LEVERAGING THE INTRINSIC SWITCHING BEHAVIORS OF SPINTRONIC DEVICES FOR DIGITAL AND NEUROMORPHIC CIRCUITS

by

STEVEN D. PYLE
M.S. University of Central Florida 2015

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Electrical and Computer Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2019

Major Professor: Ronald F. DeMara

ABSTRACT

With semiconductor technology scaling approaching atomic limits, novel approaches utilizing new memory and computation elements are sought in order to realize increased density, enhanced functionality, and new computational paradigms. Spintronic devices offer intriguing avenues to improve digital circuits by leveraging *non-volatility* to reduce static power dissipation and vertical integration for increased density. Novel hybrid spintronic-CMOS digital circuits are developed herein that illustrate enhanced functionality at reduced static power consumption and area cost. The developed spin-CMOS D Flip-Flop offers improved power-gating strategies by achieving instant store/restore capabilities while using 10 fewer transistors than typical CMOS-only implementations. The spin-CMOS Muller C-Element developed herein improves asynchronous pipelines by reducing the area overhead while adding enhanced functionality such as instant data store/restore and delay-element-free bundled data asynchronous pipelines.

Spintronic devices also provide improved scaling for neuromorphic circuits by enabling compact and low power neuron and non-volatile synapse implementations while enabling new neuromorphic paradigms leveraging the stochastic behavior of spintronic devices to realize stochastic spiking neurons, which are more akin to biological neurons and commensurate with theories from computational neuroscience and probabilistic learning rules. Spintronic-based Probabilistic Activation Function circuits are utilized herein to provide a compact and low-power neuron for Binarized Neural Networks. Two implementations of stochastic spiking neurons with alternative speed, power, and area benefits are realized. Finally, a comprehensive neuromorphic architecture comprising stochastic spiking neurons, low-precision synapses with *Probabilistic*

Hebbian Plasticity, and a novel non-volatile homeostasis mechanism is realized for subthreshold ultra-low-power unsupervised learning with robustness to process variations. Along with several case studies, implications for future spintronic digital and neuromorphic circuits are presented.

ACKNOWLEDGMENTS

Life is a funny thing. In a universe compelled towards maximizing entropy, life seems to boldly run up stream. The life of each individual is more peculiar still, for by no will of their own they are given a will and thrust into reality and consciousness, expected to make something useful out of it. Most people find their meaning traveling down the path well worn, changing things little, but making the most of it where they can. Some people are lucky enough to meet the right people at the right time who give them the tools to forge new paths into a bigger world of possibilities then they ever thought was possible. I consider myself ridiculously lucky and eternally grateful for the people that chance, fate, God, or what-have-you graciously brought into my life at just the right times to elicit the immense personal growth that made this dissertation, and my future, possible.

I want to specifically thank my parents, who gave me the freedom and support to explore any direction my heart desired. I must thank Dr. Ronald F. DeMara and the shear seconds of coincidence that lead me to his class, for his unrelenting willfulness to teach, inspire, discuss, include, help, support, and take-the-kid-gloves-off was so much of what I needed that I will never be able to fully express my gratefulness for that man. I want to thank Dr. Deliang Fan, who's research and classwork on neuromorphic architectures helped to lay the bedrock for much of this dissertation. I want to thank Dr. Pamela Douglas, who introduced me to the circuits and large-scale behaviors of the brain and encouraged me in my exploration of bold new paths. I want to thank Drs. Reza Abdolvand, Mingjie Lin, and Vikram Kapoor for their kind words, encouragement, and support on my dissertation committee. I want to thank Ramtin Zand, Soheil Salehi, Arman Roohi, and Navid Khoshavi for the endless hours of quality conversation that undoubtable shaped myself

and my work. Finally, I must thank Orlando Arias, who was always able to help me debug what stack exchange could not.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Scalable, energy-efficient, and enhanced functionality over CMOS technology are all desirable characteristics for future computational devices. Emerging spintronic devices achieve greater functionality through nonvolatility and improved scalability via Back End of Line (BEoL) compatibility, which enables vertical integration [1]. By utilizing these features, enhancements to contemporary computational architectures can be realized, such as area-efficient digital circuits with instant store/restore functionality for aggressive power gating, while new neuromorphic computational architectures can leverage the dense arrays of non-volatile memory, and the intrinsic properties of spintronic devices can be used to realize entirely new computational paradigms at even greater energy efficiency [2, 3]. This Chapter introduces the need for spintronic architectures, provides an overview of current spintronic approaches for digital and neuromorphic circuits, and delineates the contributions of this Dissertation.

Need for Spintronic Circuits

The Moore's Law scaling of CMOS devices has enabled the proliferation of computational technology in every facet of the information processing revolution since the 1960s. However, the fundamental limitations of CMOS scaling has necessitated the semiconductor industry to formally acknowledge that transistors will stop shrinking by the early 2020s, as emphasized by the chairman of the road-mapping organization [4]. Means for continuing Moore's Law or enhancing the capabilities achievable with current integration densities could be through the development of new

nanodevices. Among promising devices, the 2015 International Technology Roadmap for Semiconductors (ITRS) identifies nanomagnetic, or spintronic, devices as capable post-CMOS candidates [5].

Spintronic devices have the potential to operate at frequencies above 1 GHz and at energies approaching 1aJ [6]. Thus, they provide a direction for a scalable universal memory technology, which is in contrast to today's current memory systems that are segmented between SRAM for high speed/high area, DRAM for moderate speed/low area, and Flash or magnetic platters for very low speed/very low area data storage. Additionally, the features of spintronic devices enable new classes of circuits that, in contrast with CMOS, intrinsically hold their state without any external power signals. This feature can provide novel circuit and architectural strategies leveraging power-gating for reduced energy consumption and heat generation, as well as the development of new computational strategies beyond typical von-Neumann approaches, such as neuromorphic circuits, where the improved integration density of high-speed non-volatile devices can decrease area overheads for the memory-intensive nature of neural network paradigms, while the stochastic switching properties of the devices enable the implementation of stochastic behaviors found in biological neural networks and probabilistic unsupervised learning rules using a minimal number of discrete binary memory devices [2, 7, 8].

Spintronic Technology

In this Section, the spintronic switching mechanisms outlined in Figure 3 are delineated.

To begin, almost all currently-commercializable spintronic devices utilize a Magnetic Tunnel Junction (MTJ) for switching, reading the magnetic state, or both, and the MTJ is realized with a material stack of a thin insulating oxide, which is typically MgO, sandwiched between two Ferromagnets (FM) [1, 9-25].

Figure 1: Illustration of the MTJ stack, high resistance state, and low resistance state.

As shown in Figure 1, one FM is deemed the fixed layer since it is engineered such that its magnetic orientation will not change, and the other FM is called the free layer since its magnetic orientation will rotate according to the principles of the specific device it is a part of. The MTJ acts as a mechanism to read the state of magnetic orientation of a FM since the resistance across the MTJ is based on the relative magnetic orientations of the fixed- and free-layer FMs in the stack. If the two FMs are parallel (P), then the MTJ has its lowest resistance, and if they are anti-parallel (AP), then the MTJ has its highest resistance. The relative resistance ratio between the P and AP states is called the Tunneling Magnetoresistance (TMR) ratio. MTJs will be used throughout the dissertation to interface between spin-states and electrical circuits.

One of the primary mechanisms by which the free-layer of an MTJ can be switched is with Spin Transfer Torque (STT) [26, 27]. STT operates by injecting a spin-polarized charge current into the free-layer. If electrons of one spin orientation meet an electron of differing orientation,

then a net magnetic torque is applied, which is STT. MTJs can be utilized to polarize the spin of electrons through its fixed-layer before passing into the free-layer much like how a polarizer only allows photons of a particular polarization to pass through it [28]. Other physical phenomena, such as Spin Orbit Coupling discussed later, are also able to inject spin-polarized currents into free layers to switch the magnetic state. STT will be used throughout this dissertation as a mechanism to switch the magnetic state of a free layer.

Domain Wall Motion (DWM) is a phenomena whereby two regions with opposing magnetic orientations in the same FM wire have a region between the two called the domain wall [29, 30]. By applying STT to the domain wall it will move in the direction of the electron flow, which allows charge current-based manipulation of a magnetic region in a continuous fashion as opposed to the discrete nature of mono-domain FMs used in typical STT-MTJs. DWM will be used in Chapter 3 to realize novel hybrid spin-CMOS digital circuits.

Spin Orbit Coupling (SOC) effects such as the Spin Hall Effect (SHE) are mechanisms by which particular materials cause electrons of one spin orientation to flow opposite to those of opposite spin orientation [31-38]. The SHE is a bulk phenomena found in heavy metals such as Pt or beta-W and works by passing a charge current through the material, and due to SOC electrons with opposing spins are pushed to opposite sides of the heavy metal, which provides a mechanism for injecting spin current into FMs interfaced with the material. SOC effects can be engineered to realize very high efficiency spin injection and as such has been at the forefront of the latest in spintronic device research. SOC will be used in Chapters 4, 5, and 6 to realize neuromorphic synapses and neurons.

Spintronic Approaches for Digital Circuits

Spintronic devices can be beneficial for standard digital circuits since they can reduce the overheads by reducing the number of devices needed while migrating some circuitry to BEoL, alleviating area overheads. The current-based switching of the devices also affords the straightforward implementation of majority logic gates, which can reduce total gate counts in some circuits [17]. Furthermore, the non-volatility of the devices allows for replacement of in-circuit memory structures, such as SRAM, with spintronic devices that can be readily power-gated when not being used to conserve static power consumption. Two novel hybrid spin-cmos circuits are developed in Chapter 3 that utilize the overhead-reduction and non-volatility capabilities of spintronic devices to improve both synchronous and asynchronous Von-Neumann architectures.

Spintronic Approaches for Neuromorphic Circuits

The dense, non-volatile, resistive, and stochastic switching properties of spintronic devices can all be utilized for neuromorphic architectures. Neuromorphic architectures are a paradigm shift away from Von-Neumann architectures that focuses on combining computation and memory elements within neural network computational schemes. The primary components of neuromorphic architectures are the synapses, which modulate the strength of input signals, effectively implementing a multiplication between the synaptic weight and the input signal, and the neurons, which implement some form of non-linearity applied to the summation of pre-synaptic input signals multiplied by the connecting synaptic weights. The synapses and neurons are typically organized into a crossbar configuration as shown in Figure 2.

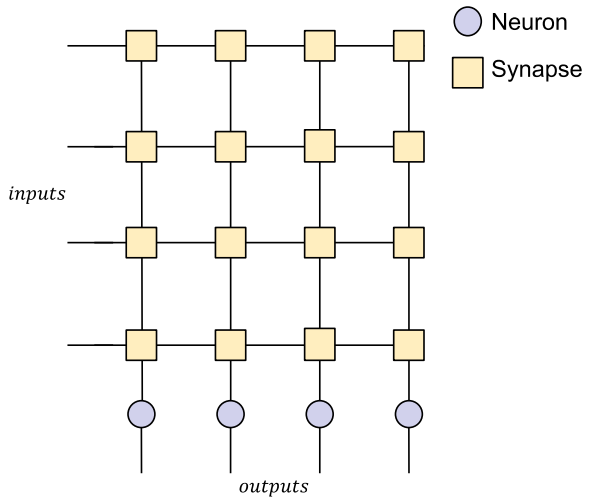


Figure 2: Typical neuromorphic crossbar architecture with synapses and neurons.

There are a variety of neural network algorithms that can be used to develop neuromorphic architectures such as standard Deep Neural Networks (DNNs) that use continuous valued weights and neuron activations [39], Binarized Neural Networks (BNNs) that use binary neuron activations and synaptic weights [40], and Spiking Neural Networks (SNNs) that attempt to mimic the spiking behavior found in biological neurons, use discrete binary spike events to represent the outputs of neurons and can have either discrete, continuous, or binary synaptic weights [3, 41-43].

Spintronic devices can be utilized to implement both the synapse and neuron components of neuromorphic architectures. The resistive nature of spintronic devices lends themselves to naturally implement synaptic weighting of input signals in an analog fashion in crossbar arrays for parallel computation of all input signals and synapses to be applied to the input of the neuron. Additional benefits are realized by the vertical integration of spintronic devices, reducing the area overheads of synapses, which make up the majority of neuromorphic architecture footprints. Also,

the non-volatility of spintronic devices mitigates static power consumption within the crossbar array. SNNs also receive additional benefits from the utilization of spintronic devices because they can leverage the intrinsic stochastic switching behavior of spintronic devices to realize efficient probabilistic update rules with binary or low-bit synapses for hardware-efficient neuromorphic architectures with unsupervised learning, as shown in Chapter 6. For neuron circuits, spintronic devices can naturally implement sigmoidal Probabilistic Activation Functions (PAFs), which can be implemented with resistive crossbar arrays for compact low-power neuron designs as developed in Chapter 4. Spintronic devices also pave a way towards more biologically-mimetic stochastic spiking neuron behaviors that realize powerful Bayesian computations as prescribed from several works in computational neuroscience [44-46]. Such stochastically spiking neurons typically require expensive overheads to be realized with deterministic CMOS circuits, but by leveraging the intrinsic stochasticity of spintronic devices to implement such stochastic spiking behaviors, compact low-power circuits can be realized, as developed in Chapters 5 and 6.

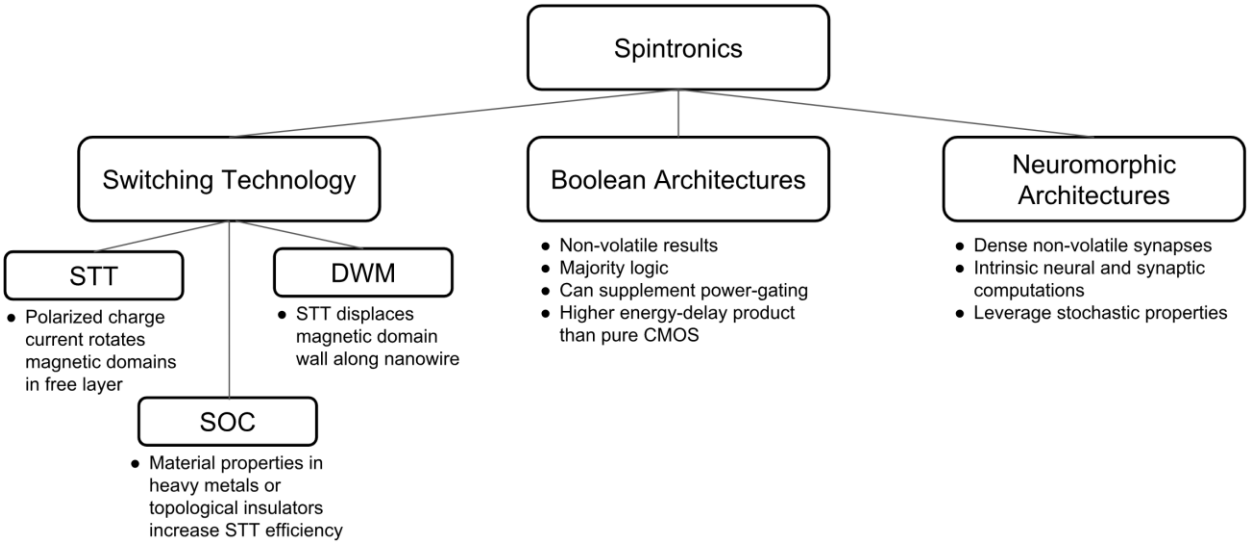


Figure 3: Taxonomy of spintronic technologies and architectures with their associated characteristics.

Contributions of the Dissertation

The primary contributions of this dissertation are detailed below.

Hybrid Spin-CMOS circuits for area reduction and enhanced functionality in Von-Neumann architectures: The direct integration of spintronic devices in traditional CMOS circuits achieves reductions in device counts as well as enhanced functionality due to the non-volatility imparted into the circuits, such as instant store and restore capabilities for aggressive power-gating schemes. As CMOS devices have been scaled, static power consumption has increased to a significant proportion of the total power consumption, prompting chip architects to utilize power-gating schemes whereby the power supplies to idle circuits are shut off to reduce static power consumption. However, due to the volatility of typical SRAM memory used in digital circuits, the

data needs to be migrated to either non-volatile storage or non-power-gated circuits prior to turning the supply off, and then the data needs to be restored from storage after power is resupplied before the circuit can resume operation. The work presented in Chapter 3 demonstrates how to utilize spintronic devices to impart edge-triggered flip-flops and Muller C-elements with instant store and restore capabilities while reducing device counts.

Stochastic neurons in area- and power-efficient circuits: Stochastic neuron circuits can be beneficial for both BNN and SNN architectures. The binarization of DNNs requires neurons that provide a non-linearity to a wide-ranging input, which can be either a deterministic sign function, or a sigmoidal Probabilistic Activation Function. Chapter 4 introduces a spintronic Probabilistic Activation Function that has reduced power and area overheads compared to the deterministic approach. A ubiquitous behavior found in biological neural networks is the stochasticity of the firing rate, which arises from the stochastic opening and closing of the ion channels that drive spiking behavior and is determined to be an important component of theoretical computational frameworks developed in computational neuroscience, such as Neural Sampling. Most spiking neuron circuits utilize a simple Leaky Integrate and Fire (LIF) model that does not accurately model the stochastic spiking found in-vivo and is incompatible with Neural Sampling. The works presented in Chapter 5 detail two approaches that utilize the stochastic behaviors of spintronic devices to realize compact, high-speed, and low-energy stochastic spiking neuron circuits with a spike rate determined by the input signal, which can be utilized for neuromorphic architectures implementing computational models that leverage stochastic spiking neurons.

Comprehensive hardware and algorithm co-design for process variation resilient and low-power neuromorphic architecture: The design of robust, hardware-efficient, and low-power neuromorphic architectures with unsupervised learning requires an integration of knowledge from device physics, low-power circuits, and computational neuroscience to realize a holistic device-aware circuit, architecture, and algorithm co-design. Several emerging devices are promising for advancing neuromorphic architectures in part due to their non-volatility and stochastic switching behaviors, of which spintronic devices have the advantage of nearly unlimited write endurance. Typical approaches to neuromorphic architectures using spintronic devices attempt to impose ideal algorithmic constraints, such as exponential Spike-Timing Dependent Plasticity learning rules onto non-ideal circuits in an ad-hoc fashion, requiring precise parameter tuning that would be challenging when considering the effects of process variations, especially at subthreshold voltages. The Neural Sampling Core developed in Chapter 6 leverages the intrinsic properties of spintronic devices and subthreshold CMOS under the effects of process variation to design ultra-low-power neuron, synapse, and homeostasis circuits that are hardware-efficient and implement principles from Neural Sampling, while utilizing the stochastic switching properties of spintronic devices to realize a new unsupervised learning rule called *Probabilistic Hebbian Plasticity* that is shown to be robust under various forms of parameter and circuit variations, input sizes, and input noise.

CHAPTER TWO: PREVIOUS WORKS

This Chapter starts by introducing previous approaches for implementing spintronic devices in standard digital circuits to reduce static power consumption through power-gating, due to the non-volatile nature of spintronic devices. Then, previous approaches utilizing emerging devices for compact and low-power neuromorphic circuits, such as neurons and synapses, are detailed.

Previous Spintronic Approaches for Digital Circuits

The non-volatile, high-speed, high write-endurance, and BEoL properties of spintronic devices offers the ability to improve digital circuits, particularly by ameliorating the ever-growing static power consumption overhead of highly scaled CMOS circuits through power-gating. In this Section, several approaches for utilizing spintronic devices to imbue D F/F and Muller C-Element circuits with low-overhead non-volatile storage for standby power reduction are delineated. A direct comparison between the works can be found in Table 1.

D F/F circuits, which are the primary intra- and inter-computation memory elements in standard synchronous architectures, are an ideal candidate for the non-volatile, high-speed, and high-endurance properties of spintronic devices since non-volatility within these circuits can allow entire portions of the overall architecture to be power-gated without loss of data. Several works have used these properties to develop such non-volatile D F/Fs, which are listed in Table 1. In particular, Ryu et al. [47] utilized a master-slave D F/F configuration in which the slave latch is

connected to a sensing circuit containing two MTJs, which also includes associated write circuitry. Prior to power-gating, a write-enable signal is sent to the write circuitry, which then writes the state of the D F/F into the MTJs in a complementary fashion. Once power is restored to the circuit, a sense-enable signal is sent to the sensing circuitry, which senses the data in the MTJs and loads that data into the D F/F. Suzuki et al. [48] used a master-slave D F/F configuration with a DWM device-based non-volatile storage cell with associated store and recall circuitry. Prior to power-gating a store signal is sent to the non-volatile storage cell, causing the state of the D F/F to be written to the DWM device, and once power has been restored, a recall signal causes the data in the DWM device to be sensed and loaded into the slave latch of the D F/F. Both of these non-volatile D F/F circuit designs require significant circuitry overhead as well as additional signaling wires and timing overheads to store and restore the data to and from non-volatile storage prior and after power-gating. Thus, the work presented in [8] and Chapter 3 removes these overheads, allowing very compact circuits with instant store and restore capabilities without any additional signaling wires or timing overheads.

The asynchronous counterpart to the D F/F for synchronous architectures is the Muller C-Element, which is involved with storing inter-computational data and is critical for performing asynchronous handshaking protocols in lieu of a global clock signal [49]. Therefore, several works have targeted the Muller C-Element for integration with spintronic devices to imbue asynchronous architectures with non-volatility for power-gating capabilities, as listed in Table 1. Zianbetov et al. [50] developed a hybrid spin-CMOS Muller C-Element with body biasing and a silicon-on-insulator design. Their design typically operates in an ordinary CMOS-only fashion for high speed,

with a propagation delay of just 32 ps, and it can then backup the state data to non-volatile MTJ cells prior to powering the circuit down for power-gating. The metrics listed in Table 1 compare only the backup delay and power instead of the standard CMOS-only high-speed operation in order to compare the non-volatile operation of the design with the intrinsically non-volatile operation of the design proposed herein. Storing data in spintronic devices can also increase radiation-induced soft error immunity. For instance, Onizawai et al. proposed their design to address Single Event Upsets using the resilience of MTJs [51]. Their design lacks the CMOS-only high speed operation of Zianbetov et al. as their focus was to improve reliability, and as such, needed to write to the non-volatile cells every operation.

Table 1: Comparison of Hybrid Spin-CMOS Digital Circuits.

