015.

[121] E. Aarts, E. H. Aarts, and J. K. Lenstra, *Local search in combinatorial optimization*. Princeton University Press, 2003.

[122] M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, *et al.*, "Memristor-based analog computation and neural network classification with a dot product engine," *Advanced Materials*, vol. 30, no. 9, p. 1 705 914, 2018.

[123] H. Gyoten, M. Hiromoto, and T. Sato, "Area efficient annealing processor for ising model without random number generator," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 2, pp. 314–323, 2018.

[124] S. Aggarwal, H. Almasi, M. DeHerrera, B. Hughes, S. Ikegawa, J. Janesky, H. Lee, H. Lu, F. Mancoff, K. Nagel, *et al.*, "Demonstration of a reliable 1 gb standalone spin-transfer torque mram for industrial applications," in *2019 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2019, pp. 2–1.

[125] X. Zhang, R. Bashizade, Y. Wang, C. Lyu, S. Mukherjee, and A. R. Lebeck, "Beyond application end-point results: Quantifying statistical robustness of mcmc accelerators," *arXiv preprint arXiv:2003.04223*, 2020.

[126] S. Nasrin, J. L. Drobitch, S. Bandyopadhyay, and A. R. Trivedi, "Low power restricted boltzmann machine using mixed-mode magneto-tunneling junctions," *IEEE Electron Device Letters*, vol. 40, no. 2, pp. 345–348, 2019.

[127] C. D. Schuman, T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, G. S. Rose, and J. S. Plank, "A survey of neuromorphic computing and neural networks in hardware," *arXiv preprint arXiv:1705.06963*, 2017.

[128] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training neural networks with weights and activations constrained to+ 1 or-1," *arXiv preprint arXiv:1602.02830*, vol. 2, 2016.

[129] C.-H. Tsai, W.-J. Yu, W. H. Wong, and C.-Y. Lee, "A 41.3/26.7 pj per neuron weight rbm processor supporting on-chip learning/inference for iot applications," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 10, pp. 2601–2612, 2017.

[130] S. Park, K. Bong, D. Shin, J. Lee, S. Choi, and H.-J. Yoo, "93tops/w scalable deep learning/inference processor with tetra-parallel mimd architecture for big-data applications," in *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*, IEEE, 2015, pp. 1–3.

[131] T. Shibata and T. Ohmi, "A functional mos transistor featuring gate-level weighted sum and threshold operations," *IEEE Transactions on Electron devices*, vol. 39, no. 6, pp. 1444–1455, 1992.

[132] N. Nakamura, K. Shimada, T. Matsuda, and M. Kimura, "Neuron mos inverter and source follower using thin-film transistors," in *Future of Electron Devices, Kansai (IMFEDK), 2015 IEEE International Meeting for*, IEEE, 2015, pp. 90–91.

[133] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2009.

[134] F. L. Traversa and M. Di Ventra, "Polynomial-time solution of prime factorization and np-complete problems with digital memcomputing machines," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 2, p. 023 107, 2017.

[135] R. Zand, K. Y. Camsari, I. Ahmed, S. D. Pyle, C. H. Kim, S. Datta, and R. F. DeMara, "R-dbn: A resistive deep belief network architecture leveraging the intrinsic behavior of probabilistic devices," *arXiv preprint arXiv:1710.00249*, 2017.

[136] K. Y. Camsari, R. Faria, O. Hassan, B. M. Sutton, and S. Datta, "Equivalent circuit for magnetoelectric read and write operations," *Physical Review Applied*, vol. 9, no. 4, p. 044 020, 2018.

[137] A. K. Biswas, H. Ahmad, J. Atulasimha, and S. Bandyopadhyay, "Experimental demonstration of complete 180° reversal of magnetization in isolated co nanomagnets on a pmn–pt substrate with voltage generated strain," *Nano letters*, vol. 17, no. 6, pp. 3478–3484, 2017.

[138] K. Roy, S. Bandyopadhyay, and J. Atulasimha, "Hybrid spintronics and straintronics: A magnetic technology for ultra low energy computing and signal processing," *Applied Physics Letters*, vol. 99, no. 6, p. 063 108, 2011.