Ref.	Circuit	Device Count	Delay	Power
[47]	D F/F	56T + 2S	203.3 ps	N/A
[48]	D F/F	27T + 1S	62.2 ps	12.7 uW
[8] *Herein*	*D F/F*	*10T + 1S*	*~1-10 ns*	*~3-9 uW*
[51]	Muller C-Element	38T + 2S	1.05 ns	263.8 uW
[50]	Muller C-Element	17T + 2S	1 ns	50 uW
[2] *Herein*	*Muller C-Element*	*8T + 1S*	*801 ps*	*34.04 uW*

Previous Approaches for Neuromorphic Circuits

Spintronic devices have been utilized to realize high-speed, low-power, and compact neuromorphic circuits including neurons and synapses, of which we review a recent selection in this Section. Although spintronics have been utilized to realize both deterministic and stochastic neuromorphic circuits, since the works developed in this dissertation focus on the stochastic approach for its potential power and area savings as well as biological plausibility and compatibility with theories from computational neuroscience. Where appropriate, spintronic approaches are also compared to standard CMOS and alternative emerging technologies, such as memristor in order to demonstrate the benefits of spintronics for stochastic circuit implementations.

The implementation of stochastic neuron circuits requires a circuit to take an input signal, which is typically voltage or current, and compute a probability of spiking based on that signal. Most approaches utilize a sigmoidal spiking probability, which is commensurate with the behavior of emerging device switching probabilities as well as biological behavior and computational neuroscience theories [42, 44, 45, 52]. Implementing stochastic neuron circuits in standard digital CMOS designs is not very natural due to the determinism inherent in the designs, and requires a Pseudo-Random Number Generator (PRNG) and the associated power, area, and timing overheads, such as in [53]. Emerging devices can implement a sigmoidal probability of switching much more naturally through their stochastic switching properties. Wijesinghe et al. [54] utilized the stochastic filament formation behavior of amorphous silicon-based metal filament formation memristor devices to realize a compact stochastic spiking neuron that operates with a three-phase

approach. The first phase is the write phase where a voltage is applied to the memristor, representing the input signal corresponding to the previous synaptic weighted summation computation, to the device for a period of time, which may or may not switch based on the input voltage. After the write phase, the read phase senses the state of the memristor, and if the device switched during the write phase, a spike signal is generated. After the read phase, the device is reset with a voltage pulse with an amplitude and duration chosen such that the device is reset with a very high probability. Spintronic approaches to designing stochastically spiking neural circuits utilize the thermally-driven random magnetic excitations found in nanomagnets to switch the device probabilistically based on the applied input signal [42, 43, 52]. Previous approaches using both MTJs and SHE-MTJs have a similar three-phase approach to the memristor design previously described. The first phase is considered the write phase, where a current pulse is applied to the MTJ or the SHE layer of the SHE-MTJ with a current pulse equivalent to the input strength for a pre-defined pulse duration, and the device may or may not have switched. After the write phase, the device is sensed, and if it was switched, an output spike is generated. Once the read phase is complete, the MTJ or SHE-MTJ must be reset with a strong current pulse to reset the state of the device prior to the next write phase. Of all the previous memristor and spintronic stochastic spiking neuron approaches, they require the write-read-reset phases, which introduces additional time and power overheads. In Chapters 5 and 6, spintronic stochastic spiking neuron circuits are developed that leverage the properties of a novel spintronic device to realize intrinsically spiking hardware that does not need any additional read or reset phase overheads.

In addition to stochastic spiking neuron circuits, emerging devices have been utilized for implementing stochastically switching synapses for realizing either binary or multi-bit synapses with probabilistic update rules for compact and low-power unsupervised learning algorithms in hardware. Suri et al. [55] fabricated a SNN chip with conductive-bridge RAM (CBRAM) binary synapses with probabilistic unsupervised learning rules that were capable of realizing auditory and visual pattern extraction. Bill et al. [56] showed that compound memristive synapses using 1-100 devices with unsupervised probabilistic learning rules is capable of recognizing handwritten digits with diminishing returns on the number of devices in each synapse, showing that high precision is not required. Zhang et al. demonstrated a similar scheme with compound MTJ synapses consisting of 1 to 9 MTJs and showed that with probabilistic unsupervised learning rules, the network can learn to recognize handwritten digits with respectable accuracies and some robustness to process variation. Srinivasan et al. [42, 57] demonstrated two approaches to using SHE-MTJs as synapses with probabilistic unsupervised learning rules resembling exponential Spike-Timing-Dependent-Plasticity (STDP) curves; one showed that single SHE-MTJs are capable of achieving respectable accuracies for handwritten digits and a stochastic spiking neuron, and the other showed that two SHE-MTJs per synapse, each with slightly different circuity, could realize long-term and short-term memory, which reduces the total energy required to train the network. In Chapter 6, a low-precision subthreshold hybrid spin-CMOS synapse using three SHE-MTJs and a new probabilistic unsupervised learning rule called *Probabilistic Hebbian Plasticity* is developed to realize an ultra-low power synapse that is robust to process variations.

Table 2: Comparison of Stochastic Spiking Neuron Approaches

	[53]	[54]	[42, 43, 52]	*S3N Herein*	*S4N Herein*
Technology	CMOS	Hybrid CMOS/Memristor	Hybrid Spin-CMOS	*Hybrid Spin-CMOS*	*Hybrid Spin-CMOS*
Source of Stochasticity	PRNG	Memristor Switching Probability	Thermal Energy	*Thermal Energy*	*Thermal Energy*
Spike Implementation	Event Signal	Write-Read-Reset Cycle	Write-Read-Reset Cycle,	*Intrinsic Circuit Behavior*	*Intrinsic Circuit Behavior*
Spike Time-Scale Order	1 ms	10 ns	1 ns	*10 ps*	*1-10 ns*
Energy per spike	~10 pJ	~1-10 pJ	~1 nJ	*~1 nJ*	*~10-100 aJ*
Normalized Device Count	>10x	~1x	~0.5×	*1×*	*2x*

Probabilistic Bits

An important family of spintronic approaches to leveraging thermally-driven true random behavior in circuits, which are used extensively in Chapters 4, 5, and 6, are called Probabilistic Bits (p-bit) [58]. P-bit circuits come in a couple different flavors that center around a spintronic device with a very low energy barrier free layer as will be described subsequently. Although these specific devices have not been demonstrated comprehensively yet, the field of probabilistic spintronics is an active area of research with very promising experimental results demonstrating the utilization of stochastic nanomagnets as a tunable random number generator [59].

Spin-Hall Effect Probabilistic Bit

The SHE-driven p-bit, shown in Figure 4a, combines a heavy metal (HM) exhibiting SHE and an MTJ whose free layer magnetization is modulated by the SHE layer. Unlike standard experiments combining the HM with an MTJ that utilizes ferromagnets with energy barriers of the order of 40 to 60 kT [58, 60-62], the p-bit uses an unstable ferromagnet, with an energy barrier of 0 to 1 kT, which can be obtained by either reducing the volume of a stable magnet [63] or by using circular magnets that effectively have no barrier in the absence of a geometrically preferred easy axis [64]. In the absence of any SHE current, the magnetization fluctuates with average $< m_z >= 0$ and the inverter chain amplifies this signal to produce rail-to-rail voltage swings between 0 and VDD. An input current into the SHE layer generates a spin current that influences the magnetization of the circular magnet, which effectively biases the probability that the output is 0 or VDD. The SHE-driven p-bit is utilized in Chapters 4 and 5 to provide intrinsic thermally-generated stochasticity for various stochastic neuron implementations.

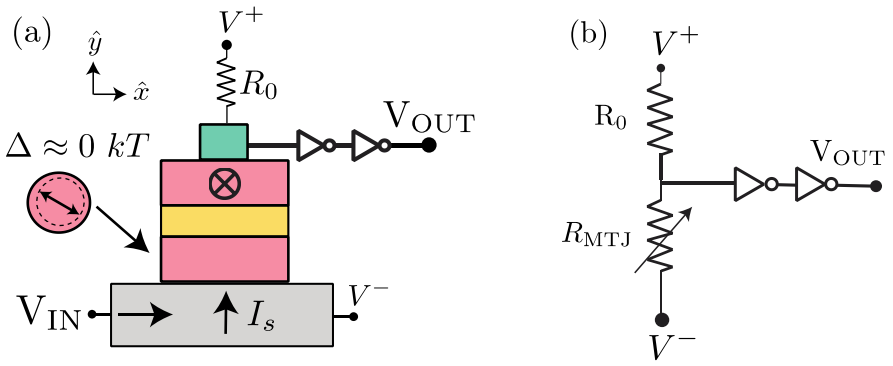


Figure 4: p-bit device schematic and equivalent READ circuit: (a) The gray layer represents a heavy metal (HM) exhibiting the Spin Hall Effect (SHE) that injects a spin current into an adjacent "free layer" of a Magnetic Tunnel Junction. The free layer is a circular magnet with no preferred easy axis (EB=0 kT) that fluctuates in the z-x plane in the presence of thermal noise. The MTJ is connected to an average resistance R0 creating a fluctuating voltage that is amplified by two inverters. (b) The circuit equivalent READ circuit is also shown.

Embedded Probabilistic Bit

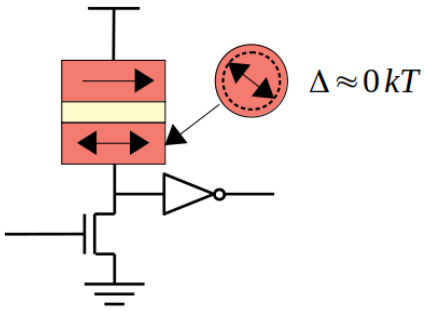


Figure 5: Embedded p-bit circuit.

The Embedded p-bit circuit developed in [65] is a straightforward adaptation of the standard 1-transistor with 1-MTJ circuit typically used in spintronic memory arrays but with a low-energy barrier free layer, as shown in Figure 5. Due to the very low energy barrier (Δ) of the free

layer, the MTJ of the embedded p-bit will stochastically switch between its AP and P states, where the mean retention time for an MTJ (τ) is given by:

$$\tau = \tau_0 e^{\Delta/kT}$$

Where τ_0 is a material dependent parameter called the attempt time, k is Boltzmann's constant, and T is the temperature in Kelvin [65]. Since the input to the inverter is a voltage divider between the MTJ and the NMOS transistor, the output of the inverter will be a function of the voltage applied to the gate of the NMOS and the probabilistic state of the MTJ, providing a sigmoidal probability of outputting a logic 1 based on the voltage applied to the NMOS. The embedded p-bit is utilized in Chapters 5 and 6 to realize compact and low-power voltage-controlled stochastic spiking neuron circuits.

CHAPTER THREE: HYBRID SPIN-CMOS DIGITAL CIRCUITS

In this Chapter, the *Domain Wall Coupled Spin Transfer Torque (DWCSTT)* device developed in [30] is described and then utilized to design two new hybrid digital circuits that introduce a novel direct interfacing between spintronic and CMOS devices to realize critical datapath components with enhanced functionality and reduced area costs. The first circuit developed is an edge triggered flip flop utilizing spintronic devices to enable instant store and restore functionality for aggressive power-gating schemes © 2016 IET, reprinted, with permission, from [8]. The next circuit leverages the state-holding properties of a spintronic device to reduce the area overhead of the Muller C-element, which is a critical component in asynchronous architectures, while also imbuing it with instant store and restore functionality, © 2018 IEEE, reprinted, with permission, from [2]. Both circuits illustrate how the use of complementary MTJs in a single device allows a seemless voltage-divider-based integration with CMOS.

Domain Wall Coupled Spin Transfer Torque Device

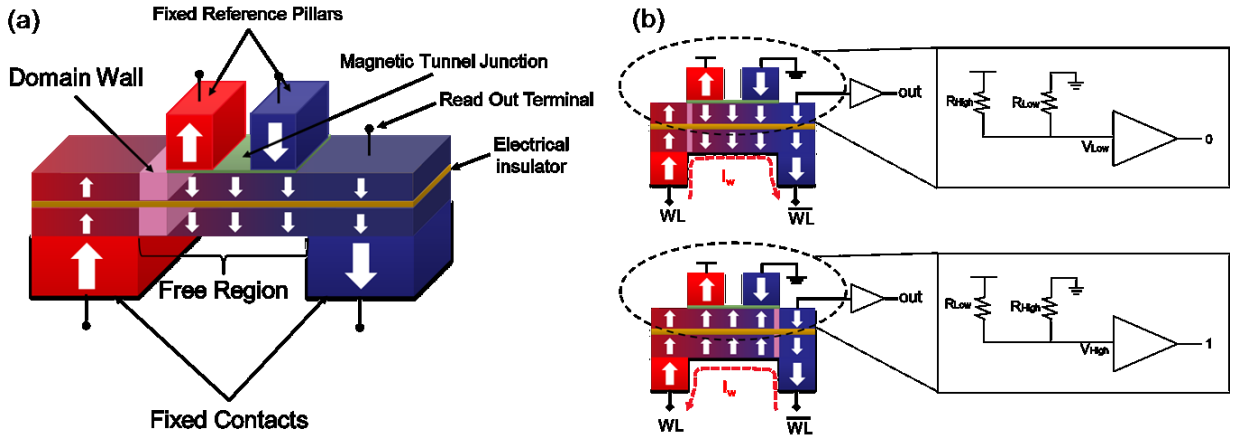


Figure 6 (a): Domain Wall Coupled Spin Transfer Torque Device; (b): Low and High states.

The DWCSTT device shown in Figure 6a uses two electrically isolated, but magnetically coupled FM domain wall layers to isolate the read and write mechanisms of the device [30]. The device state is sensed through the two anti-parallel fixed reference pillars, which have MTJs with the underlying domain-wall-based free layer. The domain wall only has two stable states as shown in Figure 6b, and therefore one fixed reference pillar will always be R_{High} and the other will be R_{Low}, exclusively, and they will alternate depending upon the location of the domain wall. If the TMR of the MTJs is large enough (~100%), then we can use the two MTJs of the DWCSTT device as a voltage divider to output a V_{low} or a V_{high}, which can then be used to switch a CMOS inverter as shown in Figure 7b [8]. The write operation of the device is performed by passing a current through the lower domain wall FM, denoted as the write layer, which is first spin-polarized through the fixed contact layers, and then exerts a STT on the write layer, which moves the domain wall.

Since the upper domain wall FM layer, denoted as the free layer, has strong dipolar coupling with the write layer, its magnetization will rotate in conjunction with the write layer as it undergoes STT. The velocity of the domain wall is linearly related to the current density applied to the write layer, and experimental results show that domain wall velocities up to 125 m/s is achievable with current densities near 1.8e8 A/cm^2 [66]. Using 16nm PTR CMOS models [67] with a supply voltage of 0.7V, respectable write speeds are achieved by simply applying logic "1" or "0" (*vdd* or *gnd*) to the device inputs. By varying the driving transistor widths, the speed and power draw of the device are able to be adjusted. Seo et al. utilized the self-referencing differential nature of the device to read the state of the device by fixing the read out terminal to ground and then comparing the currents of the two fixed reference pillars when a fixed voltage is applied to both [30]. However, with proper read and write path optimization of the DWCSTT, 16nm CMOS gates with balanced transistor widths are capable of both writing to and reading from the device in lieu of using a sense amplifier to compare relative current levels.

Compact Instant Store/Restore D Flip-Flop

The Hybrid Spin-CMOS edge triggered Flip-Flop (D F/F) focuses on using the non-volatile properties of the DWCSTT to reduce the number of transistors needed compared to the standard pure-CMOS implementation of a master-slave D F/F while providing instant store/restore functionality with full data retention, simplifying the requirements of power-gating techniques [8]. The D F/F developed herein is shown in Figure 7 and consists of a Static Random Access Memory (SRAM)-based master latch, a DWCSTT device as the slave latch, an output inverter, and two

pass gates used for clocking control. While the clock signal is low, the pass gate leading into the master latch from the input is conducting, allowing the master latch to poll the data arriving at the input terminal D. Once the clock signal goes high, the master latch becomes isolated from terminal D and the data stored in the master latch is latched into the DWCSTT slave latch as shown in Figure 7. With this circuit, power gating is achieved by simply disconnecting the entire circuit from VDD since the data is already latched inside the NV element, and no pre-sleep data-storing strategies are necessary. However, since the data stored inside the SRAM master-latch is non-deterministic upon re-powering the circuit, power restoration must commence with the negative edge of the CLK signal to ensure that the proper data stored in the NV slave-latch is propagated through the circuit and the result is ready at the master-latch of the following D F/F before CLK may go high and write the data from the master-latch into the slave-latch. The delay between the clock signal going high and the output updating based on the new data is called the clock-to-Q (C-Q) delay [47], and is dependent upon the speed of the STT-driven domain wall motion in the DWCSTT device, which is proportional to the write current provided by the transistors in the SRAM master-latch. Thus, the C-Q delay can be adjusted by varying the width of the transistors in the SRAM master-latch cell. By increasing the transistor width, we can reduce the C-Q delay for an increased power and area overhead. The relationship between transistor width, power, and C-Q delay is shown in Figure 8, where the x-axis is multiples of the minimum feature size (F) corresponding to $W_{NMOS} = xF$ and $W_{PMOS} = 2xF$. For these simulations, F is taken to be 16nm.

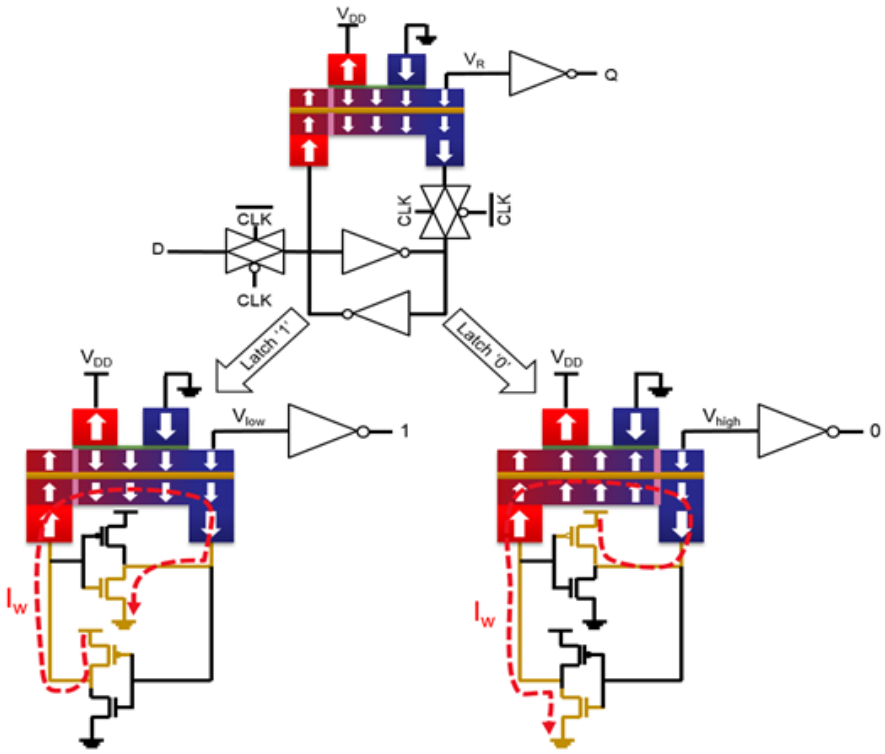


Figure 7: The compact hybrid spin-CMOS D F/F circuit with non-volatile input latching.

Results

We simulated the Hybrid Spin-CMOS D F/F circuit in HSPICE using the 16nm high-performance transistor models available from Arizona State University [67]. The DWCSTT device was simulated by taking a Verilog-A model from the mCell model available online [68] and modifying it such that it accurately performs the operation of the DWCSTT device. Essentially, the only change from the mCell to the DWCSTT model was to flip one of the reference layers such that the two fixed reference pillars are complementary and to connect a terminal to the free layer shared between the two MTJs. The circuit parameters for the simulation are found in Table 3. The value of R_P is chosen to be of a high resistance but not out of the range of feasibility [69] in order

to reduce the read power overhead. W_{NMOS} and W_{PMOS} are chosen to minimize the C-Q delay, and if one's application can relax the C-Q delay for improved area and power metrics, they may reduce the transistor sizing. The device width, length, TMR, and write path resistivity were all chosen as the base values included in the model of the mCell [68].

The simulated waveforms for the developed D F/F are shown in Figure 9. At the positive CLK edge, the data present at D is written into the DWCSTT slave-latch and is then outputted at terminal Q as depicted. The functionality of this design is critically dependent upon V_R switching above and below the threshold voltage for an inverter, V_{Th}, which is shown. Upon power-gating V_{DD}, the data stored in the slave-latch is saved and immediately restored upon restoration of V_{DD} to the circuit, illustrating the instant store/restore functionality of the D F/F.

Table 3: Simulation parameters used for the compact hybrid spin-CMOS D F/F

Parameter	Value
V_{DD}	$0.7V$
R_P	$40k\Omega$
TMR	100%
Device Length	12nm
Device Width	10nm
W_{NMOS}	96nm
W_{PMOS}	192nm
Write Path Resistivity	$200\Omega nm$

Compared to previous works [47, 48], the proposed design has a reduced number of transistors and negates the need for store/restore circuitry and signaling prior to power-gating. However, since the C-Q delay of this design is impacted by the relatively slow write speed of the

DWCSTT compared to an SRAM cell, some trade-offs are observed. In particular, compared to [48] the proposed D F/F uses 17 fewer transistors, but has a 1.2ns longer C-Q delay. Although such an increase in C-Q delay is unfavorable for many applications, applications with relaxed speed requirements that utilize power-gating schemes can be benefited with the compact size and simplified power-gating requirements of the proposed design. In addition, further advancements in spintronic research may lead to faster switching designs, which can improve the C-Q delay.

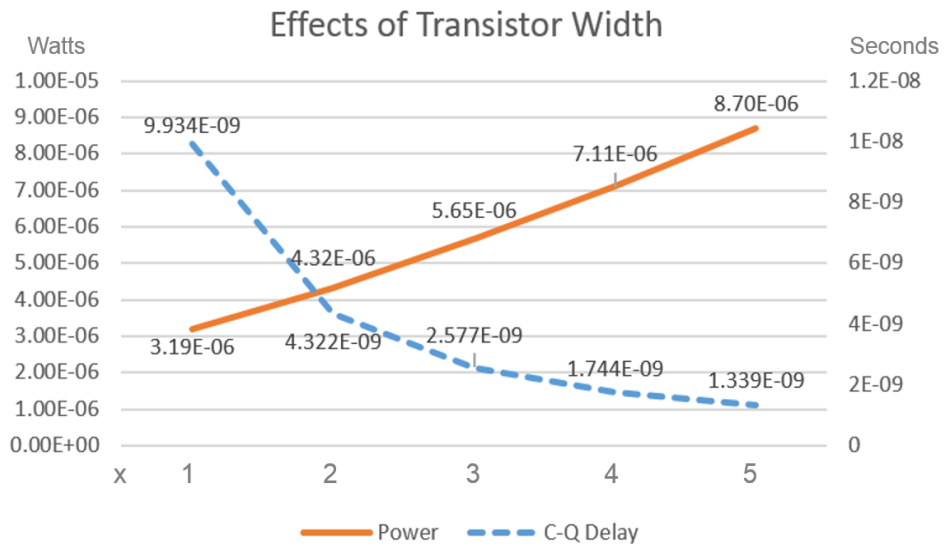


Figure 8: The effects of transistor width on power and C-Q delay of the compact hybrid Spin-CMOS D F/F

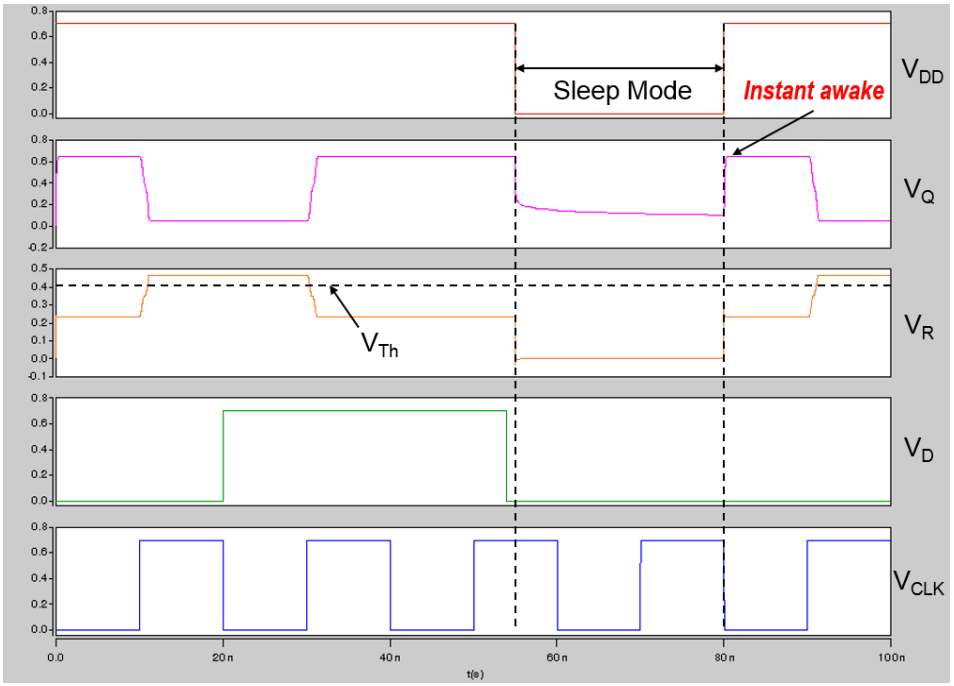


Figure 9: Simulated waveforms of the compact hybrid spin-CMOS D F/F

Discussion

The proposed D F/F design herein is shown to retain its data in a non-volatile spintronic device as a part of its operation, which allows instant store/restore functionality without the need for store/restore signaling or overhead control circuitry. Furthermore, the proposed design uses 10 fewer transistors than a traditional pure-CMOS-based master-slave D F/F [48]. Additionally, we showed that by varying the transistor widths in the SRAM master-latch, it is possible to tune the circuit for the power, delay, and area needs of one's application. The functionality of the design was demonstrated by using 16nm CMOS models and a Verilog-A model of the DWCSTT device in HSPICE. Area results were favorable compared to previous works, but C-Q delay was shown to be worse.

Spintronic Muller C-Element

The complementary roles of asynchronous architecture with nonvolatile spintronic devices are explored in this section to realize novel improvements for a logic element critical for asynchronous architectures. By redesigning the Muller C-Element to take advantage of spintronic device non-volatility and area-efficiency, benefits such as reduced asynchronous handshaking area overhead, are achieved in addition to instant on/off capabilities for reduced static-power dissipation through power-gating. A novel 8 transistor and 1 spintronic device Muller C-Element design is developed which is 20% faster and uses 68% of the power of previous non-volatile Muller C-Element designs. This spintronic Muller C-Element is demonstrated within a 4-phase dual-rail asynchronous First In First Out (FIFO) pipeline resulting in 48% fewer transistors in comparison with the previous designs. Additionally, bundled-data protocol overheads are shown to be reduced by using the spintronic Muller C-Element proposed herein. Detailed analysis of the effects of driving transistor width and the TMR ratio on device performance characteristics are included.

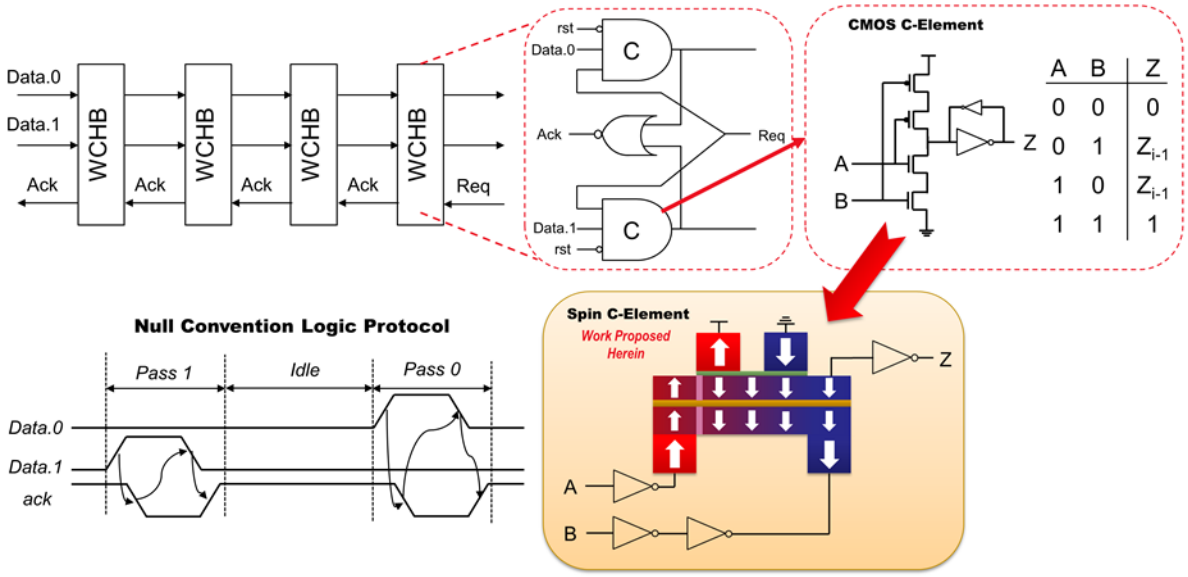


Figure 10: Null Convention Logic, Weak Conditioned Half Buffer, Muller C-Element gate, CMOS implementation, and the proposed implementation. © 2018 IEEE.

Overview of Asynchronous Pipelines

Asynchronous pipelines utilize local clock generation control circuits instead of the global synchronization clock found in synchronous pipelines [70]. Local clock generation in lieu of global synchronization provides robust, elastic, and high performance circuits with on-demand power consumption by design [70]. Systems with similar levels of robustness, performance, and power efficiency are difficult, if even possible, to realize using synchronous design [70]. This is due to synchronous designs being limited by the worst case logic delays, register setup and hold times, process variation, and aging-based circuit degradation [70]. Additionally, propagation delays, multiple clock domains, and the exponentially increasing circuit complexity of today's

microprocessors all contribute to increasingly complex clock trees and the associated area and energy overheads relative to the circuits they are controlling in synchronous pipelines [71]. Heat, power, security, and electromagnetic radiation issues also arise from the large power spikes following the clock edge, which is avoided with asynchronous design [72].

Asynchronous design is an established field with a large variety of protocols available, and it is out of the scope of this dissertation to describe them in detail completely. Thus, we will introduce the two protocols used herein. The 4-phase dual-rail quasi-delay-insensitive Null Convention Logic (NCL) pipeline as depicted in Figure 10 is based on dual-rail logic, which uses 2 wires, noted as *Data.0* and *Data.1*, to transfer every bit of data [70]. If *Data.0* is asserted logic high, then a logic 0 is transferred. If *Data.1* is asserted logic high, then a 1 is transferred. If both *Data.0* and *Data.1* are asserted logic low, then a NULL value is transferred, and it is not allowed for both *Data.0* and *Data.1* to be asserted logic high at the same time. The basic register element for our particular NCL implementation is called the Weak-Conditioned Half Buffer (WCHB) and is shown in Figure 10. The WCHB works by correlating the request (*req*) signal of the following stage with the input dual-rail data of the previous stage to determine if it can record the previous stage's data. When a data signal is asserted high, the WCHB resets the *ack* signal, indicating that it is not ready to accept data, other than a NULL value. Once a NULL value is received, the *ack* signal is set and indicates that it is ready to accept new data. With this inter-stage handshaking, data flows through the pipeline in a coordinated accurate-by-design method without the need for global synchronization.

The asynchronous Bundled-Data (BD) protocol lacks the dual-rail logic of NCL architectures, which alleviates the overhead needed for two wires per bit, but negates the intrinsic completion detection of NCL designs [70]. By contrast, BD is implemented with standard pipelining of combinational logic and local clock generation by inserting delay elements equivalent to the delay of the combinational logic between clock generating circuits.

Regardless of the protocol, the key element for implementing the inter-stage handshaking that many asynchronous pipelines depend on is the *Muller C-Element* [49]. As depicted in Figure 10, the Muller C-Element asserts a logic 1 when both inputs are logic 1 and asserts a logic 0 when both inputs are logic 0; if the inputs are different, then its output does not change. The key principle of this operation is that an output change indicates that both of the inputs are identical to the output at that transition. One particular low-area CMOS implementation of the Muller C-Element is shown in Figure 10 and utilizes a weak-inverter based SRAM cell to store the output data until a (0,0) or (1,1) condition is reached [73]. The volatile state of this design leads to increased static power dissipation and power-gating overheads by requiring additional store and restore circuitry and delays if power-gating is desired. The work proposed in this Section utilizes the non-volatile memory properties of a particular spintronic device for implementing a compact Muller-C element with instant store/restore functionality for reduced asynchronous pipeline area and power-gating requirements.

The following research contributions are provided for this design:

1) a novel, compact spintronic-based Muller C-element design for reducing asynchronous control area overhead,

2) reducing power-gating store and restore delay and circuit overheads by operating in an intrinsically non-volatile manner, and

3) realization of a delay-element-free asynchronous Bundled Data pipeline.

Proposed Spintronic Muller C-Element Designs

When developing the spintronic Muller C-Element, two functionally-correct designs were obtained. The first design developed is shown in Figure 11a and operates by only allowing one p-MOS branch and one n-MOS branch to be "on" when *A* and *B* are either (0,0) or (1,1), and all CMOS branches to be in a high-impedance state when *A* and *B* are (1,0) and (0,1). This causes current to pass through the write terminals of the DWCSTT device only when the output transitions according the Muller C-element functionality. It also restricts current flow through the write terminals when output transitions do not occur.

The second design iteration of the spintronic Muller C-Element is shown in Figure 11b. This design operates by using the pMOS or nMOS branches of the inverters driving *A* and *B* to connect the write terminals of the DWCSTT device to either V_{DD} or *GND*. If A and B are (0,0) or (1,1), then a potential difference occurs between the write terminals of the DWCSTT device, which generates a current through the device to change its state. If *A* and *B* are either (0,1) or (1,0), then both write terminals will be either *GND* or V_{DD}, thus eliminating any potential difference necessary to generate current flow.

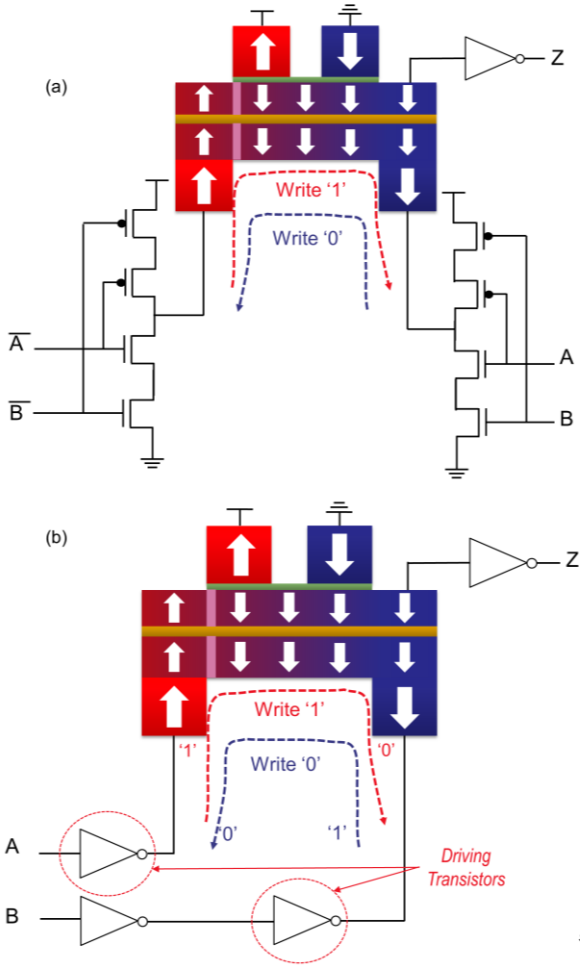


Figure 11: Two Spintronic Muller C-Element Designs. © 2018 IEEE.

Spintronic Muller C-Element Results

The two design iterations of the spintronic Muller C-Element were implemented with 16nm PTR transistor models [67] in HSPICE. The DWCSTT device was simulated by altering the Verilog-A model of [68] to have opposite fixed reference pillars as well as a *Read Out Terminal* according to the DWCSTT device operation; exactly the same as for the spintronic D F/F developed previously in this Chapter.

The circuits simulated with their corresponding functional verification waveforms and power and delay metrics are shown in Figure 12. For both circuits, V_{DD} is 0.7V, nMOS width is 3F, pMOS width is 6F (where F is the minimum feature size), the TMR is 100%, and the value of R_{Low} used herein is $40k\Omega$. Having a large R_{Low}, which is correlated with the thickness of the oxide layer in the MTJ, minimizes the reading current used in the voltage-divider sensing scheme. Both circuits were also power-gated by turning off V_{DD} for 5 ns when the output is both a logical 0 and logical 1. In both instances, when V_{DD} is restored, the correct logical value is restored, demonstrating the simple instant-on/off nature of the circuits. Additionally, V_R, which is the voltage read at the *Read Out Terminal,* is shown to switch above and below V_{Th}, which is the threshold voltage for a 16nm CMOS inverter using PTR models.

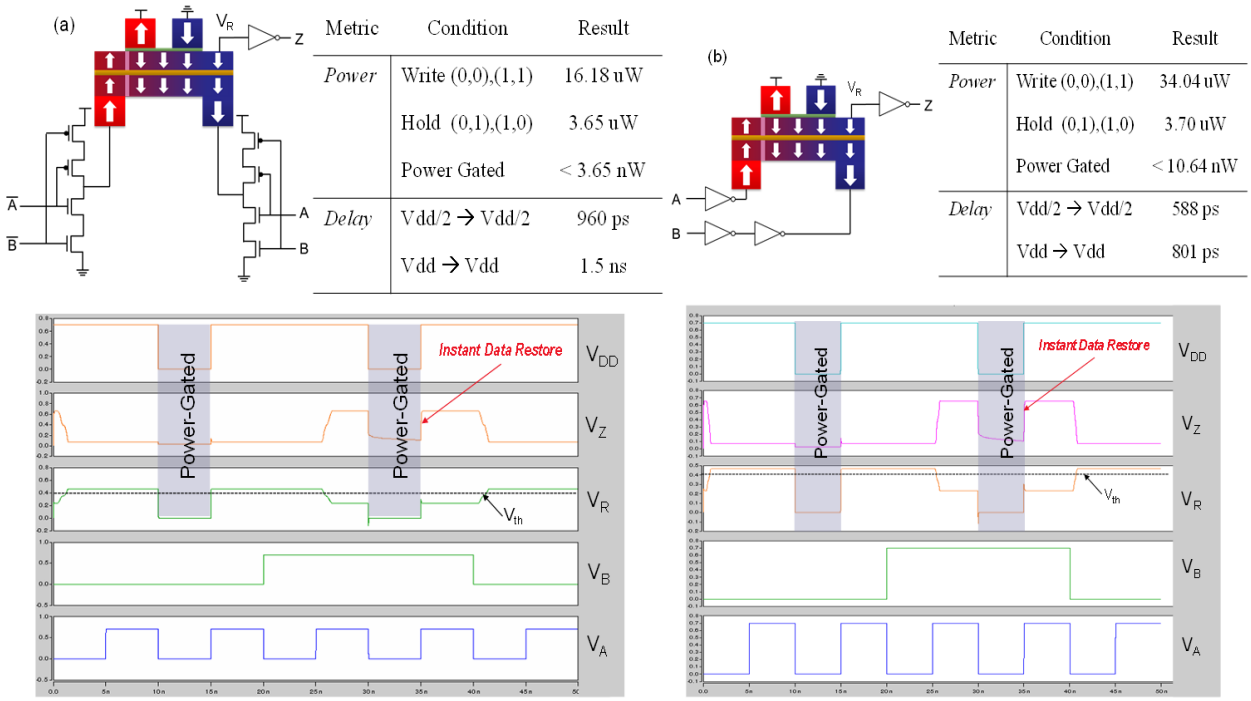


Figure 12: Spintronic C-Element designs, functional verification, and performance metrics.

Although the first design iteration in Figure 12a showed much lower power consumption than the second design in Figure 12b, the nearly double rail-to-rail swing delay causes the total energy consumption to be similar between the two designs. Since both designs use a comparable amount of energy, yet, the second design is nearly twice as fast and reduces the transistor count, it is deemed the better design and is the one used for pipeline simulations in the following sub-section.

In Figure 13, we analyze the relationship between TMR and write power, energy, and delay in order to extrapolate how the spintronic Muller C-Element will benefit with improved MTJ manufacturing techniques, which can improve TMR. For instance, room temperature TMR has

been experimentally reported to be as high as 604% [74]. Additionally, in Figure 13, we analyze how the widths of driving transistors, which determine the magnitude of the current through the DWCSTT, affects the performance characteristics of the Muller C-Element. As the TMR is increased, improvements in all performance characteristics are observed due to the greater difference between V_{low} and V_{high} achieved when increasing the TMR. As the driving transistor width is increased, which increases the driving current, write power is increased significantly, delay is decreased significantly until w=3, and then steadily decreases, and write energy is increased.

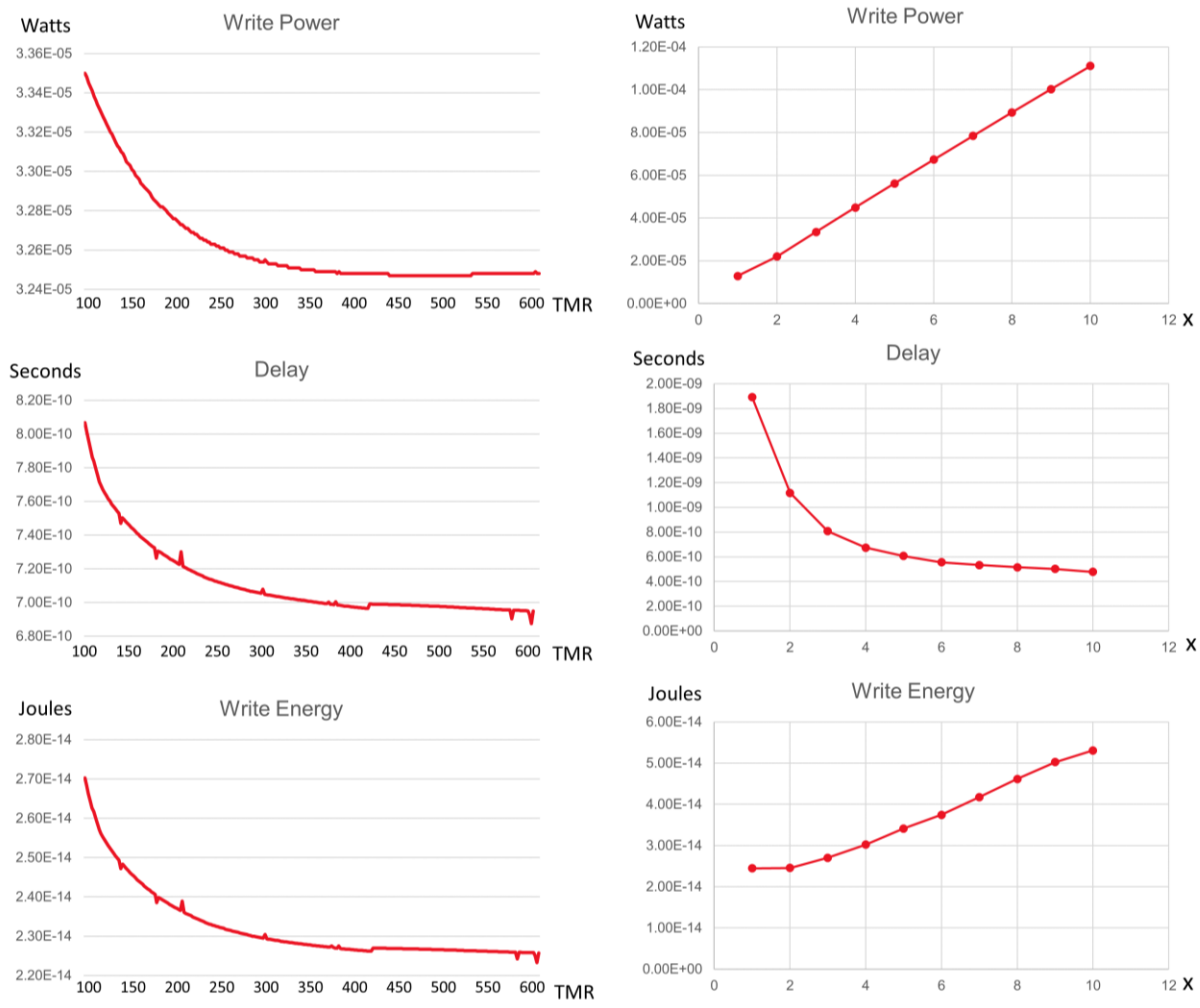
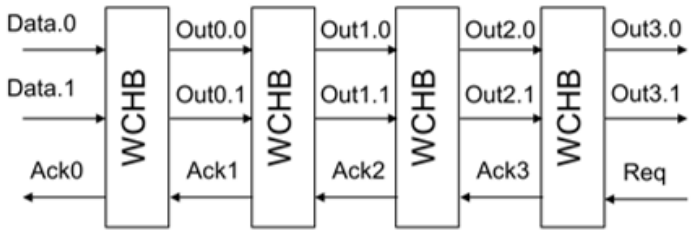


Figure 13: Left side – relation of performance characteristics to TMR. Right side – relation of performance characteristics to driving transistor width (nMOS width = *Fw* and pMOS with = *2Fw* where *F* is the minimum feature size). © 2018 IEEE.