[139] N. Kani, J. T. Heron, and A. Naeemi, "Strain-mediated magnetization reversal through spin-transfer torque," *IEEE Transactions on Magnetics*, vol. 53, no. 11, pp. 1–8, 2017.

[140] A. Jaiswal and K. Roy, "Mesl: Proposal for a non-volatile cascadable magneto-electric spin logic," *Scientific reports*, vol. 7, p. 39 793, 2017.

[141] S. Manipatruni, D. E. Nikonov, R. Ramesh, H. Li, and I. A. Young, "Spin-orbit logic with magnetoelectric nodes: A scalable charge mediated nonvolatile spintronic logic," *arXiv preprint arXiv:1512.05428*, 2015.

[142] T. Gao, X. Zhang, W. Ratcliff, S. Maruyama, M. Murakami, A. Varatharajan, Z. Yamani, P. Chen, K. Wang, H. Zhang, *et al.*, "Electric-field induced reversible switching of the magnetic easy axis in co/bifeo3 on srtio3," *Nano letters*, vol. 17, no. 5, pp. 2825–2832, 2017.

[143] J. Heron, J. Bosse, Q. He, Y. Gao, M. Trassin, L. Ye, J. Clarkson, C. Wang, J. Liu, S. Salahuddin, *et al.*, "Deterministic switching of ferromagnetism at room temperature using an electric field," *Nature*, vol. 516, no. 7531, p. 370, 2014.

[144] X. He, Y. Wang, N. Wu, A. N. Caruso, E. Vescovo, K. D. Belashchenko, P. A. Dowben, and C. Binek, "Robust isothermal electric control of exchange bias at room temperature," *Nature materials*, vol. 9, no. 7, p. 579, 2010.

[145] Z. Zhao, W. Echtenkamp, M. Street, C. Binek, and J.-P. Wang, "Magnetoelectric device feasibility demonstration—voltage control of exchange bias in perpendicular cr 2 o 3 hall bar device," in *Device Research Conference (DRC), 2016 74th Annual*, IEEE, 2016, pp. 1–2.

[146] P. K. Amiri and K. L. Wang, "Voltage-controlled magnetic anisotropy in spintronic devices," in *Spin*, World Scientific, vol. 2, 2012, p. 1 240 002.

[147] D. Chien, X. Li, K. Wong, M. A. Zurbuchen, S. Robbennolt, G. Yu, S. Tolbert, N. Kioussis, P. Khalili Amiri, K. L. Wang, *et al.*, "Enhanced voltage-controlled magnetic anisotropy in magnetic tunnel junctions with an mgo/pzt/mgo tunnel barrier," *Applied Physics Letters*, vol. 108, no. 11, p. 112 402, 2016.

[148] S. K. Piotrowski, M. Bapna, S. D. Oberdick, S. A. Majetich, M. Li, C. Chien, R. Ahmed, and R. Victora, "Size and voltage dependence of effective anisotropy in sub-100-nm perpendicular magnetic tunnel junctions," *Physical Review B*, vol. 94, no. 1, p. 014 404, 2016.

[149] M. G. Mankalale, Z. Liang, Z. Zhao, C. H. Kim, J.-P. Wang, and S. S. Sapatnekar, "Comet: Composite-input magnetoelectric-based logic technology," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 3, pp. 27–36, 2017.

[150] A. Khan, D. E. Nikonov, S. Manipatruni, T. Ghani, and I. A. Young, "Voltage induced magnetostrictive switching of nanomagnets: Strain assisted strain transfer torque random access memory," *Applied Physics Letters*, vol. 104, no. 26, p. 262 407, 2014.

[151] N. Pertsev, "Giant magnetoelectric effect via strain-induced spin reorientation transitions in ferromagnetic films," *Physical Review B*, vol. 78, no. 21, p. 212 102, 2008.

[152] R.-C. Peng, J.-M. Hu, L.-Q. Chen, and C.-W. Nan, "On the speed of piezostrain-mediated voltage-driven perpendicular magnetization reversal: A computational elastodynamics-micromagnetic phase-field study," *NPG Asia Materials*, vol. 9, no. 7, e404, 2017.