Asynchronous Pipeline Simulation and Results

In this Section, two different pipelines utilizing alternative asynchronous protocols are implemented and discussed. First, a simple 4-bit dual-rail NCL FIFO pipeline is implemented with the spintronic Muller C-Element demonstrating functional correctness as well as instant on/off power-gating potential. Next, a pipelined 4-bit Ripple Carry Adder is implemented in a Bundled Data fashion to illustrate how the delay of the Spintronic Muller C-Element compared to typical CMOS gate delays can be utilized to reduce circuitry overhead of the local clock generating circuits.

In order to demonstrate the functionality and results of the developed Muller C-Element, the asynchronous NCL pipeline shown in Figure 14 is simulated in HSPICE. It consists of a simple FIFO pipeline using WCHBs as the intermediate storage and control. Each WCHB uses two Muller C-Elements with reset control and a two-input NOR gate as shown in Figure 10. The total transistor count for each WCHB is 24, which is 52% fewer transistors than in [70]. The included waveform demonstrates the pipeline handshaking protocol, which deterministically ensures correct propagation of data, typically called tokens, between the stages. Since V_{Req} is held low for a period, tokens are only allowed to propagate to the 3rd stage until V_{Req} goes high. Until this happens each stage holds onto its previous token, even after power-gating the entire circuit for 5 ns, demonstrating the NV functionality of the pipeline.



Metric	Condition	Result
Power	Average	120.83 uW
	Peak	202.95 uW
	Power Gated	39.38 nW
Delay	Token Passing	641 ps

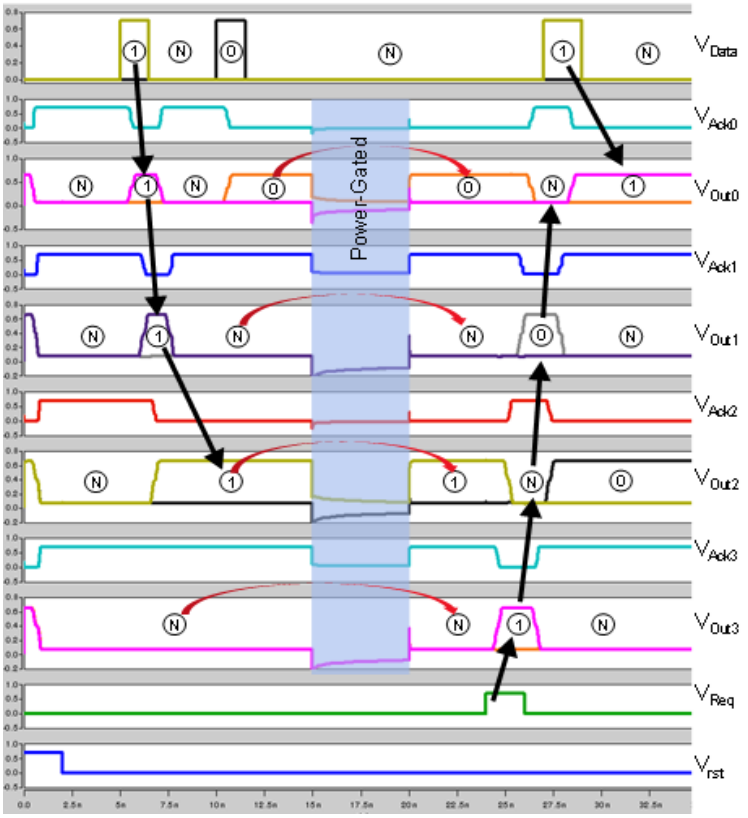


Figure 14: Asynchronous 4-phase dual-rail FIFO pipeline design, performance characteristics, and functional verification showing instant on/off after power gating. © 2018 IEEE.

Typical CMOS-only implementations of Bundled Data pipelines require delay elements equating to the delay of the combinational logic to be inserted between the local clock generating Muller C-Elements as shown in Figure 15. By simulating the clocking circuitry using Spintronic Muller C-Elements without any combinational logic, it is determined that the propagation delay between pipeline stages is about 605ps. Additionally, the propagation delay of a two-input NAND gate was found to be about 9ps. By extrapolation, it can be said that up to 67 CMOS gate delays can fit within the delay margin of the Spintronic Muller C-Element, and therefore, it is not obligatory to insert delay elements between Muller C-Elements, as illustrated in Figure 15. This concept was then simulated with a four-stage pipelined 4-bit ripple carry adder, and it was found to be functionally correct and had additional delay slack to accommodate larger combinational circuits between pipeline stages. The stage delay for an input pattern of all zeros for both operands was 605 ps, which corresponds to the minimum delay since no carry logic is required. The maximum stage delay for an input pattern of all ones for both operands was about 2.4 ns, which corresponds to the maximum delay since each adder will compute a carry that must be propagated.

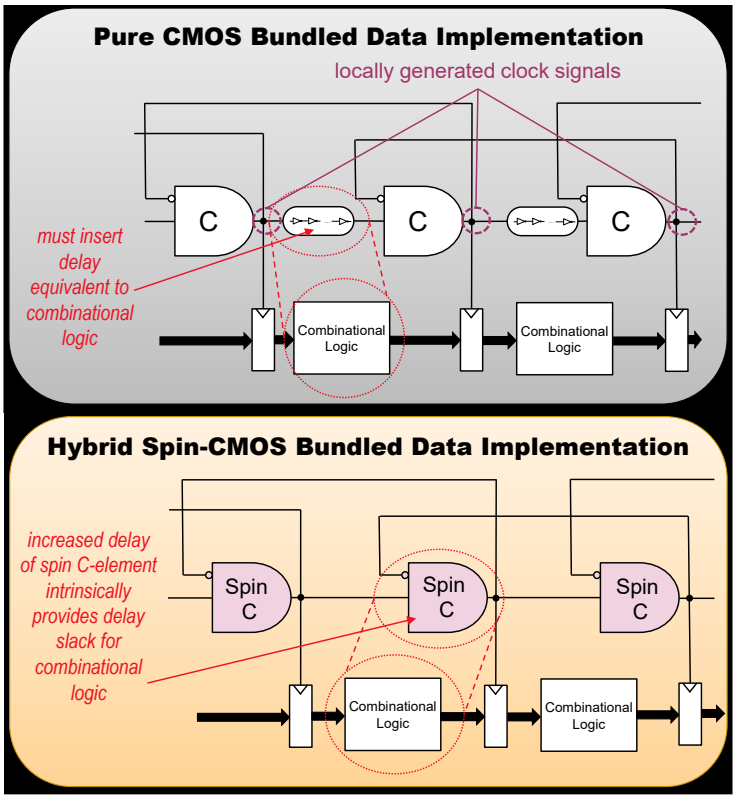


Figure 15: Comparison between pure CMOS and hybrid spin-CMOS bundled data implementations. © 2018 IEEE.

Discussion

Non-volatile spintronic device-based Muller C-Element designs have beneficial characteristics relating to asynchronous architectures. The designs proposed through the course of this Section realizes a compact non-volatile Muller C-Element that's capable of instant on/off functionality with higher speed and lower power compared to previous non-volatile Muller C-Element designs. Additionally, analysis of driving-transistor widths and TMR effects on performance characteristics are detailed. The best performing design was simulated using HSPICE within a four-phase NCL pipeline and demonstrated instant store/restore power-gating

functionality. Furthermore, Bundled Data protocols have been shown to have reduced overheads when using the Spintronic Muller C-Element proposed herein.

Summary

This Chapter detailed two new digital circuit elements utilizing a hybrid spintronic-CMOS approach to achieve intriguing benefits, specifically these contributions are provided:

- utilizing the self-complementing nature of the DWCSTT to directly interface with CMOS to provide low-area non-volatile logic,
- the development of a novel compact nonvolatile edge triggered flip flow that requires no additional store and restore circuitry or signals prior to power-gating,
- the development of a novel compact spintronic Muller C-element that requires no additional store and restore circuitry or signals prior to power-gating, and
- the utilization of the spintronic Muller C-element to reduce Bundled Data asynchronous protocol overheads by negating the requirement of CMOS delay elements.

This Chapter also details the effects of improved TMR on performance from improved MTJ manufacturing techniques as well as the effects of transistor sizing on performance characteristics.

CHAPTER FOUR: BINARIZED DEEP NEURAL NETWORKS WITH STOCHASTIC SPINTRONIC NEURONS

Deep Neural Networks (DNNs) have realized impressive feats of intelligence, surpassing humans in specific tasks and achieving high efficacy at speech recognition, machine translation, and higher-level human-like reasoning activities such as interpretation and classification of visual art [75]. Although there are many successful architectural models used by DNNs, their common characteristic is the use of many layers of hidden nodes to realize "deep" topologies of non-linear neurons with linear weighted connections trained by backpropagation to distinguish complex inputs with high degrees of accuracy [39]. However, this beneficial characteristic of having many layers results in high computational demands and a large memory footprint. Thus, recent works into reducing the computation and memory overheads of DNNs have investigated the possibility of attaining similar recognition capabilities using reduced-precision approaches which incur significantly lower computation and memory demands. By reducing these overheads, in-situ networks could potentially be realized on resource-constrained platforms such as mobile and Internet of Things (IoT) devices [76].

Promising advancements towards this goal have focused on the substitution of high-precision floating-point parameters with binary representations, which replace expensive multiply-and-accumulate computations with bitwise logical operations and bit counting. These Binary Neural Networks (BNNs) have replicated some incredible feats of narrow intelligence demonstrated by Deep Learning with energy profiles that could be implemented on more resource-

constrained systems by using efficient custom neuromorphic hardware with emerging computing devices [40, 77-80]. This has led to the development of neuromorphic hardware accelerators that can implement such networks in a highly efficient manner [77-80]. The work presented in this Chapter extends these works by utilizing an emerging compact stochastic device, called the probabilistic bit (p-bit) [58], that naturally implements a non-linear Probabilistic Activation Function (PAF), [81]. The low-current operation of the p-bit allows for a seamless integration with low-voltage and high-resistance Resistive Random Access Memory (RRAM) crossbar and pseudo-crossbar arrays, which leads to very low power. In addition to the novel PAF proposed herein, we analyze how process variations in RRAM devices impact the performance of BNNs that are implemented using our scheme, and how such variations can be mitigated with in-situ training.

To summarize, this work provides the following contributions:

1. We demonstrate the feasibility of a compact PAF that uses just 4.98 μW combined with parallel binary RRAM pseudo-crossbar arrays for a low power of 75 nW per each weighted connection having an excitatory input, and
2. We evaluate the effects of RRAM process variation rates up to 50% on the recognition rate for the CIFAR-10 image recognition dataset using a convolutional neural network and demonstrate how on-chip learning can mitigate the resulting performance degradation.

Background

In this section, the relevant background information on BNNs and recent works on accelerating BNNs in hardware using emerging devices are reviewed.

Binary Neural Networks

Convolutional Neural Networks (CNNs), the popular class of DNNs that we investigate herein, are multi-layered networks that typically consist of several convolution layers, which convert high dimensional data, such as RGB images, into features, followed by a number of fully connected layers and terminated with a Log SoftMax layer for classifying the input data objects into labels based on their features, as depicted in Figure 16 [39]. Within each layer, either convolutional or fully-connected, the primary computations are realized by abstract neurons that calculate the weighted-summation of their input connections. Each neuron then computes a non-linear activation function of that weighted-sum. There are many different activation functions used with DNNs, such as rectified linear units, tanh, sigmoid, and others. Since DNNs traditionally utilize high precision floating-point representations of millions and sometimes billions of parameters, they incur significant memory requirements and computational operations during their training and deployment phases. They also are subject to the high latency overhead of transferring data between the memory and processor in traditional von-Neumann architectures. Thus, the development of BNNs which discretize the weights and activations of DNNs to binary values, as shown in Figure 16, can greatly reduce the memory and computation overheads of training and utilizing DNNs. Namely, the expensive floating-point operations can be simplified into highly-

efficient bitwise computations using bit-counting [40]. If we realize these memory and computation overhead reductions by implementing BNNs with custom neuromorphic accelerators using resistive devices in crossbar and pseudo-crossbar topologies [78-80], we can also reduce physical chip area, energy, and computation time requirements, which we expound upon in the following sections.

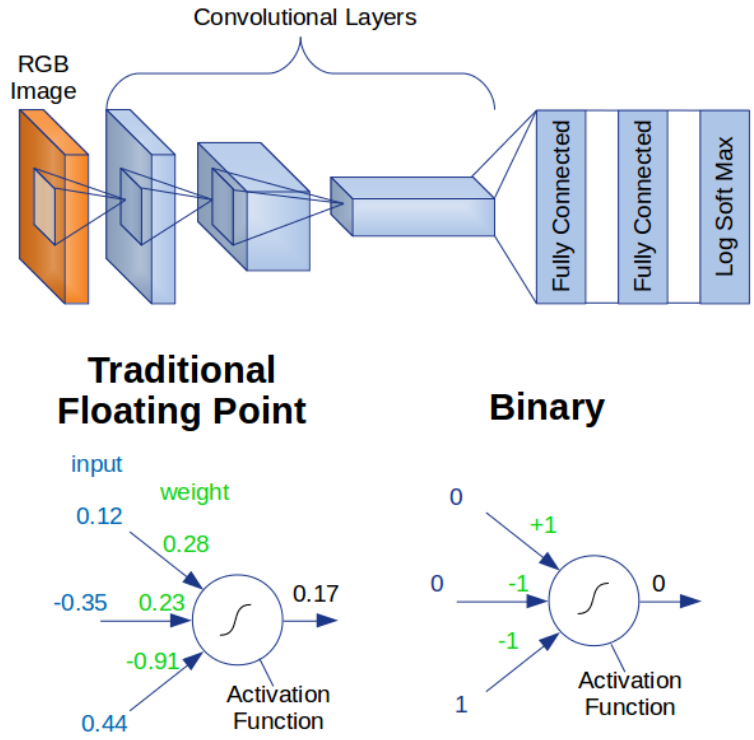


Figure 16: Convolutional DNN structure along with representative neurons for both floating-point and binary representations. © 2019 IEEE.

The binarization of weights in BNNs are typically constrained to (-1, +1) values, while activation functions have been explored with both (0, 1) and (-1, +1) constraints [78-80].

Additionally, the activation function is typically implemented with a deterministic sign function due to its straightforward implementation on most hardware [78-80]. Even though stochastic binarization has been suggested to be more appealing, the generation of random bits in hardware normally requires large pseudo-random number generators [40]. However, the approach proposed herein utilizes a novel stochastic spintronic device that uses ubiquitous thermal noise to generate random bits in a compact and low-power PAF circuit. Another consideration for BNN accelerators using resistive crossbar arrays is the effect that process variation has on the accuracy of the network. With deviations from the device's ideal resistance values, the weights effectively shift from their intended values, which as we show later, can cause a significant increase in the error rate. However, we demonstrate that by incorporating the hardware into the backpropagation training loop, it is possible to mitigate almost all degradations of accuracy associated with process variation.

Recent Work on Binary Neural Network Hardware Acceleration

Several recent works have aimed towards realizing BNN acceleration through the utilization of emerging devices, such as RRAM and spintronics, to compute the necessary binary operations in-memory. This frees up chip area and eliminates bottlenecks in the data pathways between memory and computational resources. A selection of these works in Table 4 is compared on the basis of Transistor Count, which determines the silicon die area that is needed to interface with and facilitate computational operations in the memory array, the Sequential/Parallel operation of the in-memory computations, and the Variation Degradation Factor. The latter quantifies the

accuracy degradation, defined as the increase in mean percent recognition error across the CIFAR-10 dataset divided by the percentage value of single-sigma variation in resistance of the RRAM elements, as described subsequently herein. The influential work of Sun et al. proposed an RRAM-based XNOR BNN that is capable of both sequential operation by computing the XNOR of inputs and weights one input bit at a time, summing the result, and then applying the sign activation function and also parallel operation by using two input lines per input bit and two single-transistor/single-resistive-element (1T1R) cells per weighted input to compute parallel bitwise XNOR operations and then using a Sense Amplifier output to realize the sign activation function [80]. Ni et al. proposed a sneak-path-free binary crossbar using two types of RRAM devices that do not require a select transistor for each bit cell, and their activation function is determined by a voltage comparator, which uses 16 transistors [78]. They found that a 29% variation rate in the resistance values of the bit cells degraded accuracy by 4%, leading to a Variation Degradation Factor of 0.138. The work presented herein uses a compact, low-power PAF and an RRAM array to conduct parallel BNN computations that are resilient to RRAM variations. With a Variation Degradation Factor of 0.02 when used with on-chip training, concerns about process variation are practically eliminated.

Table 4: Neuron attributes of recent BNN approaches. © 2019 IEEE.

Ref.	Transistor Count	Sequential/Parallel	Variation Degradation Factor
[80]	14	Mixed	N/A
[78]	16	Parallel	0.138
[77]	>10	Sequential	N/A
Herein	*4*	*Parallel*	*0.02*

Accelerator Design

The binarized Deep Neural Network (DNN) neuromorphic accelerator developed in this Chapter is depicted in Figure 17. It consists of multiple Binary Neural Network (BNN) layers, where each layer contains a pseudo-crossbar array, along with the associated Probabilistic Activation Functions (PAFs) at the outputs, that corresponds to either a convolution kernel or a fully-connected layer, as determined by the BNN architecture that is being implemented. In addition to the BNN layers, our simulations assume that a rudimentary on-board CPU or ASIC with access to sufficient RAM is used for handling the training logic and backpropagation calculations, as well as storing and delivering the training/test data and labels. Those resources do not significantly impact the computational burden, as the majority of the calculation workload is engaged when computing each DNN layer's weighted sums and activation functions.

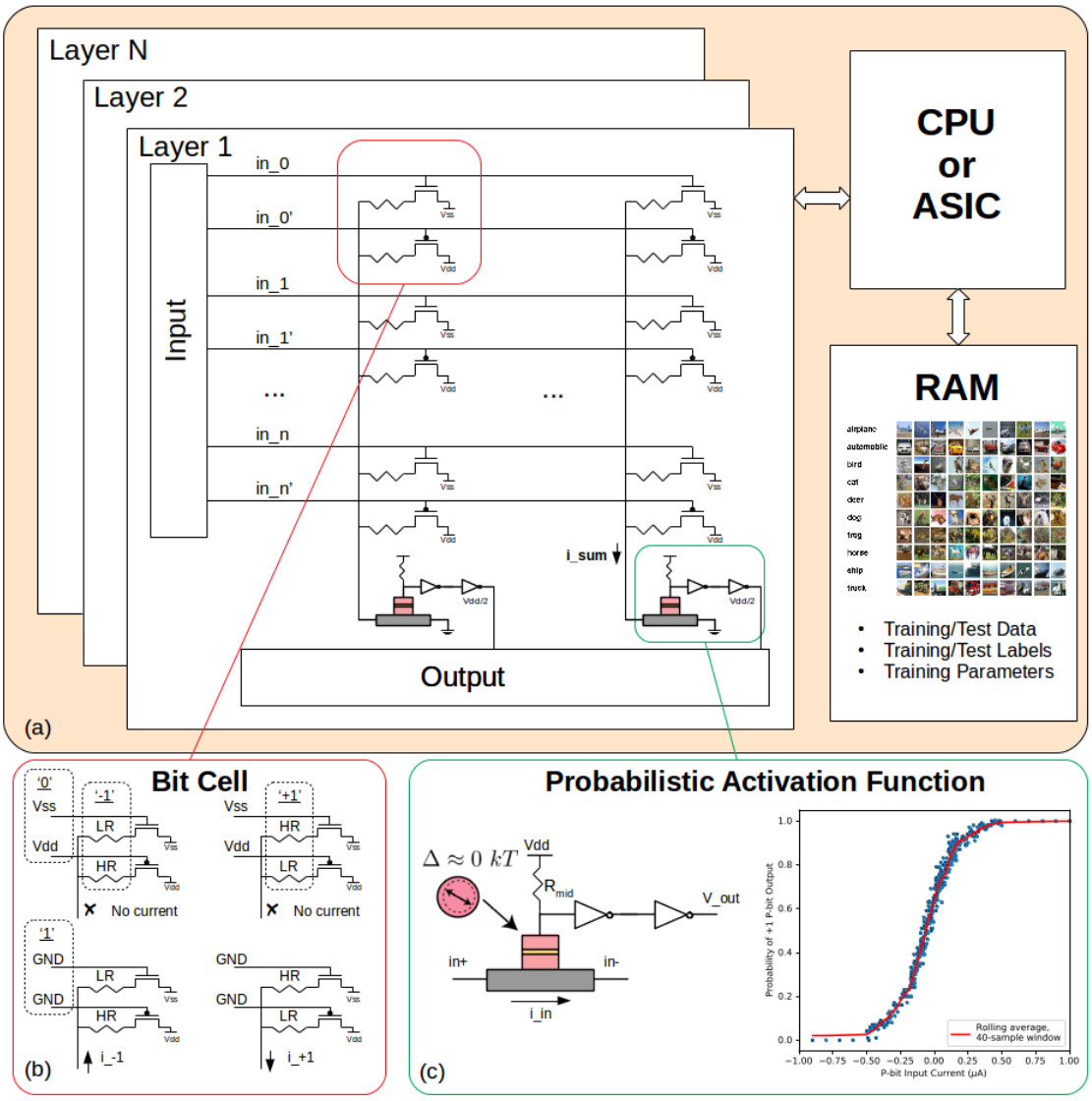


Figure 17: Neuromorphic accelerator proposed in this Chapter. © 2019 IEEE.

Pseudo-Crossbar Array

Each layer in the neuromorphic accelerator performs three computations in parallel between all input and output bits. The first computation is the bitwise multiplication of the binary input and the stored binary weight between each input and each output. This computation is

realized using two complementary memristors per input/output pair as well as two signal lines per input, which is similar to resistive pseudo-crossbars previously developed for processing BNNs [78-80]. As shown in Figure 17b, the paired input wire (in_i, in_i') voltages signify a value of either '0' or '1'. The signal level '0' is represented by an input pair value of (Vss, Vdd) (here Vss is chosen to be -Vdd) which deactivates both the PMOS and NMOS transistors in the bit cell, allowing no current to flow regardless of the weight value. An input value of '1' is represented by an input pair value of (GND, GND), which activates both transistors, allowing current to flow through both branches. Each memristor pair corresponds to a '-1' or '+1' weight, depending on which memristor is in the High Resistance (HR) state and which memristor is in the Low Resistance (LR) state, as depicted in Figure 17b. When the input value is '1', meaning both transistors in the bit cell are on, the branch with the LR memristor sinks/sources the vast majority of the current flow in the bit cell, due to the large resistance ratio between the LR and HR states. Thus, if the LR branch connects to the NMOS and Vss, it will sink current, corresponding to an output value of '-1', and if the LR branch connects to the PMOS and Vdd, it will source current, corresponding to an output value of '+1'. Thus, the three possible output values of the bitwise input and weight multiplications are 0, -1, and +1.

The second parallel calculation performed in the BNN layer is the bit-counting operation, which is the summation of the parallel bitwise multiplications described previously along a single output path. This is accomplished by connecting the in- terminal of the p-bit to GND and the in+ terminal to the bit line that connects all of the bit-cells within a single column, whereby the accumulation of currents according to Kirchoff's current law corresponds to parallel input-weight

multiplications, which equates to one of the (0, -1, +1) current-level values described previously. The final calculation in the BNN layer uses the accumulated current summation from the previously described computation as the input to a PAF, which outputs a '0' or '1' with a probability according to a sigmoidal function as shown in Figure 17c. The implementation of the PAF is described next.

Probabilistic Activation Function Circuit

In order implement the PAF, we leverage the probabilistic-bit (p-bit) design proposed by Camsari et al. [58] and shown in Figure 17c, which is a low-power and compact circuit capable of generating random bits from thermal fluctuations. The p-bit is a Spin-Hall-Effect driven Magnetic Tunnel Junction (MTJ) device with a very low energy barrier, which allows thermal agitations to stochastically switch the free layer of the MTJ between its parallel and antiparallel states on sub-nanosecond timescales. Since a current flowing through the bottommost heavy metal layer can effectively bias the free layer of the MTJ, the probability of the MTJ being in HR or LR states can be tuned along a sigmoidal function via the input current, as shown in Figure 17c. By placing a resistor, R_{mid}, having a fixed resistance equal to the average of the HR and LR states of the MTJ between the MTJ and VDD, we can use a voltage divider to switch a pair of inverters, giving us a digital representation of the state of the MTJ, which represents the output of the PAF. Since the energy barrier of the p-bit is very low, it requires current on the order of 100's of nA to bias the PAF, which allows for low-energy BNN calculations using low voltage and highly-resistive scaled RRAM devices for weighted connections.

Simulation Framework

The simulation framework utilized herein is depicted in Figure 18 and consists of HSPICE modeling of the circuit-level parallel bitwise and bit-counting operations of the RRAM pseudo-crossbar as well as the p-bit PAF for determining the circuit behavior under different RRAM variations and PAF properties, which are then modeled in PyTorch for training and testing on Nvidia Tesla V100 GPU clusters.

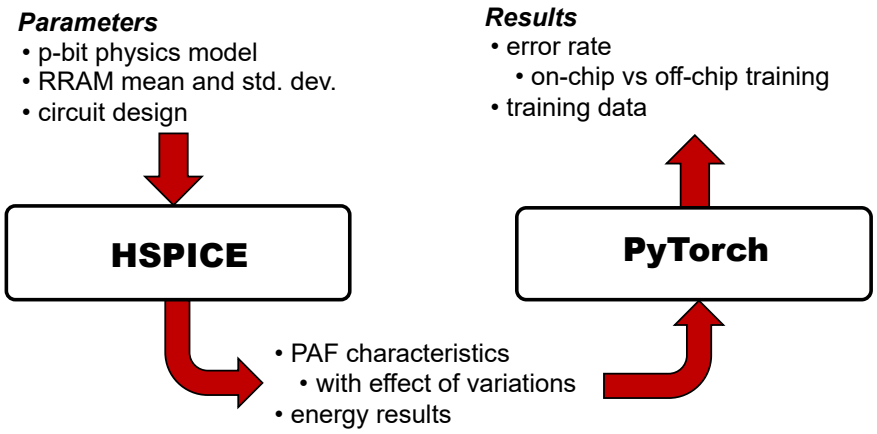


Figure 18: Simulation Framework for Stochastic Binarized Deep Neural Network Accelerator.

HSPICE Simulations

On the HSPICE platform utilized, 14nm PTM transistor models [82] were deployed along with RRAM resistance values of N(5 MΩ,σ) for the LR state, and N(50 MΩ,σ) for the HR state, where N(μ,σ) is a normal distribution with a mean of μ and a standard deviation of σ, which is varied from 0%-50% of μ in 5% increments, and the p-bit is modeled using experimentally verified

physics modules from the Modular Spintronics Library [83] at a Vdd of 0.6V. PAF activation probability was measured using Monte Carlo simulations of 100 samples each for 480 different p-bit input currents. The results, shown in Figure 17c, depict the sigmoidal probability of the p-bit's output voltage representing a '1' signal level.

PyTorch Simulations

For training BNNs using the PyTorch framework, we extend the code developed in [40] to include the impacts of weight-resistance distortion resulting from device variations, as well as implementing our p-bit based PAF. During training, high-precision weight values are stored in RAM, which are used for gradient calculations as per the training algorithm for BNNs developed by Hubara et al [40]. However, the activations in the forward pass are determined strictly by the binary weight values with their associated variations, as if the computation was done on-chip. Although some mismatch may still occur between the high-precision weights used for the backward pass and the on-chip weight variations used for the forward pass, we found that performance is still improved by using the real on-chip weights during the forward pass. The p-bit PAF circuit from the HSPICE simulation is approximated with a probabilistic hard sigmoid function.

The CNN used herein is the same one used for [80], which has six convolutional layers and three fully-connected layers and was trained on the CIFAR-10 dataset [84]. We do not include any pooling layers, however we utilize a stride of 2 on CNN layers 2, 4, and 6. Although it is not explicitly claimed within other BNN hardware accelerator literature, we found that binarizing the

final layer results in a significant increase in the error rate, whereby if only the last layer is chosen to be LogSoftmax, the error rate reduces to very tractable levels. Thus, we use LogSoftmax for the final layer, which would be computed using the CPU or ASIC in our scheme. Since the final layer is a small fraction of the total size of the network, the large accuracy improvement can be worthwhile given the acceptable minimal performance degradation incurred.

On-Chip vs Off-Chip Training

When developing BNN accelerator hardware, one has the choice to either train the network off-chip using ideal (-1, +1) weight values and then download the final weight configuration to the as-built hardware, or train the network on-chip, using the actual circuit and its associated process variations within the training loop. The latter approach is a distinguishing feature of our design scheme developed herein. We leverage the useful property that device-to-device resistance variations are effectively variations to the (-1, +1) weight values computed when training and testing the network in PyTorch. We compare the performance of on-chip and off-chip training techniques with respect to the final error rates under multiple levels of resistance variation, and we show that even with the anticipated rates of variation measured experimentally for fabricated devices, on-chip training allows the learning mechanism to produce BNN configurations that achieve error rates that are nearly identical to the error rates corresponding to idealized networks without device variation [85].

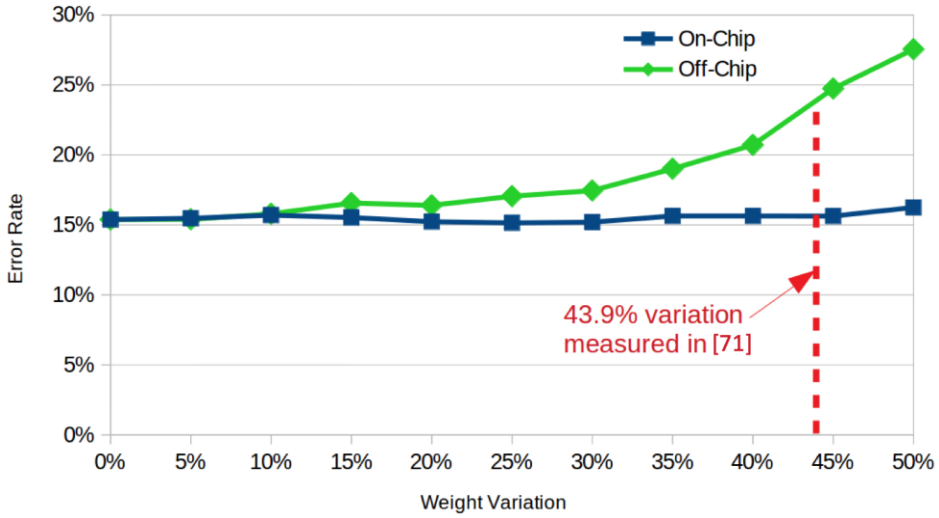


Figure 19: Effect of weight variations for on-chip vs off-chip training. © 2019 IEEE.

Results

The results for the simulations conducted as described in the previous sections can be summarized in Figure 19 as follows. We simulated at least 100 epochs for on-chip and off-chip training with weight variations from 0% to 50% in 5% increments for as-built process variation and/or degradation over the devices' lifetime of operation. The lowest error rate for each condition is shown in Figure 19, and we show the error rate per epoch for a selection of test cases in Figure 20. The lowest error rate for ideal weights with no variation is determined to be 15.38%. The increase in error rate due to weight variations for the off-chip training condition is negligible under 15% variation, staying within a 2% error rate increase for up to 30% variation, but increases significantly thereafter, reaching a maximum of 12.17% increase in error rate at 50% variation. However, for the on-chip training approach developed herein, the increase in error rate fluctuates within 1% for all weight variations, demonstrating the robustness that is achieved by utilizing the

intrinsic device variations within the forward pass for training. Such a result is crucial for deeply-scaled beyond-CMOS devices that exhibit significant process variation. In addition to the error rate analysis, our HSPICE simulations demonstrated that each bit cell that has an input value of '1' consumes an average power of 75 nW, and that the p-bit PAF uses just 4.98 µW, demonstrating a very low-power scheme for accelerating BNNs.

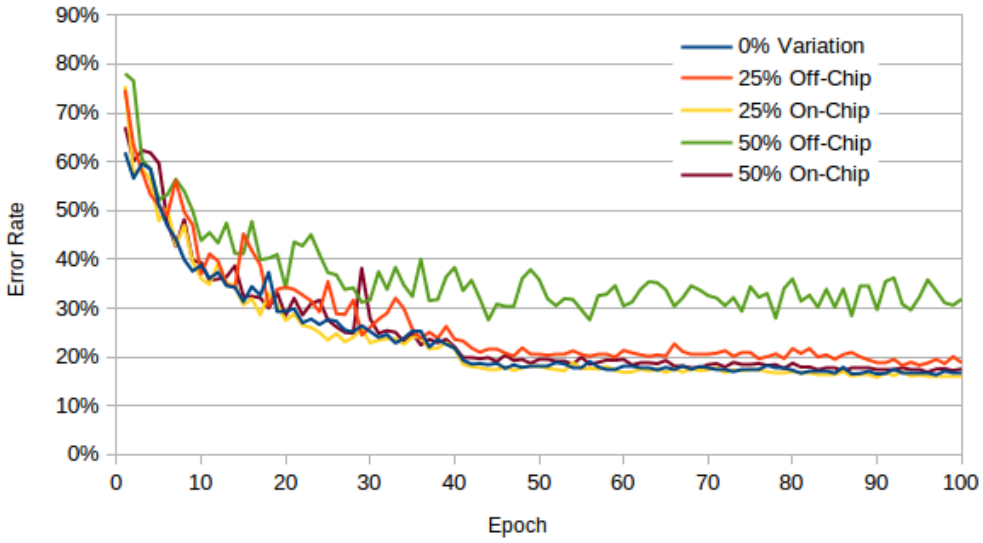


Figure 20: Error rate per epoch for a selection of off-chip and on-chip test cases. © 2019 IEEE.

Summary

The BNN hardware accelerator proposed in this Chapter uses a novel spintronic device to realize a compact implementation of a PAF that naturally integrates with current-summation-based crossbar and pseudo-crossbar arrays for low-power parallel BNN computations of just 75 nW per

activated bit cell and 4.98 uW per PAF. Additionally, we demonstrated that this approach is highly resilient, even to extreme process variation, when utilizing an on-chip training framework. Taken altogether, such a scheme is well-situated for highly-scaled BNN acceleration on resource-constrained platforms such as mobile and IoT.

CHAPTER FIVE: SPINTRONIC STOCHASTIC SPIKING NEURONS

In this Chapter, two approaches for implementing asynchronous stochastic spiking neuron circuits with alternative benefits are presented. First, the Spintronic Stochastic Spiking Neuron (S3N), is designed with a preference towards high-speed and low device count, Second, the Subthreshold Spintronic Stochastic Spiking Neuron (S4N) is designed for ultra-low-power subthreshold operation.

Spintronic Stochastic Spiking Neuron

The S3N circuit is depicted in Figure 21. It consists of a spintronic p-bit device to provide a tunable stochastic output via a bias driven by the input current at i_{IN}, a capacitor (C_{MEM}) representing the membrane potential of a neuron to accumulate temporal information about the state of the p-bit, two inverters preceding C_{MEM} to sense the state of the p-bit, two inverters proceeding C_{MEM} to detect if the voltage V_{MEM} has reached a threshold (V_{th}), which is the same as the threshold for a CMOS inverter, and an NMOS (M_0) to discharge C_{MEM} upon detecting an output spike. The overall circuit operation is as follows: by tuning the input current, i_{IN}, the p-bit will be stochastically biased towards either its high-state or low-state, with a statistically equal amount of time between the two states if i_{IN} is zero. Based on the state of the p-bit, C_{MEM} will either charge or discharge, and if charged enough, $spike_{OUT}$ will go high, subsequently turning on M_0, which discharges C_{MEM} and then sets $spike_{OUT}$ low. Thus, generating brief pulses, or spikes,

at $spike_{OUT}$ with timing characteristics dependent upon transistor parameters and the capacitance of C_{MEM}.

It is worthy to note that the stochasticity of the p-bit device is due to the effects of thermal noise on low-energy barrier nanomagnets, and the tunability is introduced by methods of magnetic bias, such as the SHE used herein. Therefore, alternative designs of p-bit devices that utilize alternative methods of magnetic bias, such as the magnetoelectric effect [87], can be readily implemented with the S3N scheme, providing future avenues of exploration and improvement to the design.

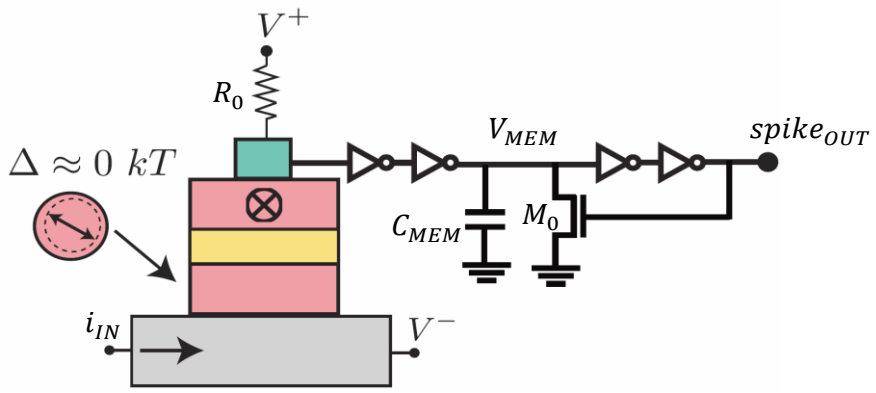


Figure 21: The Spintronic Stochastic Spiking Neuron circuit with Spin-Hall driven p-bit.

Second Order Synapse

In order to emulate the postsynaptic transient currents found in biological neurons following a preceding spike, the Neuromorphic VLSI second-order synapse developed in [88] and

depicted in Figure 22 is utilized to convert incoming spikes into linearly additive temporally extended current pulses. The circuit is essentially a cascade of two current-mode lowpass filters, whereby the effective weight of the circuit can be tuned by adjusting V_W, and the temporal characteristics of the circuit can be tuned by adjusting V_f, V_s, C_1, and C_2. For the results demonstrated in the following section, V_W is either fixed or varied deterministically in order to demonstrate the effect of varying the weight. As such, no learning mechanism has been implemented. This synapse circuit was chosen due to its high degree of biological mimicry, demonstrating full neuron-synapse-neuron communication as similar to biological structures in addition to its utility in demonstrating an elementary computational network in the following Section. Although we utilize the synaptic circuit herein in a completely excitatory sense, such a circuit could be used for inhibitory currents by connecting i_{OUT} to the V^- terminal of an S3N. As will be discussed later, alternative synaptic architectures that prefer area efficiency over biological mimicry, such as crossbar arrays, could be utilized with the S3N for dense Stochastic Spiking Neural Network computational paradigms.

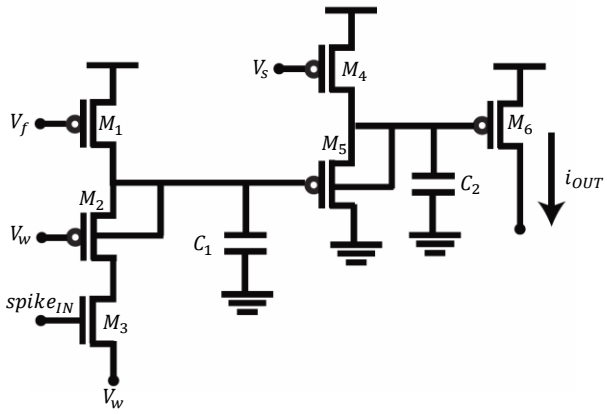


Figure 22: Second-order neuromorphic synapse used herein.

Results

The simulations provided in this Section were performed in HSPICE using 14nm FinFET transistor models [67]. The p-bit device was simulated using benchmarked SPICE spin-circuit models from the Modular Spintronics Library [83]. All inverter and synapse transistors are minimally sized with just a single fin, and M_0 has 5 fins. It is critical that M_0 has stronger driving characteristics than the inverter transistors since it must pull down V_{MEM} regardless of the state of the p-bit device. After testing multiple values for C_{MEM}, 10fF provided the desired circuit characteristics. V_f, V_s, C_1, and C_2 were found by experimentation and are included with the parameters provided in Table 6.

Table 5: S3N and Second-Order Synapse Simulation Parameters.

Parameter	Value
V_{dd}	0.7 V
C_{MEM}	$10\ fF$
R_0	$500\ k\Omega$
C_1	$1\ fF$
C_2	$1\ fF$
V_f	$0.54\ V$
V_s	$0.47\ V$

a) Stochastic Spiking Neuron

Figure 23 shows the results of a single S3N neuron receiving a stepwise increasing input current. Every nanosecond, the input current is increased by 50nA. $\widehat{m_z}$ is the z component unit vector of the magnetization of the free layer of the p-bit. As shown, when i_{IN} is low, $\widehat{m_z}$ stochastically switches between +1 and -1 in about equal amounts. As i_{IN} is increased, $\widehat{m_z}$ becomes increasingly biased towards -1, which is the high-state of the p-bit in this configuration, while still exhibiting stochastic switching. When the p-bit is in the high state, V_{MEM} begins to charge, and if it is asserted for a long enough period, V_{MEM} will reach V_{th} and a spike is generated at $spike_{OUT}$ and V_{MEM} is subsequently pulled down. Thus, the poissonian spike rate of the S3N can be controlled via i_{IN}. The power consumption of the S3N with an input current of 0.8uA, which ellicites a very high rate of spiking is 9.6uW, and with an input current of 0uA, which ellicits almost no spiking, the SSN uses just 0.6uW. The average spike width from $\frac{V_{dd}}{2}$ to $\frac{V_{dd}}{2}$ is just 15

ps. The average spike interval during high rates of spiking, such as with an input current of 0.8uA, is about 120 ps.

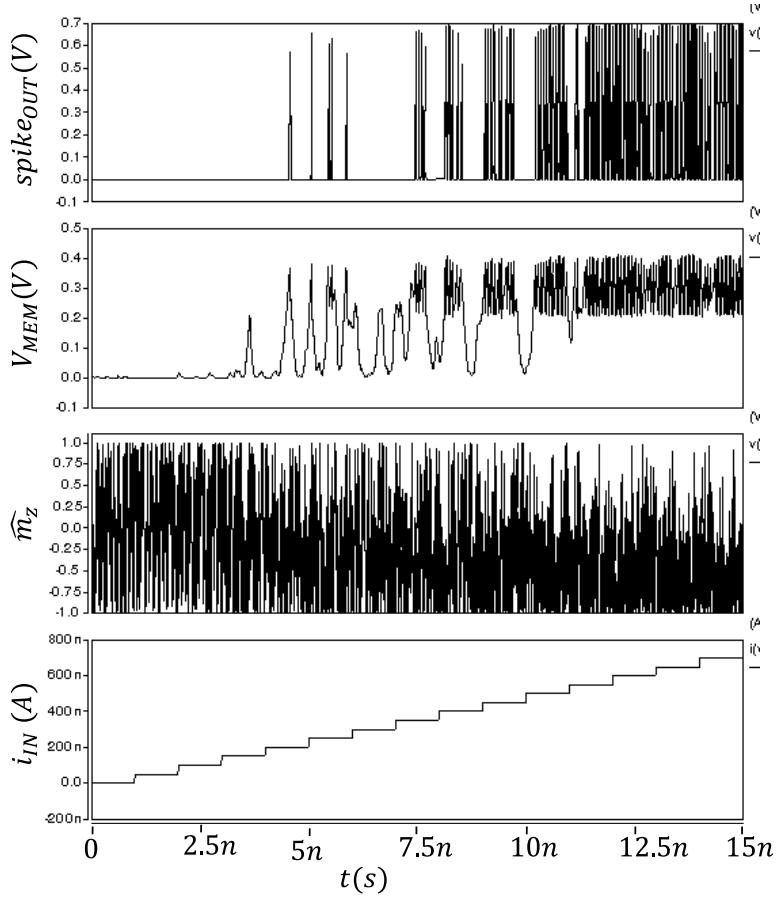


Figure 23: Stochastic Spiking Neuron simulation graphs illustrating, from the bottom up, i_{IN}, $\widehat{m_z}$, V_{MEM}, and $spike_{OUT}$.

b) Synaptic Dynamics and Weight Control

The S3N combined with the second order synapse circuit was simulated by connecting the $spike_{OUT}$ of the S3N circuit to the $spike_{IN}$ terminal of the synapse and applying a fixed current

of $0.5uA$ at the i_{IN} terminal of the S3N with V_w set to 0.14V, which can be considered as a strong weight, which means that the output current is significant enough to elicit a high rate of spikes in the post-synaptic S3N if the pre-synaptic S3N is strongly spiking. As shown in Figure 24a, the output current of the second order synapse, i_{OUT}, follows a prolonged and slightly delayed integration of the incoming spikes. Single spikes or a few dispersed spikes have little effect on i_{OUT}, but prolonged periods of intense spiking ellicit a strong increase in output current, similar to the EPSPs found in biological neurons. The saturation current of i_{OUT} depends upon the weight of the synapse, which is determined by the voltage at V_w. Figure 24b shows the effect of decreasing V_w, which effectively increases the weight of the synapse. A single-input S3N with an input current of 0.7uA is used to generate the spike pattern $spike_1$, which is then fed to the $spike_{IN}$ terminal of a synapse, whose resulting output current is used as the input current for another S3N to generate the spike pattern $spike_2$. All other parameters are the same as previous simulations. As shown, when V_w is decreased, the synaptic output current, i_{OUT}, saturation point is increased, and thus, the spiking rate of $spike_2$ increases as well. The potential range of current output from the synapse circuit is quite large compared to the effect it has on proceeding S3Ns since V_w could potentially be varied from gnd to V_{dd}, and our results show that just varying V_w from 0.14V-0.2V is enough to modulate the output from very high spiking to almost no spiking.