[153] R. M. Iraei, S. Dutta, S. Manipatruni, D. E. Nikonov, I. A. Young, J. T. Heron, and A. Naeemi, "A proposal for a magnetostriction-assisted all-spin logic device," in *Device Research Conference (DRC), 2017 75th Annual*, IEEE, 2017, pp. 1–2.

[154] S. Sharmin, Y. Shim, and K. Roy, "Magnetoelectric oxide based stochastic spin device towards solving combinatorial optimization problems," *Scientific Reports*, vol. 7, no. 1, p. 11 276, 2017.

[155] N. Tiercelin, Y. Dusch, A. Klimov, S. Giordano, V. Preobrazhensky, and P. Pernod, "Room temperature magnetoelectric memory cell using stress-mediated magnetoelastic switching in nanostructured multilayers," *Applied Physics Letters*, vol. 99, no. 19, p. 192 507, 2011.

[156] A. Klimov, N. Tiercelin, Y. Dusch, S. Giordano, T. Mathurin, P. Pernod, V. Preobrazhensky, A. Churbanov, and S. Nikitov, "Magnetoelectric write and read operations in a stress-mediated multiferroic memory cell," *Applied Physics Letters*, vol. 110, no. 22, p. 222 401, 2017.

[157] S. Manipatruni, D. E. Nikonov, and I. A. Young, "Beyond cmos computing with spin and polarization," *Nature Physics*, vol. 14, no. 4, pp. 338–343, 2018.

[158] P. K. Amiri, J. G. Alzate, X. Q. Cai, F. Ebrahimi, Q. Hu, K. Wong, C. Grèzes, H. Lee, G. Yu, X. Li, *et al.*, "Electric-field-controlled magnetoelectric ram: Progress, challenges, and scaling," *IEEE Transactions on Magnetics*, vol. 51, no. 11, pp. 1–7, 2015.

[159] K. Roy, S. Bandyopadhyay, and J. Atulasimha, "Binary switching in a 'symmetric'potential landscape," *Scientific reports*, vol. 3, p. 3038, 2013.

[160] T. Shen, V. Ostwal, K. Y. Camsari, and J. Appenzeller, "Demonstration of a pseudomagnetization based simultaneous write and read operation in a co 60 fe 20 b 20/pb (mg 1/3 nb 2/3) 0.7 ti 0.3 o 3 heterostructure," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.

[161] A. Sheikholeslami and P. G. Gulak, "A survey of circuit innovations in ferroelectric random-access memories," *Proceedings of the IEEE*, vol. 88, no. 5, pp. 667–689, 2000.

[162] e. a. Tingting Shen Orchi Hassan, "Demonstration of pseudo-magnetization based write operation in cofeb films and nanodots using ferromagnetic resonance," *(in preparation)*,

A. DERIVATION OF PINNING FIELD OF LBM

Magnets are generally used to store information putting the focus on the evaluating and predicting characteristics of stable high-barrier magnets. It is interesting to note that theoretical predictions and analytical derivations regarding low-barrier magnet ($\Delta \leq k_B T$) dynamics typically receive less attention as cases of 'least practical interest'[81]. We document the analytical expressions associated with LBM in fig. 3.10. The expressions for correlation time and biasing current can be found in ref.[60], [82], [83], [85], in this appendix we derive the bias field.

We derive the expressions for external magnetic field H_0 required to pin the magnetization of an LBM with $\Delta \leq k_B T$ here. We start from the energy expression for the magnet (E) and derive the expressions presented in fig. 3.10 from the steady-state average magnetization defined by:

$$\langle m \rangle = \frac{\int_{\theta=0}^{\theta=\pi} \int_{\phi=-\pi}^{\phi=\pi} \sin\theta \; d\phi \; d\theta \; m \exp(-E/k_B T)}{\int_{\theta=0}^{\theta=\pi/2} \int_{\phi=-\pi}^{\phi=\pi} \sin\theta \; d\phi \; d\theta \; \exp(-E/k_B T)} \tag{A.1}$$

where $(m_x, m_y, m_z) \equiv (\cos\theta, \sin\theta \sin\phi, \sin\theta \cos\phi)$.

A.1 Perpendicular Magnetic Anisotropy (PMA)

In case of LBM with perpendicular magnetization, the anisotropy field along x-axis $H_{kp} \to 0$ and thus for a field applied in the x-direction the energy expression eq. 3.1 is reduced to :

$$E = -H_{ext} M_S \Omega \; m_x \tag{A.2}$$

Evaluation eq. A.1 wrt to this energy gives us: $\langle m_x \rangle = \coth(H_{ext} M_S \Omega / k_B T) - (H_{ext} M_S \Omega / k_B T) \approx \tanh(H_{ext} M_S \Omega / 3 k_B T)$. So to pin the magnetization to any of its state $\langle m_x \rangle = \pm 1$, the required external field for PMA magnets can be approximated by:

$$|H_{ext(PMA)}| = \frac{3 k_B T}{M_s \Omega} \tag{A.3}$$

A.2 In-plane Magnetic Anisotropy (IMA)

For LBM with in-plane magnets, the anisotropy field along z-axis $H_{ki} \rightarrow 0$ and a large demagnetization field H_D exists along the z-axis which keeps the magnetization in-plane. The energy expression from eq. 3.1 in this case is :

$$E = H_D M_S \Omega \, m_x^2 - H_{ext} M_S \Omega \, m_z. \tag{A.4}$$

Once again evaluating eq. A.1 wrt to this energy for very large demagnetizing field ($H_D \rightarrow \infty$) can be simplified to $\langle m_z \rangle \approx H_{ext} M_S \Omega / 2k_B T$. So to pin the magnetization to any of its state $\langle m_z \rangle = \pm 1$, the required external field for IMA magnets can be approximated by:

$$|H_{ext(IMA)}| = \frac{2k_B T}{M_s \Omega} \tag{A.5}$$

The expression is independent of the demagnetization field. These empirical expressions match our SPICE simulation results quite well as shown in fig. A.1.

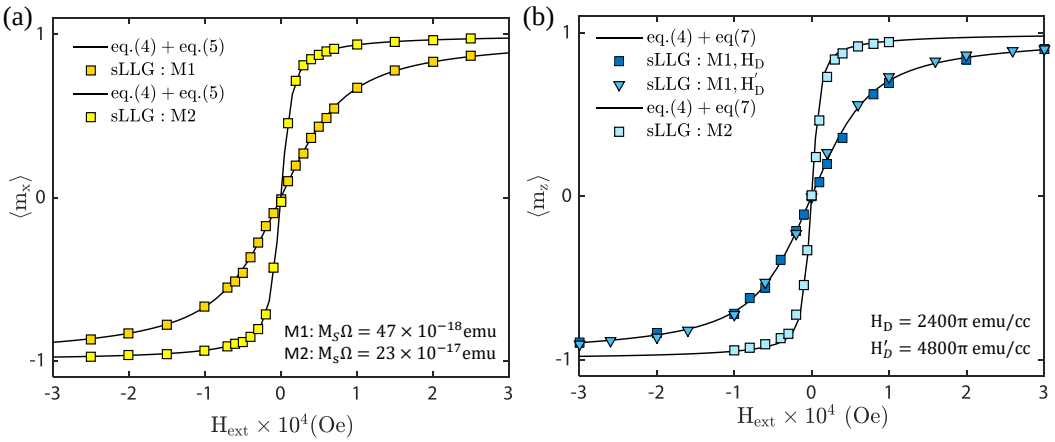


Figure A.1. Pinning Field of low-barrier magnets The numerical evaluations of equations are compared to SPICE simulation for (a) Isotropic magnets and (b) circular IMA magnets which have $\Delta \leq k_B T$. The pinning fields are shown to be a function of $M_S\Omega$ only where M_S = 600 emu/cc and the volume of magnet Ω is varied, The pinning field values for IMA magnets indicate that it is independent of the large demagnetization field, H_D. The precise correspondence between the analytical formulas and the numerical simulation also constitutes as a benchmark to our finite temperature (stochastic) LLG formulation.