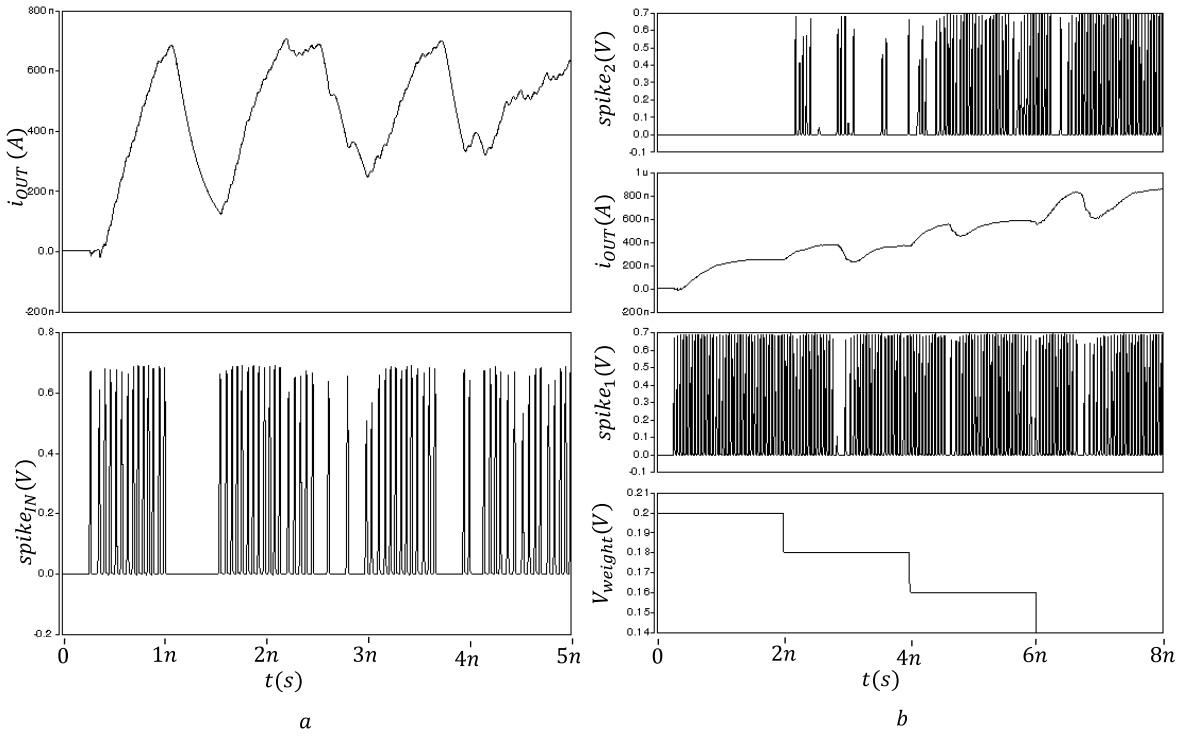


Figure 24: Simulation transients of Stochastic Spiking Neuron with Neuromorphic Synapse.

c) Boolean Two-Input Perceptron

In order to demonstrate rudimentary computational capabilities utilizing the S3N, we simulated a two-input one-output perceptron implementing AND and OR logic functions. For this demonstration, a high rate of spikes indicates a logic '1', and a low rate of spikes indicates a logic '0'. The circuit consists of two (input) S3Ns whos output terminals are connected to two synapses, whose outputs are combined into the input of a third (output) S3N. For both functions, the circuit topology is the same, and just the weight, V_w, is changed for both synapses, effectively changing the network operation. By using a high weight of 0.14V, the output SSN will spike at a high rate when either of the inputs spike at a high rate; thus, implementing OR logic. This is shown in Figure

25a, where input current pulses of 0.7uA are applied to input 1 from 2ns-4ns and from 6ns-8ns and to input 2 from 4ns-8ns. As shown, a high rate of spiking activity at the output ($spike_3$) occurs when either input is firing. Figure 25b shows the results of the same simulation, but with V_w at 0.2V, which is equivalent to a low synaptic strength. As shown, the output only becomes highly active when both inputs are active with a high rate of spikes, implementing AND logic.

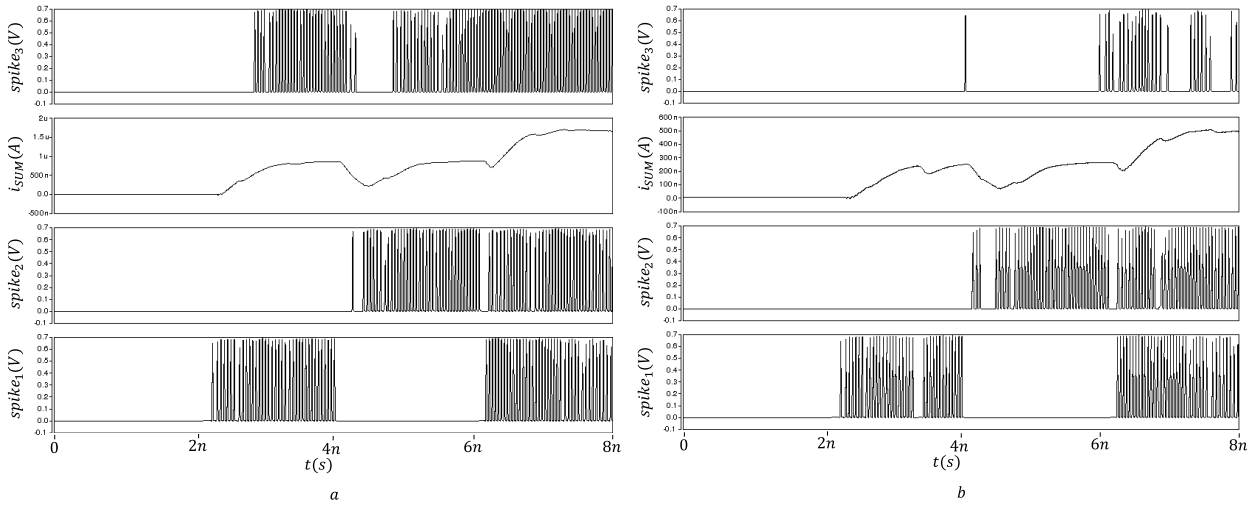


Figure 25: Simulation transients of S3N with Neuromorphic Synapse implementing Perceptron functionality.

Discussion

The primary contribution of the S3N is the demonstration of a novel compact stochastic neuron circuit that leverages true randomness from thermally-driven magnetic excitations in an ultra-low energy barrier spintronic device to generate high speed spikes in a fashion which exhibits a Poisson distribution. As such, a rather simple choice of synapse and test case was used to demonstrate and evaluate the speed, power, and biological mimicry of the design without

designing an elaborate architecture, which could be done in future works as will be discussed later. Therefore, the synapse circuit chosen, although biologically-mimetic, requires a significant device count for each synapse and utilizes voltage levels for weight-value implementation, which would require additional memory and programming elements in order to implement the weight in a programmable/learnable manner. For the perceptron test cases used herein, we assumed fixed weights that were tailored to the particular operation.

Secondly, the SHE-driven p-bit used in this work that combines an MTJ that has a thermally unstable nanomagnet with a heavy metal exhibiting SHE has not been experimentally demonstrated yet, even though each individual component has been demonstrated by different authors. 2-Terminal MTJs having unstable free layers have been experimentally utilized for TRNG applications [89-92] while SHE driven MTJs with stable free layers are also commonly demonstrated [61, 62] for memory applications. More recently, an embedded Magnetoresistive Random Access Memory (MRAM)-based implementation of a p-bit was proposed that uses a 2-Terminal MTJ with an unstable free layer along with an NMOS transistor [65]. The main results would remain essentially unchanged if such an alternative p-bit replaces the 3-Terminal device proposed herein, and whether a 3-Terminal device proves to be more flexible due to the separate control terminal deserves further study.

Theoretical neuroscience has demonstrated that networks of stochastic neurons having firing rates which follow a Poisson distribution can achieve Markov Chain Monte Carlo (MCMC) sampling of an underlying probability distribution as encoded by their weights. Referred to as *Neural Sampling*, various aspects of probabilistic inference become feasible, which provides a

particularly interesting explanation being elaborated for various cognitive processes [44]. Thus, a natural application and extension to the S3N functionality developed herein is to leverage it to implement hardware-based neural sampling networks using intrinsic thermally-driven stochasticity in a low area hardware design operating with low energy consumption. Realization of hardware-based artificial neural network acceleration leveraging the stochastic properties of spintronics leads to a fresh direction towards increasing performance and efficiency. Namely, software-based approaches to artificial intelligence systems suffering from massive switching plurality due to an underlying binary-value representation and layers of software bloat are reduced substantially.

With regards to the underlying learning paradigms, it has been demonstrated that competitive networks of stochastic neurons with lateral inhibition, a structural organization prevalent throughout the mammalian cortex, in conjunction with very simple Hebbian learning rules converges high-dimensional stochastic spiking inputs to an implicit generative model through Expectation Maximization [93]. With such a generative model, Bayesian computations are readily implemented for probabilistic inference in both the spatial and spatio-temporal regimes, giving the ability to make predictions and classifications on new data. Therefore, utilizing the S3N with an appropriate synaptic architecture, one could realize a computational system that intrinsically "learns" a generative model of high dimensional input distributions with improved performance and efficiency over software-only based approaches or CMOS-only hardware accelerators.

In order to alleviate the utilization of floating point weights, which either require a large amount of memory per synapse (32-64 bits), or are difficult to reliably encode intrinsically in

hardware, such as through memristors, several works have demonstrated impressive results of classification and detection utilizing binary synaptic weights with probabilistic Hebbian learning rules [57, 94-96]. Hence, it should be possible to implement S3Ns with dense arrays of binary stochastically switching memory devices, such as STT-MTJs or CBRAM, to realize dense and fast unsupervised learning architectures for future cognitive systems.

Subthreshold Spintronic Stochastic Spiking Neuron

The S4N is inspired by the principle that the cortex consists of noisy and imprecise components in order to realize an ultra-low-power stochastic spiking neural circuit that resembles biological neuronal behavior, which can be used as a building block for future biologically-inspired computational paradigms. By utilizing probabilistic spintronics to provide true stochasticity in a compact CMOS-compatible device, an Adaptive Ring Oscillator for as-needed discrete sampling, and a homeostasis mechanism to reduce power consumption, provide additional biological characteristics, and improve process variation resilience, this subthreshold circuit is able to generate sub-nanosecond spiking behavior with biological characteristics at 200mV, using less than 80nW, and with good behavioral robustness to process variation.

Circuit Overview

The S4N is motivated by the desire to realize a minimal-complexity, ultra-low-power circuit that intrinsically behaves similar to the noisy heterogeneous neurons in the cortex, at least in the sense that it can be relevant for implementing Neural Sampling. This has led to a circuit that

appears at first glance rather different than traditional rate-based spiking neuron schemes where the output is purely a Poissonian spike rate, yet the S4N is still relevant in the following ways.

1. The S4N generates samples (or spikes) where the rate is somewhat deterministic and periodic, but the 'strength' of the samples is determined by a sigmoidal relationship with the input voltage and a random variable.
2. The S4N output bears little resemblance to the spike signals found in typical spiking neuron designs, but they strongly resemble the double-exponential Post-Synaptic-Potentials (PSPs) found in biology that result from pre-synaptic spike trains.
3. A fast homeostasis mechanism not only modulates the sample strength in a fashion that closely resembles spike-frequency-adaptation found in biology [97], but also assists in balancing the network to be sensitive, but not too sensitive, even in the presence of process variation.
4. Process variation effects don't cause the circuit to fail, but simply modify the sigmoidal relationship between the input and output, such that the behavior of multiple neurons is heterogeneous, which is found in cortical neurons of the exact same type and region [98].

The S4N circuit shown in Figure 26 is implemented by what is essentially a voltage divider between an sMTJ and three transistors, $M_1 - M_3$, modulating the input to M_4, which acts like a voltage-controlled current source since its operating in the subthreshold region. The input voltage, V_{input}, modulates the resistance of M_1 in an exponential fashion, while also modulating the

Adaptive Ring Oscillator (ARO). The ARO, which is a five-inverter ring oscillator with an additional nmos Transistor in the second inverter controlled by V_{input}, as shown in Figure 26, oscillates at a frequency depended upon V_{input}, generating voltage pulses applied to M_2, which are considered to be samples. The ARO is used in place of a standard ring oscillator in order to save energy by sampling more frequently only when V_{input} is significant, and less when it is not. The resistance of M_3 is related to the homeostasis mechanism and modulated by V_b, which is a leaky exponential inverted integration of the output activity. During periods of high activity, V_{out} reduces the resistance of M_b enough to pull down V_b, increasing the resistance of M_3, increasing V_{state}, and lowering the current through M_4 during samples, resulting in a negative feedback to balance periods of high activity. By leveraging the high resistance of subthreshold CMOS devices, which results in low current operation, an ultra-low-power scheme is realized.

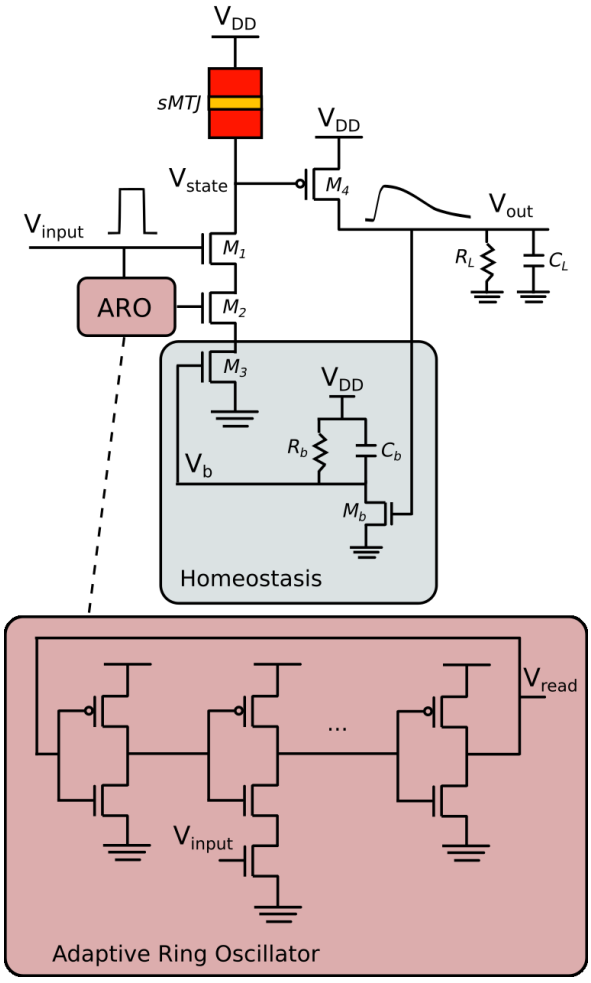


Figure 26: The Subthreshold Spintronic Stochastic Spiking Neuron circuit.

The stochasticity of the circuit arises from the stochastic switching of the sMTJ between Anti-Parallel (R_{high}) and Parallel (R_{low}) resistance states due to thermal noise. Although typical ratios forR_{high} and R_{low} is 100-150%, which is small compared to the exponential resistance changes of subthreshold CMOS, when the resistance of the lower branch is close to that of the sMTJ, the state of the sMTJ becomes significant in determining the strength of the output current

through M_4, where R_{low} will result in a significantly weaker signal than R_{high}. This results in three primary operating regions of the S4N that resembles the saturating and linear regions of a sigmoid:

1. When the resistance of the lower branch is $>> R_{high}$, such as when the ARO output is low, V_{input} is low, or V_b is low, then the output is saturated at the lower bound, providing little to no activity regardless of the sMTJ state.
2. When the resistance of the lower branch is $<< R_{low}$, such as when the ARO output is high, V_{input} is high, and V_b is high, then the output is saturated at the upper bound, providing maximum output activity regardless of the sMTJ state.
3. When the lower branch is $\sim[R_{low}, R_{high}]$, such as when the ARO output is high and V_{input}, V_b take intermediate values, the state of the sMTJ has a large influence on the output signal, resulting in stochastic spiking behavior.

An interesting observation detailed in the following section is that when a large constant input voltage is applied for a long enough time, the homeostasis mechanism balances V_b so that the resistance of the lower branch remains sensitive to the state of the sMTJ.

The output resistor R_L is used to leak V_{out} over time, and the output capacitor C_L is a very small value used in place of downstream CMOS devices in synaptic circuits that the circuit may drive. The signals shown above V_{input} and V_{out} in Figure 26 give an example of a single sample whereby a brief pulse equivalent to V_{DD} is applied to the input for enough time to elicit a single sample, and the resulting output waveform is shown, resembling a PSP.

Results

This Section analyzes the results of our simulations, which were performed using HSPICE with high-performance 7nm FinFET PTM Transistor models [67]. The sMTJ was modeled using physically benchmarked spintronic modules from the Modular Spintronic Library [83]. The other circuit parameters are listed in Table 7.

Table 6: Circuit Parameters for the Subthreshold Spintronic Stochastic Spiking Neuron.

Parameter	Value
V_{DD}	200 mV
R_{low}, R_{high}	6 MΩ, 15 MΩ
R_L	2 MΩ
C_L	0.5 fF
R_b	5 MΩ
C_b	2 fF

Figure 27 illustrates the S4N circuit behavior when applying voltage pulses of 50mV, 100mV, 150mV, and 200mV to V_{input} for 20ns, 20ns, 50ns, and 50ns, respectively, with 15ns periods of 0V in between. Since square voltage pulses are not the typical input voltage signals that would be propagated in networks of S4N circuits, the output of the S4N, V_{out1}, is connected to another S4N, and the output of that S4N, V_{out2}, is shown to illustrate how the circuit operates with in-situ signals. This can be considered a 1-to-1 network with a synaptic weight of 1. As shown,

V_{read}, which is the output of the ARO, oscillates with a rate proportional to V_{input}, and when the input is too low, such as for 50mV and 100mV, almost no output signal is generated at V_{out1}. For the case where V_{input} is 150mV, it takes a few samples from V_{read} before V_{out1} reaches its peak at just below 200mV, which is when the homeostasis mechanism reduces V_b so that V_{out1} decreases and stochastically jitters from higher and lower voltages due to the interplay between the homeostasis mechanism and the sMTJ, which corresponds to operational region 3 described in the previous Section; V_{out2} appears to only generate a single significant spike when V_{out1} is at its highest, although there are additional minor fluctuations. For the case where V_{input} is 200mV, only a single sample is needed to elicit a maximum voltage at V_{out1}, which subsequently reduces V_b such that the circuit operates in region 3 as described in the Previous Section; V_{out2} generates a larger initial spike than the 150mV case, and has additional minor stochastic fluctuations.

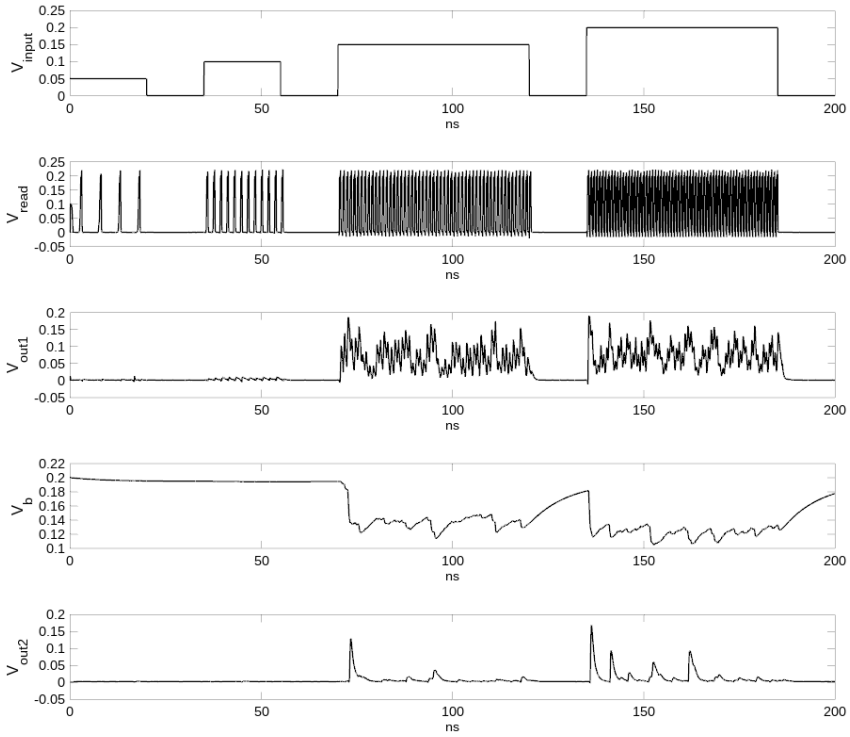


Figure 27: Operational waveforms of the Subthreshold Spintronic Stochastic Spiking neuron.

In order to analyze the effects of process variations on the S4N circuit, we performed monte-carlo analysis with 50 samples for values of V_{input} ranging from 0mV to 200mV with 10mV increments for 50ns, varying the threshold voltage of each transistor with a standard deviation of 75mV and all resistances and capacitances listed in Table 7 with a standard deviation of 20%. As shown in Figure 28, the mean output voltage follows a sigmoidal behavior, which is commensurate with biological characteristics, and the behavior is maintained even in the presence of process variation, although it may be shifted and skewed to a degree. We argue that this does not constitute an issue for biologically-inspired computational paradigms since neurons of the

exact same type and similar location in the brain have similar heterogeneous sigmoidal spiking responses to inputs [98].