B. P-BIT DESIGN CRITERIA FROM BEHAVIORAL MODEL

Independent of the technology being used, probabilistic bits need to fulfill three key criteria. The necessary conditions that the p-bit needs to satisfy can be interpreted from the mathematical description of the p-bit given by:

$$m_{\mathrm{i}} = sgn[tanh(I_{\mathrm{i}}) + r_{\mathrm{i}}] \tag{B.1}$$

where I_{i} is the input to the p-bit which tunes its probability and r_{i} is a continuous random variable that provides stochasticity. The *sgn* function ensures that the m_{i} values are thresholded between ± 1 values. But is it necessary for r_{i} to be continuous and m_{i} to be thresholded? In chapter. 3 we showed that bipolar r_{i} did not work. In this section we elaborate on these necessary conditions.

Eq. B.1 can be modified to represent a bipolar and un-thresholded p-bit design:

$$\begin{aligned} \text{Bipolar } \mathrm{r_i} \ &: \ m_{\mathrm{i}} = sgn[tanh(I_{\mathrm{i}}) + sgn(r_{\mathrm{i}})]; \\ \text{Unsigned } \mathrm{m_i} \ &: \ m_{\mathrm{i}} = tanh(I_{\mathrm{i}}) + r_{\mathrm{i}}; \end{aligned} \tag{B.2}$$

We look at the performance of each of the p-bit behavioral models in a probabilistic spin logic framework by designing two Boltzmann machines (BMs) for performing invertible Boolean logic, namely the AND gate and 1 bit full-adder (FA). The p-bits are correlated through their individual inputs $\mathrm{I_i} = \sum \mathrm{J_{ij} m_j} + \mathrm{h_i}$, where J_{ij}s are the coupling coefficients and h_{i} are the individual bias terms [31]. The numerical probabilities of the system should agree with the probabilities predicted from the energy function defined by

$$E(\mathbf{m}) = -\frac{1}{2}\sum_{\mathrm{ij}} J_{\mathrm{ij}} m_{\mathrm{i}} m_{\mathrm{j}} - \sum_{\mathrm{i}} h_{\mathrm{i}} m_{\mathrm{i}} \tag{B.3}$$

using the Boltzmann law:

$$p(\mathbf{m}) = \frac{\exp\left(-E\right)}{\sum_{\mathrm{i,j}} \exp\left(-E\right)} \tag{B.4}$$

Fig. B.1 shows the transfer characteristic and probability distribution for the gates for each p-bit model along with the expected Boltzmann distributions. The bipolar and unsigned