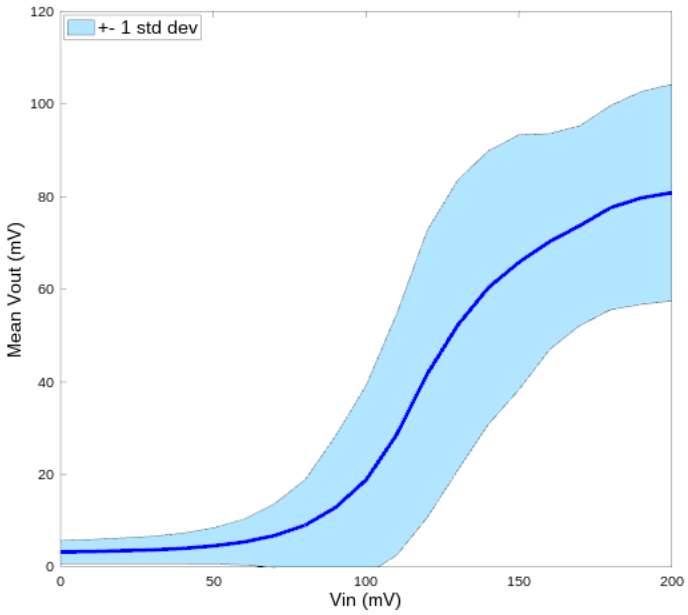


Figure 28: Mean output voltage of the S4N versus input voltage including process variation.

The S4N circuit, operating at 200mV, in the presence of process variation, uses a maximum power of just 77nW, as shown in Figure 29, which is incredibly efficient for a spiking neuron design operating at the nanosecond time-scale. Additionally, the power consumption scales in an almost sigmoidal fashion to the input voltage, using up to about an 8x reduction in power at low input voltages, which would be the most likely operating region for most S4Ns in a large network architecture. For a simple back-of-the-envelope comparison to biology, the human brain uses 100 billion Neurons at just 20W; if 100 billion S4Ns were operating with 10% using 70nW, and the rest using 10nW, then it would require 1600W of power. Although this is two orders of magnitude

larger than the human brain, the S4N operates six orders of magnitude faster. Of course, this doesn't include synapses and routing, which would use a significant fraction of the total power, but begs the question that perhaps leveraging these noisy and imprecise subthreshold circuits could be an avenue to realizing the computational efficiency of biological brains?

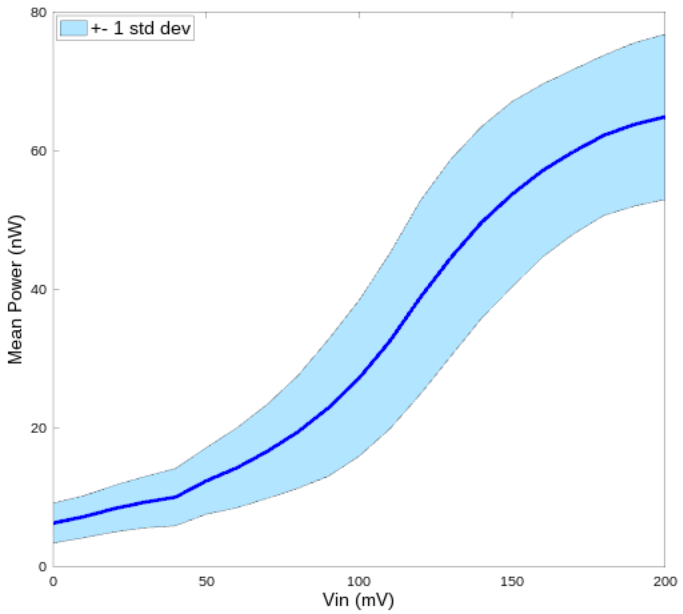


Figure 29: Mean power of the S4N versus input voltage including process variation.

Summary

The Spintronic Stochastic Spiking Neuron introduced herein was demonstrated to achieve tunable high-speed poisson-distributed spike generation within a compact hybrid spin-CMOS circuit using just 0.8-9.6uW. The circuit, when combined with a neuromorphic 2nd-order synapse,

is capable of realizing perceptron functionality such as AND and OR logic, for a readily implementable test of the computational capabilities of such a circuit. A variety of potential future works for the S3N design draw inspiration from theoretical neuroscience to focus on the realization of hardware-based online learning architectures utilizing stochastic learning algorithms.

The S4N demonstrates that circuits of noisy and imprecise components can realize biologically-inspired computational primitives at ultra-low-power. The Subthreshold Spintronic Stochastic Spiking Neuron circuit combines an Adaptive Ring Oscillator for as-needed sampling, probabilistic spintronics for thermally-driven stochasticity, and a homeostasis mechanism in order to realize biologically-inspired signals at nanosecond time scales using less than 80nW. Good behavioral robustness to process variation in line with biological observations is also demonstrated. Such a circuit could pave the way to realizing brain-like computational abilities and efficiency.

CHAPTER SIX: NEURAL SAMPLING CORE

An intriguing observation is that biological brains and nanoscale electronic circuits share characteristics that can provide insights towards designing circuits and architectures that can be utilized for biologically-inspired computation with potential power efficiency comparable to brains. The first comparable characteristic is that the primary computational structures of biological brains, neurons and synapses, are highly heterogeneous and imprecise [99, 100], which is akin to the fact that all manufactured nanodevices have behavioral variability arising from Process Variation (PV), especially CMOS devices operating at subthreshold voltages [101]. The question then arises that perhaps by designing neuromorphic circuits and architectures that can adapt to, and even utilize, such heterogeneity while trying to aggressively lower supply voltages, even greater power efficiency could be achieved compared to adhering to the strict design margins and deterministic behaviors that VLSI circuits are typically designed to realize. The second comparable characteristic is that the fundamental mechanisms underlying neural activity, ion channel opening and closing, is a stochastic process, which leads to stochasticity throughout neural activity [45]. Coincidentally, a promising framework in computational neuroscience, Neural Sampling, has theoretically proven that a particular biologically-plausible model of stochastically spiking neurons in cortical circuit motifs represent samples from an underlying conditional distribution that can be used for probabilistic inference [44, 46]. Therefore, leveraging heterogeneity and stochasticity in neuromorphic architectures using emerging devices that are

intrinsically stochastic, such as spintronics [42, 86, 102], could lead to more capable and efficient neuromorphic hardware.

The Neural Sampling Core (NSC) presented in this Chapter is motivated by the ultra-low-power and robust characteristics of biological neural networks, which utilize stochastic and heterogeneous components with local learning rules in competitive networks, The NSC is a thrust towards mimicking the underlying computational principles of the brain in nanoelectronic circuits in order to realize self-adaptive and low-power neuromorphic hardware with noisy and imprecise CMOS and spintronic devices operating at subthreshold voltages.

The following contributions are provided in this Chapter:

1. a stochastic spiking neuron circuit with protracted digital post-synaptic-potentials realizing behaviors from Neural Sampling,
2. a low-precision hybrid spintronic-CMOS synapse circuit with a new event-based *Probabilistic Hebbian Plasticity* (PHP) unsupervised learning , and
3. a novel homeostasis mechanism that regulates neural activity across multiple time-scales and process variation effects.

The above contributions are integrated into low-power neuromorphic hardware approach operating at subthreshold voltages while remaining robust to noisy and imprecise components.

Previous Work on Stochastic Spiking Neural Network Hardware with Unsupervised Learning

Several recent works have leveraged the stochastic switching properties of spintronic devices to realize unsupervised learning in SNN neuromorphic hardware as delineated in Table 8. The work developed by Zhang et al. [43] utilized multiple parallel MTJs to form a compound magnetoresistive synapse with a stochastic Spike-Timing-Dependent (STDP) learning rule in conjunction with a MTJ-based stochastic spiking neuron to realize a SNN able to achieve respectable accuracies on MNIST dataset. However, their work did not evaluate the power consumption of the design, which can be quite large for many parallel MTJs per synapse in a crossbar, nor the effect of process variation on the CMOS circuitry necessary for the neuron.

The long-term short-term stochastic synapse developed by Srinivasan et al. [57] utilizes two SHE-MTJs with distinct peripheral circuitry to realize various switching characteristics corresponding to different STDP sensitivities, enabling one SHE-MTJ to have sharper correlation sensitivity and greater synaptic strength then the other, which had moderate correlation sensitivity. They demonstrated that the scheme has faster training convergence, resulting in a reduction in total training energy consumption. However, the scheme was quite sensitive to STDP and circuit parameters, and they did not analyze the effect of process variations.

The all-spin stochastic SNN developed in [42] leverages one-bit SHE-MTJ synapses with a stochastic-STDP learning rule and SHE-MTJ based stochastic spiking neurons with a homeostasis mechanism to realize a low-energy SNN with online learning. However, the SHE-MTJ neuron requires write-read-reset cycling, which adds additional timing and energy overheads, the stochastic-STDP learning rule requires precision between the spike timing, switching

probability, and write current, the homeostasis mechanism is rather coarse since it simply cuts off neurons that reach a certain spike count during learning, and the effect of process variations are not analyzed.

Table 7: Comparison to Previous Spintronic Stochastic Spiking Neuromorphic Hardware.

Ref	Synapse Technology	Neuron Technology	Learning Rule	Homeostasis	Key Quantitative Findings
[43]	Compound MTJ	Stochastic Switching MTJ	Simplified Stochastic STDP	None	91.27% accuracy on MNIST
[57]	LT-ST SHE-MTJs	LIF	Stochastic STDP	None	10.4 uJ to train network on MNIST
[42]	SHE-MTJ	Stochastic Switching SHE-MTJ	Stochastic STDP	Spike count cutoff	682 nW per neuron
Herein	*Spin-CMOS*	*Embedded p-bit with PSP*	*PHP*	*Adaptive to fast and slow time-scales*	*311 nw per neuron 1.9-7.7 nw per synapse*

Thus, the NSC developed herein extends beyond these promising works by developing a robust subthreshold stochastic SNN approach utilizing a 3-bit hybrid spin-CMOS synapse with series and parallel SHE-MTJs, a flexible and adaptive homeostasis mechanism, and a stochastic spiking neuron with digital PSPs implementing neural sampling and enabling a simple and robust event-driven unsupervised learning mechanism, all developed and analyzed with the effect of process variations in both the spintronic and CMOS devices.

Neural Sampling Theory

This Section details the Neural Sampling Theory from computational Neuroscience [44, 46]. Neural Sampling interprets the stochastic spiking behavior of biological neurons as stochastic samples of underlying conditional distributions [44, 46]. In particular, it models the spiking behavior of neurons with an instantaneous stochastic spiking rate exponentially dependent upon the membrane potential, combined with a refractory period of duration τ and a commensurately prolonged rectangular Post-Synaptic-Potential (PSP), which approximates the PSPs found in-vivo. Combined with a Hebbian learning rule, such a model can be shown to realize a generative model of the input distribution [46]. This is in contrast to typical Leaky-Integrate and Fire (LIF) spiking neuron models, which models spikes as impulses and neurons as a leaky integration of synaptically-weighted pre-synaptic spikes that fires if a threshold is reached and then reset. For the rest of the Chapter, a 'spike' means a rectangular pulse of τ clocks, as in Neural Sampling. Several cortically-inspired circuit motifs have been developed utilizing Neural Sampling that have demonstrated impressive results of unsupervised, and reward-based learning [46, 103]. Thus, Neural Sampling provides a theoretically-accomplished and biologically-relevant framework for leveraging stochastic neural models to achieve brain-like computations

Circuits of the Neural Sampling Core

This Section delineates the constituent circuits of the NSC, such as the stochastically spiking neuron with a refractory period and prolonged digital PSPs congruent to those utilized in Neural Sampling's theoretical modeling, a three-bit synapse with event-driven probabilistic

Hebbian learning rules, and a novel homeostasis mechanism. Since an important premise of this work is that the NSC should be able to adapt and utilize the heterogeneity of components that emerges from PV, we model PV in both the spintronic and CMOS devices as described in the next Section on the simulation framework at all stages of development and analysis. This Section is organized by first detailing the operational principles of each circuit and then integrating them into a cohesive mixed-signal architecture with discussions. Although detailed later, it is worth mentioning here that there are two reciprocating phases based on the state of the clock; the read-phase occurs when the clock is low, and the update-phase occurs when the clock is high.

Stochastic Spiking Neuron with Digital Post-Synaptic Potentials

The Stochastic Spiking Neuron circuit shown in Figure 30a consists of an embedded p-bit and a digital PSP circuit that operates as follows. Based on the voltage applied to IN and the state of the stochastically switching MTJ in the embedded p-bit, $p - bit_{OUT}$ will either be high or low. If $p - bit_{OUT}$ is high at the positive edge of CLK, then the output of the PSP circuit, $Neuron_{OUT}$, will go high and hold it for eight clocks, which corresponds to a τ of eight clocks. The waveforms shown in Figure 30b is an illustrative snapshot that shows the relevant circuit signals obtained from SPICE simulations for the parameters given in the following Section on the Simulation Framework.

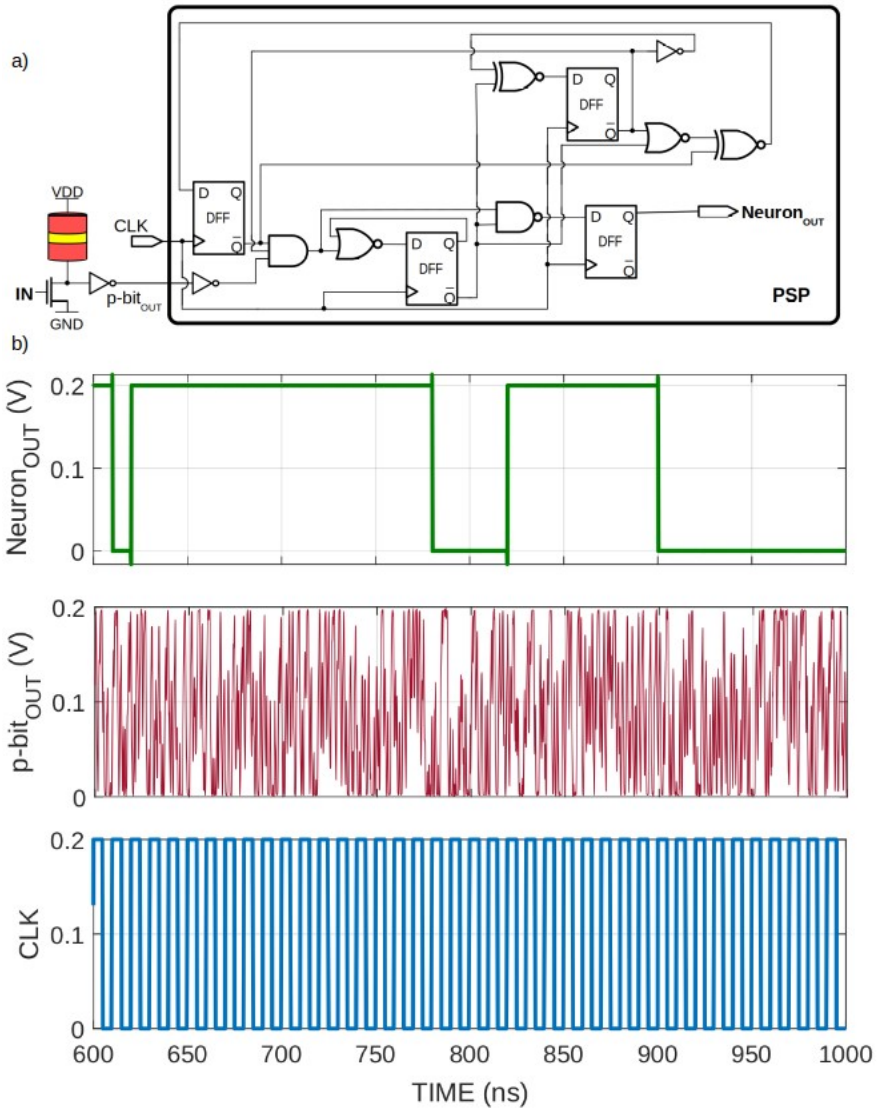


Figure 30: The stochastic spiking neuron circuit and associated waveforms. © 2019 IEEE.

Hybrid Synapse with Probabilistic Hebbian Plasticity

The hybrid spintronic-CMOS synapse shown in Figure 31 and the PHP learning rule were co-designed to take advantage of the prolonged PSP signals with the stochastic switching behavior

of spintronic devices. The synapse uses three SHE-MTJs (S1-S3) to store the synaptic weight, one PMOS transistor (M1) that operates as a voltage-controlled current source since the circuit is at subthreshold, and two NMOS transistors (M2-M3) that are used when updating the synapse. The circuit operates during the read phase as follows. If the pre-synaptic neuron is not active, which means it has not spiked within the previous τ clocks, then $\overline{IN}$ will be at VDD, N will be at VDD, and no current will flow through M1 onto SUM. If the pre-synaptic neuron has spiked within the previous τ clocks, then $\overline{IN}$ will be at GND, causing a voltage-divider between S1 and S2-S3, which determines the voltage at N, which then controls the current through M1 into SUM. The synaptic weights determined by the P or AP states of S1-S3 are shown in Table 9 where W0<W1<W2~W3<W4<W5.

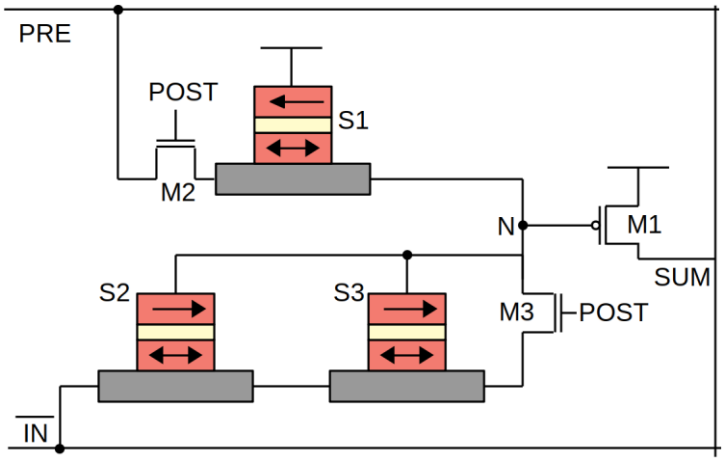


Figure 31: The three-bit hybrid spin-CMOS synapse. © 2019 IEEE.

Table 8: Synapse Weights for MTJ States. © 2019 IEEE.

S1	S2	S3	Weight
P	AP	AP	W0
P	AP	P	W1
P	P	AP	W1
P	P	P	W2
AP	AP	AP	W3
AP	AP	P	W4
AP	P	AP	W4
AP	P	P	W5

PHP modifies the synapses during the update phase in an event-driven fashion as follows. If the post-synaptic neuron, POST, has spiked during the previous τ clocks, then both M2 and M3 are turned on, allowing current to flow through the write paths of S1-S3 based on the voltages applied to PRE and $\overline{IN}$. If the pre-synaptic neuron has spiked within the previous τ clocks as well, then the synapse will update according to a *synaptic potentiation* event, that is, different voltages will be applied to PRE and $\overline{IN}$ for a given pulse duration such that S1 has a probability of switching to its anti-parallel state and S2-S3 have a probability of switching to their parallel states, which all have the effect of lowering the voltage at N and increasing the current through M1 during the read-phase. If the pre-synaptic neuron has not spiked within the previous τ clocks, then the synapse will update according to a *synaptic depression* event, that is, voltages will be applied to PRE and $\overline{IN}$ for a given pulse duration such that S1 has a probability of switching to its parallel state and S2-S3 have a probability of switching to their anti-parallel states, which all have the effect of

increasing the voltage at N and decreasing the current through M1 during the read-phase. Therefore, each time a post-synaptic neuron spikes, all associated synapses are probabilistically updated for τ clocks, and more coincident pre-synaptic spiking will have a higher chance of strengthening the synapse, while non-spiking pre-synaptic neurons will have a chance of being depressed.

Non-Volatile Homeostasis Mechanism

The homeostasis mechanism acts to increase the activity of under-active neurons and decrease the activity of over-active neurons, is implemented with a number of the homeostatic synapses shown in Figure 32 connected to the input of each neuron. The two homeostatic synapse designs shown in Figure 32 utilize alternative mechanisms for implementing homeostasis on both fast and slow time-scales, where S1 has a higher probability of switching compared to S2, and therefore adapts on a faster time-scale. The positive-feedback effect of synaptic plasticity needs a fast homeostasis mechanism to balance network activity [104, 105] while a slower homeostasis mechanism is beneficial for balancing the neuron's excitability in the presence of its intrinsic heterogeneity arising from PV. Both of the designs operate similar to the regular synapse during the read phase as follows. During the read phase $\overline{BOT}$ is pulled to GND, causing a voltage divider between S1 and S2, which determines the voltage at N, which then determines the current through M1 into SUM. The weight values are akin to the regular synapses described previously in that if S1 is AP and/or S2 is P, then the homeostatic synapse has a higher effective weight than vice-

versa. The two designs differ during the update phase as follows. The circuit in 31a requires S1 to have a lower Δ than S2, which causes it to have a higher probability of switching for the same current and pulse duration. The circuit in Figure 32b does not require S1 and S2 to have different Δs, but requires more overhead with an additional NMOS and two horizontal wires to isolate the two devices during the update phase, allowing different voltages and/or pulse durations to switch the two devices with different probabilities such that S1 switches with a higher probability than S2. During the update phase, UPDATE goes high and different voltages are applied to TOP and $\overline{BOT}$ for Figure 32a, or TOP, $\overline{TOP}$, BOT, and $\overline{BOT}$ for Figure 32b, depending on the state of the connected neuron - if the neuron is active, then a *homeostatic depression* event occurs, and if the neuron is inactive, then a *homeostatic potentiation* event occurs.

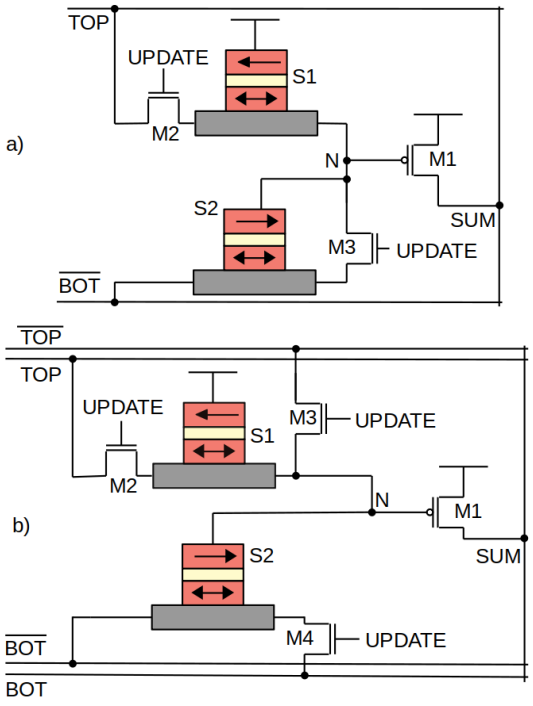


Figure 32: Two alternative implementations for the homeostatic synapse. © 2019 IEEE.