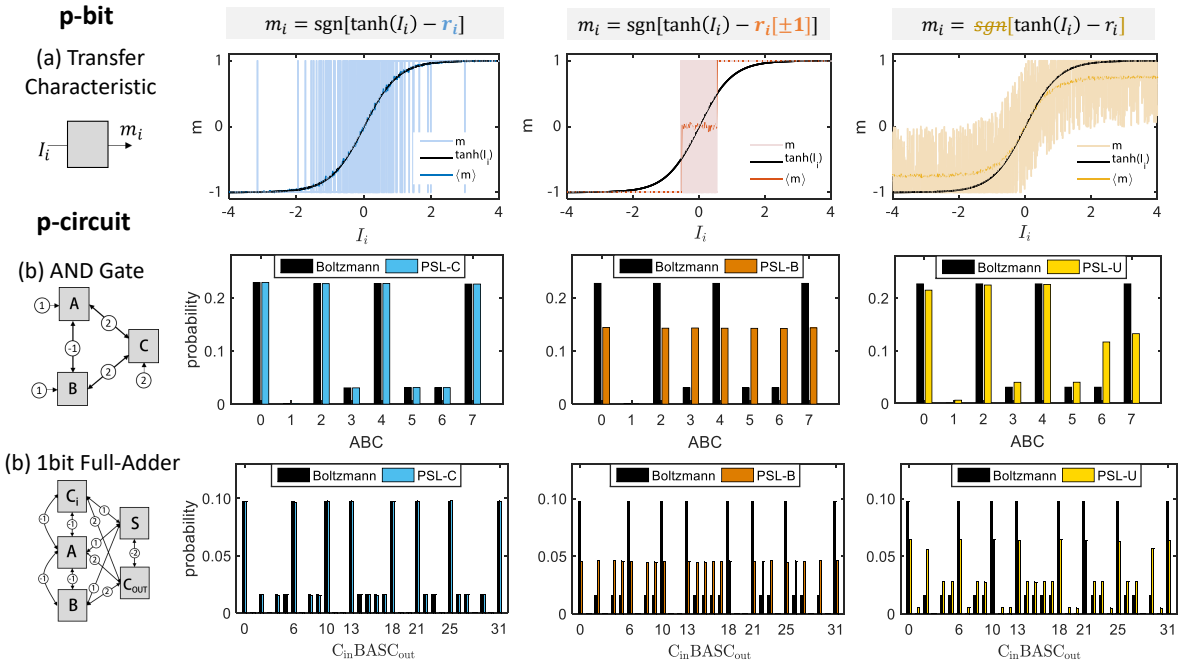


Figure B.1. Behavioral Models: p-bit (a)Transfer Characteristics and p-circuit implementations showing (b) AND Gate operation and (c) 1-bit Full Adder operation for three different behavioral representations of p-bits. Only the p-bit model expressed by eq. B.1 with thresholding and continuous random variable r_i is able to reproduce the Bolzmann distribution exactly.

p-bit models fail to agree with the expected Boltzmann distributions. We demonstrate two necessary requirements of p-bit behavior through this exercise, namely (a) the r_i needs to be continuous meaning the source of randomness in hardware has to have a continuous distribution and (b) the output m_i needs to be thresholded meaning hardware would require a thresholding circuit (like the inverter) of some-sort.

C. CODES

The SPICE modules for the spintronic device elements used in this thesis are available at https : //nanohub.org/groups/spintronics [58] and the transistor models are available at http : //ptm.asu.edu/ [59].

The codes used for generating the figures are available upon request to the author (email: hassan19@purdue.edu, orchi.hassan@gmail.com).

PUBLICATIONS

1. "Voltage driven building block for hardware belief network," O. Hassan, K. Y. Camsari, and S. Datta. *IEEE Design & Test*, vol.36, Jun 2019

2. "Low Barrier Magnet Design for Efficient Hardware Binary Stochastic Neurons," O. Hassan, R. Faria, K. Y. Camsari, J. Z. Sun, and S. Datta. *IEEE Magnetic Letters*, vol.10, Apr 2019.

3. "Quantitative Evaluation of Hardware Binary Stochastic Neurons," O. Hassan, S. Datta, and K. Y. Camsari. (*to be submitted*)

4. "Equivalent circuit for Magnetoelectric Read and Write Operations," K. Y. Camsari, R. Faria, O. Hassan, B. M. Sutton, and S. Datta, *Physical Rev. App.*, vol.9, Apr 2018.

5. "Energy Efficient Magnetoelectric Memory Device Based on Pseudo-Magnetization," O. Hassan, T. Shen, N. R. Dilley, V. Ostwal, P. Upadhyaya, J, Appenzeller, Supriyo Datta, and K. Y. Camsari. (*Unpublished*, presented at MMM 2019)

6. "Demonstration of Pseudo-Magnetization Based Write Operation in CoFeB Films and Nanodots Using Ferromagnetic Resonance," T. Shen, O. Hassan, et. al. (*in preparation*)

7. "p-transistors and p-circuits for Boolean and non-Boolean logic," K. Y. Camsari, R. Faria, O. Hassan, A. Z. Pervaiz, B. M. Sutton, and S. Datta. *Spintronics X*, vol. 10357, Sep. 2017.

VITA

Orchi Hassan is a PhD candidate in the school of Electrical and Computer Engineering (ECE) at Purdue University, West Lafayette, Indiana. She is currently working as a research assistant in Professor Supriyo Datta group. She received her Bachelor's (B.S.) and Master's (M.S.) degree in Electronic Engineering (EEE) from Bangladesh University of Engineering and Technology (BUET) in 2012 and 2014, respectively. Her undergraduate research was focused on analyzing Quantum Cascade Lasers (QCL). Her current research is on applying magnet and circuit physics for evaluating spintronic devices. Her focus is on evaluating the performance of low-barrier magnet based spintronic devices for compact and energy-efficient realization of probabilistic hardware.