Inhibition Mechanism

Inhibitory feedback is a mechanism to ensure that only a small number of output neurons are active at a time by decreasing the input strength of the others, and therefore their chances of spiking, each time one has spiked. This enforces competition between the neurons, which enforces selectivity [106]. Without it, it is likely for all neurons to become receptive to all input patterns, and therefore there is no information from the network that can be used to discern the input patterns from one another, which is key for unsupervised learning and probabilistic inference [106]. The exact inhibitory mechanisms that the brain utilizes is still an active area of research, but many SNN models utilize a fixed inhibition model such that every time a neuron spikes, a fixed decrease in input strength is applied to all other neurons [41, 46], and the same is used for the NSC. In order to minimize area overhead, the inhibition mechanism is implemented with a single NMOS connected to the SUM wire and GND. The input voltage to that NMOS is chosen such that the effect on SUM is equivalent to the negative of the strongest synaptic weight, W5, and its associated distribution according to PV.

Architectural Discussion

Figure 33 shows all of the core components of the NSC integrated into a single layer feed-forward SNN. During the read phase, which is when CLK is low, if POST is also low, and therefore M_{read} is on, all of the synapses with spiked pre-synaptic neurons, all of the homeostatic synapses, and all of the inhibitory feedback with active POST signals will source and sink current, generating a voltage at SUM due to the resistance of R_{read} and M_{read}, which is the resulting parallel analog

computation of weighted pre-synaptic spikes plus the cumulative effect of the homeostatic synapses minus any active inhibition, and is applied to the input of the stochastic spiking neuron circuit. When CLK goes high, the post-synaptic neuron may or may not have spiked, M_{read} turns off to prevent wasted current flow, all inhibitory feedback turns off for the same reason, and the synapse and homeostatic update mechanisms occur according to Algorithm 1.

Algorithm 1: Update Phase

Input neuron state: $\mathbf{y}(t) \in [0, 1]$
Output neuron state: $\mathbf{z}(t) \in [0, 1]$
for $z_i(t)$ ***in*** $\boldsymbol{z}(t)$ **do**
    // Update homeostatic synapses
    if $z_i(t) == 1$ **then**
        homeostatic-depression(i);
    else
        homeostatic-potentiation(i);
    end
    // Update input-output synapses
    if $z_i(t) == 1$ **then**
        for $y_j(t)$ ***in*** $\boldsymbol{y}(t)$ **do**
            if $y_j(t) == 1$ **then**
                synaptic-potentiation(i,j);
            else
                synaptic-depression(i,j);
            end
        end
    end
end

The event-based nature of the NSC with its non-volatile parameters affords flexibility to its operational and greater architectural needs. For instance, the NSC is described herein with two phases corresponding to different states of the clock for simplicity, but in principle, many other

operations could intermix with the two main phases, such as routing algorithms for intra- and inter-chip communications or monitoring processes. Additionally, the clock rate could be adjusted based on application needs, using a slower clock when idle and a faster clock as needed. The clock rate could also be adjusted based on as-manufactured timing considerations. Another beneficial aspect of the non-volatile nature of the NSC is that the more power-intensive update phases are only required during training and/or re-training. Once a desired capability is achieved, the update phases can be much more dispersed or stopped altogether, saving considerable power.

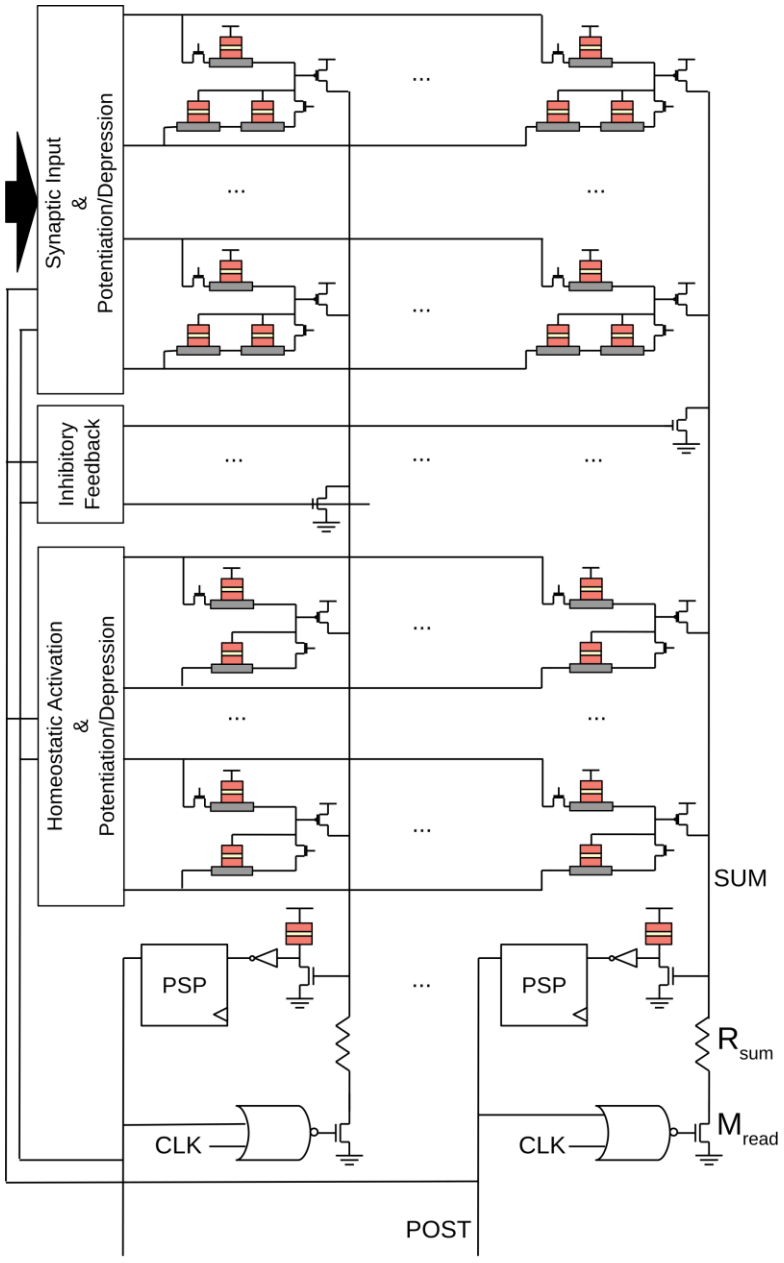


Figure 33: An architectural overview of the Neural Sampling Core. © 2019 IEEE.

Another benefit of the NSC is that it is able to learn patterns of different dimensions, as is described in the following Section, using all the same constituent circuits and devices with just an

alteration of the number of homeostatic synapses - smaller dimensional inputs require more homeostatic synapses. Therefore, fixed NSC networks of a certain size could be fabricated and then inputs and homeostatic synapses could be turned on or off depending on the application needs. Also, this could provide redundancy in the case of unusable components, providing a higher potential yield.

Simulation Framework

This section delineates the SPICE simulation parameters and results for the stochastic spiking neuron circuit, synapse circuit, and homeostatic synapse circuits, as well as the architectural simulation results from Brian2, a Spiking Neural Network Simulator [107]. All MOSFET models used 7nm high performance PTM FinFET models [67] with threshold voltages modified by a Gaussian distribution, $\mathrm{N}(0mv, 75mv)$ where $\mathrm{N}(\mu, \sigma)$ is a Gaussian distributed random variable with mean μ and standard deviation σ, to model effects of Process Variation (PV). All of the resistances of Magnetic Tunnel Junctions (MTJ) were modified from their ideal values with a Gaussian distribution with a mean of the ideal value and a standard deviation of 20% to model PV effects. The supply voltage was set to 200mV.

Stochastic Spiking Neuron Circuit Simulation Results

The stochastic spiking neuron as shown in Figure 30 was first modeled by using SPICE simulations to obtain the spiking probabilities for all input voltages and then that behavior was

modeled in Brian2 simulations. The low-energy barrier MTJ can be modeled by the stochastic Landau-Lifshitz-Gilbert (s-LLG) equation below [65].

$$\frac{(1+\alpha^2)d\hat{m}}{dt} = |\gamma|\hat{m} \times \vec{H} - \alpha|\gamma|\left(\hat{m} \times \hat{m} \times \vec{H}\right) + 1/qN\left(\hat{m} \times \overrightarrow{I_S} \times \hat{m}\right) + (\alpha/qN(\hat{m} \times \overrightarrow{I_S})) \quad (1)$$

Where α is the damping coefficient of the nanomagnet, γ is the electron gyromagnetic ratio, q is the electron charge, and $\overrightarrow{I_S}$ is the spin current applied to the free layer. The spin current's polarization, P, is equivalent to the polarization of the fixed layer, which is $\hat{z}$, and its amplitude is given by $\overrightarrow{I_S} = PI_C\hat{z}$, where I_C is the charge current flowing through the MTJ. N is the number of spins in the free layer, which is given by $N = M_s Vol.\mu_B$, where M_s is the saturation magnetization, μ_B is the Bohr magneton, and Vol. is the volume of the nanomagnet. The effective field for the monodomain circular magnet used for the free layer is $\vec{H} = -\frac{4}{\pi}M_s m_x\hat{x} + \overrightarrow{H_n}$, where $\hat{x}$ is the out-of-plane direction of the magnet and $\overrightarrow{H_n}$ is the thermal noise field in three directions: $\left(H_n^{x,y,z}\right)^2 = 2\alpha kT/(|\gamma|M_s Vol.)$. However, simulating the s-LLG equation in SPICE requires a significant amount of time. Therefore, we utilized a compact Verilog-A model for simulation speed where a resistor was modeled that stochastically switched from $0.5M\Omega$ to $1.5M\Omega$ with a retention time of $\mathrm{N}(0.5ns, 0.5ns)$, a transition time of $\mathrm{N}(0.1ns, 0.05ns)$, and a minimum retention and transition time of 0.01ns. This provided behavior that was qualitatively similar to the results provided by the s-LLG and described in [65]. The embedded p-bit with the compact stochastic MTJ model was connected to a D-Flip-Flop to estimate the probability of spiking at the clock edge

and was simulated for 100 Monte-Carlo runs with a clock period of 10ns for input voltages ranging from 0mV to 200mV with steps of 1mV for 1000ns each, and the resulting probability of spiking for each run is shown in Figure 34a. Based on this result, we modeled the instantaneous spiking probability, $\rho(t)$, of each neuron in Brian2 with equation 4.

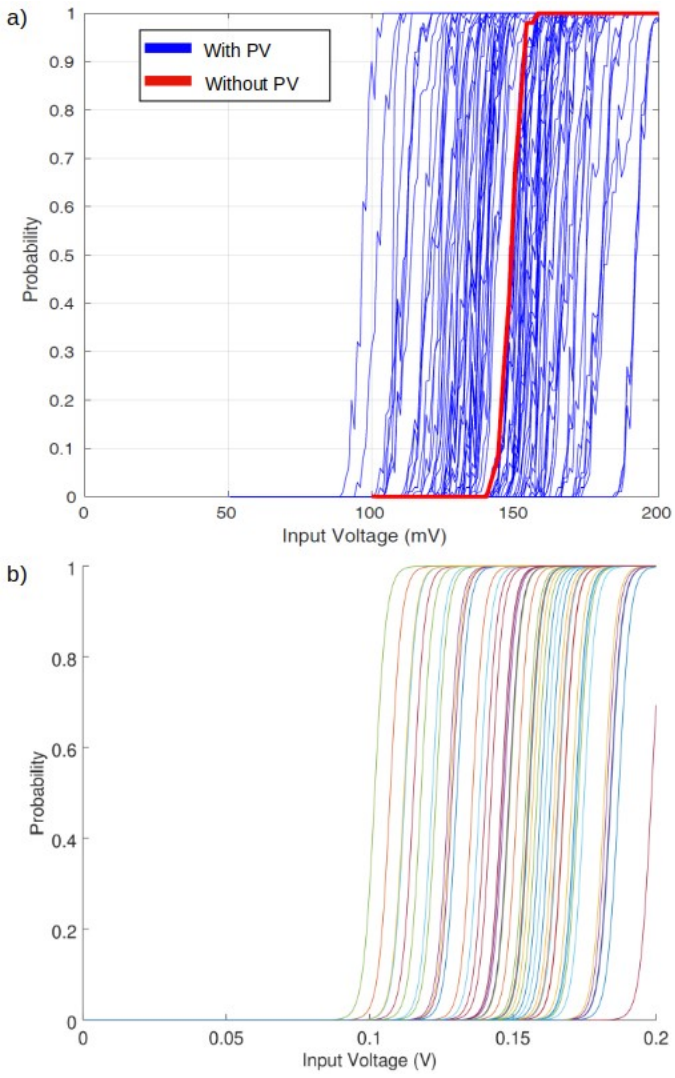


Figure 34: SPICE results and modeled sigmoids for simulation framework. © 2019 IEEE.

$$\rho(t) = \frac{1}{1+e^{-\alpha v(t)+\beta}} \qquad (2)$$

Where $\alpha = 500$, $\beta = N(75,9.75)$, and $v(t)$ is the input voltage at time t. Figure 34b shows 50 samples of equation 4 used for neurons in Brian2, which is very close to the behavior obtained from SPICE simulations.

Synapse Simulation Results

Modeling the hybrid spin-CMOS synapse and homeostatic synapse circuits in Brian2 is challenging due to the complexity of CMOS behavior, especially in the presence of process variations at subthreshold voltages. Our approach is to use 10000 monte carlo SPICE simulations for each possible synapse strength, W0-W5, for the input-output synapses and W0-W3 for the homeostatic synapses, to fit the voltage increase seen at SUM in Figure 33 of the main paper, V_{weight}, to a gamma distribution with shape parameter a and scale parameter b, and then model the synaptic strength in Brian2 with such a distribution. We used $R_{SUM} = 200k\Omega$, $R_P = N(20M\Omega, 4\text{M}\Omega)$, which is the resistance of the parallel state of the SHE-MTJs, and $R_{AP} = N(50M\Omega, 10\text{M}\Omega)$, which is the resistance of the anti-parallel state of the SHE-MTJs. The resulting fitted gamma distribution parameters for the synapse and homeostatic synapses are listed in Tables 10 and 11. The synaptic weights from the SPICE simulations as well as 10000 samples from a gamma distribution of the fitted parameters for each weight are shown in Figure 35, demonstrating conformity between the SPICE simulations and the modeled weights. The homeostatic synapse weights had a similar conformity, and which are not shown for brevity.

Table 9: Synapse Fitting Parameters. © 2019 IEEE.

Weight	Gamma Parameters	
	a	b
W0	1. 8496	1.50E-4
W1	1. 8018	2.60E-4
W2	1. 7275	4.03E-4
W3	1. 8340	4.17E-4
W4	1. 8008	9.19E-4
W5	1. 7715	1.772E-3

Table 10: Homeostatic Synapse Fitting Parameters. © 2019 IEEE.

Weight			Gamma Parameters	
	S1	S2	a	b
W0	P	AP	1. 8311	1.95E-4
W1	P	P	1. 8213	3.83E-4
W2	AP	AP	1. 8320	3.84E-4
W3	AP	P	1. 8232	1.181E-3

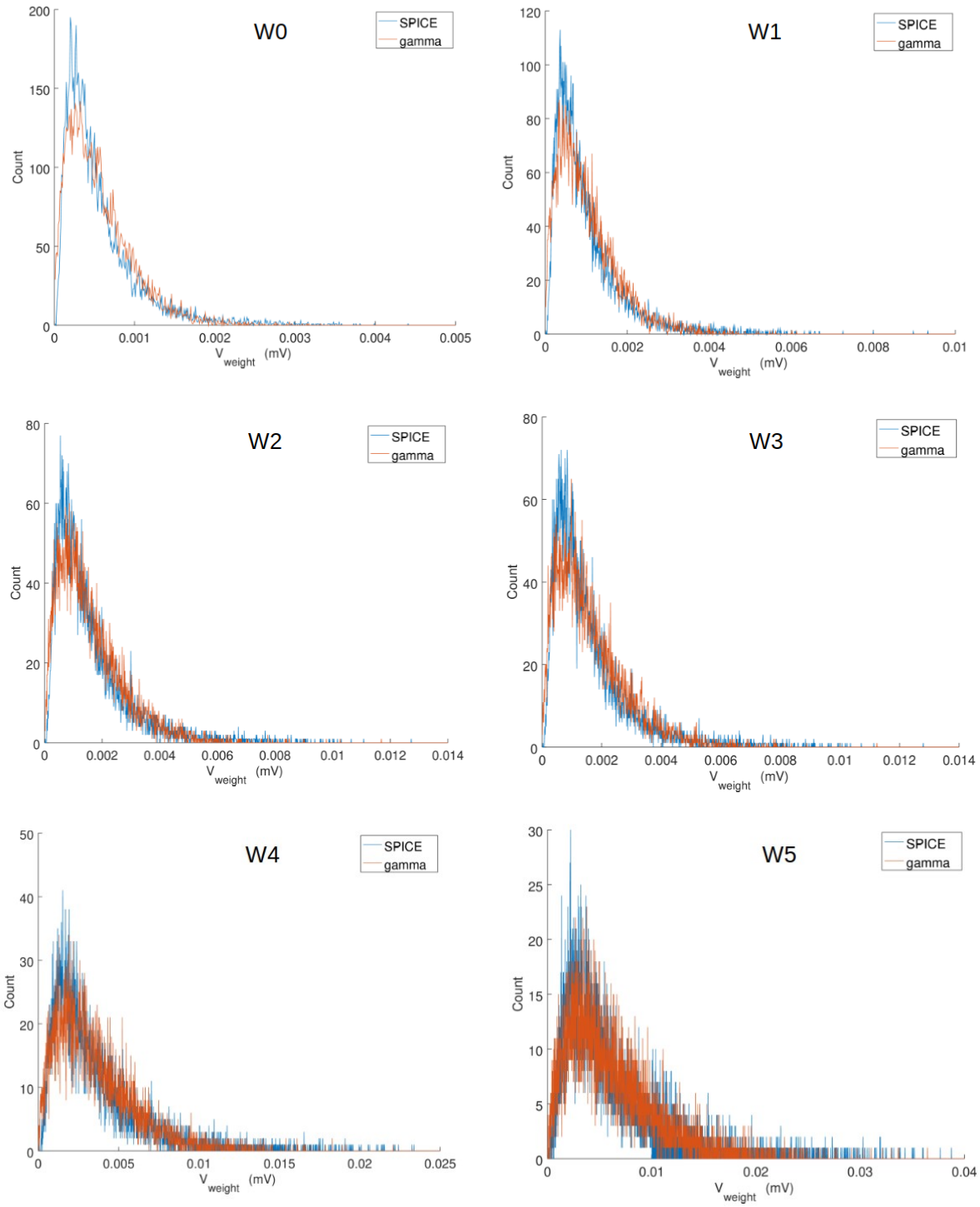


Figure 35: Synapse weight distributions for SPICE simulations and fitted gamma parameters.

Update Phase

The primary motivations and contributions of this work is the investigation of using imprecise and stochastic components to realize robust neuromorphic hardware that has very low operational power. Therefore, since the stochastic switching behavior of spintronic devices is well established [108], and yet, highly dependent upon the specific device parameters and switching mechanism, for which there is no standard SHE-MTJ foundry process yet, determining the exact voltages and pulse durations needed for a given probability of switching will differ from any assumptions that could be made herein. Therefore, we model the update mechanisms outlined in Algorithm 1 with a Gaussian-distributed probability of switching listed in Table 12 for each of the SHE-MTJs in the synapses, S1-S3, or the homeostatic synapses, S1-S2. Although the unsupervised learning results herein are obtained using the parameters listed in Table 12, we also explored switching probabilities with standard deviations of up to 500% and found no qualitative differences in the results, illustrating that the exact switching probabilities are not important as long as the average behavior is similar to those in Table 12, and therefore, such a scheme should be robust to process variations. It is also worth noting that the probability of switching for each SHE-MTJ is very small, and thus, would yield a very low-power update mechanism compared to approaches that would require larger switching probabilities. Also, if the update mechanism required too much power to update the entire NSC in parallel, time-multiplexing could be used to update smaller portions at a time, thanks to the non-volatility of the design.

Table 11: SHE-MTJ Switching Probability During Events. © 2019 IEEE.

Event	SHE-MTJ	P_{SW}
Synaptic Potentiation	S1-S3	$\mathrm{N}(0.01,0.0025)$
Synaptic Depression	S1-S3	$\mathrm{N}(0.001,0.00025)$
Homeostatic Potentiation		
	S1	$\mathrm{N}(0.0001,0.000025)$
	S2	$\mathrm{N}(0.00001,0.0000025)$
Homeostatic Depression		
	S1	$\mathrm{N}(0.01,0.0025)$
	S2	$\mathrm{N}(0.001,0.00025)$

Architecture Results

This Section describes the simulation results of the NSC. The circuits of the NSC were simulated and analyzed using SPICE and then modeled in Brian2, a SNN simulation framework [107], to obtain the unsupervised learning results.

Unsupervised Learning

The emergent unsupervised learning capabilities of the NSC are demonstrated by learning a cortically-inspired behavior, orientation selectivity, within a feed-forward SNN of 50 output neurons with 60 homeostatic synapses each and 900 Poisson spiking input neurons that each correspond to a pixel in a 30x30 stimulus window. The input pattern distribution consists of 180

28x2 bars centered and rotated in the stimulus window such that they cover the complete 180 degrees of rotation. The synaptic weights are initialized with S1-S3 randomly distributed and the homeostatic synapses are initialized with S1 in AP state and S2 in P state. Up to 10,000 randomly chosen samples from the input distribution are presented to the network for 100 clocks where each input neuron that the randomly chosen bar corresponds to has a Poisson spike rate of 75 spikes per 1,000 clocks and all others have a spike rate of 1 spike per 1,000 clocks. In between each sample is a brief period of 20 clocks whereby all input neurons have a spike rate of 1 spike per 1,000 clocks. Figure 36a shows the temporal evolution of a random selection of output neuron's receptive fields, that is, the strength of their 900 synapses shaped into a 30x30 window corresponding to the stimulus window, where a lighter color indicates a stronger synaptic strength, illustrating the emergent specialization of each neuron to a particular input pattern. Figure 36b illustrates the emergent orientation selectivity in another way, where all synapses were fixed and each input pattern was presented to the network for 100 clocks and the spikes of all output neurons were counted and shown for a random selection of 5 output neurons. The orientation selectivity of all 50 output neurons can be seen in Figure 36c, demonstrating that the entire range of possible orientations are well represented by the collection of output neurons. It can be seen that the spike counts closely resemble the tuning curves for simple cells in V1 cortex [109]. The NSC was also tested using a smaller stimulus window of 20x20 and bars of 18x2, which is shown in Figure 37, and the only needed change was an increase in the number of homeostatic synapses to 90.

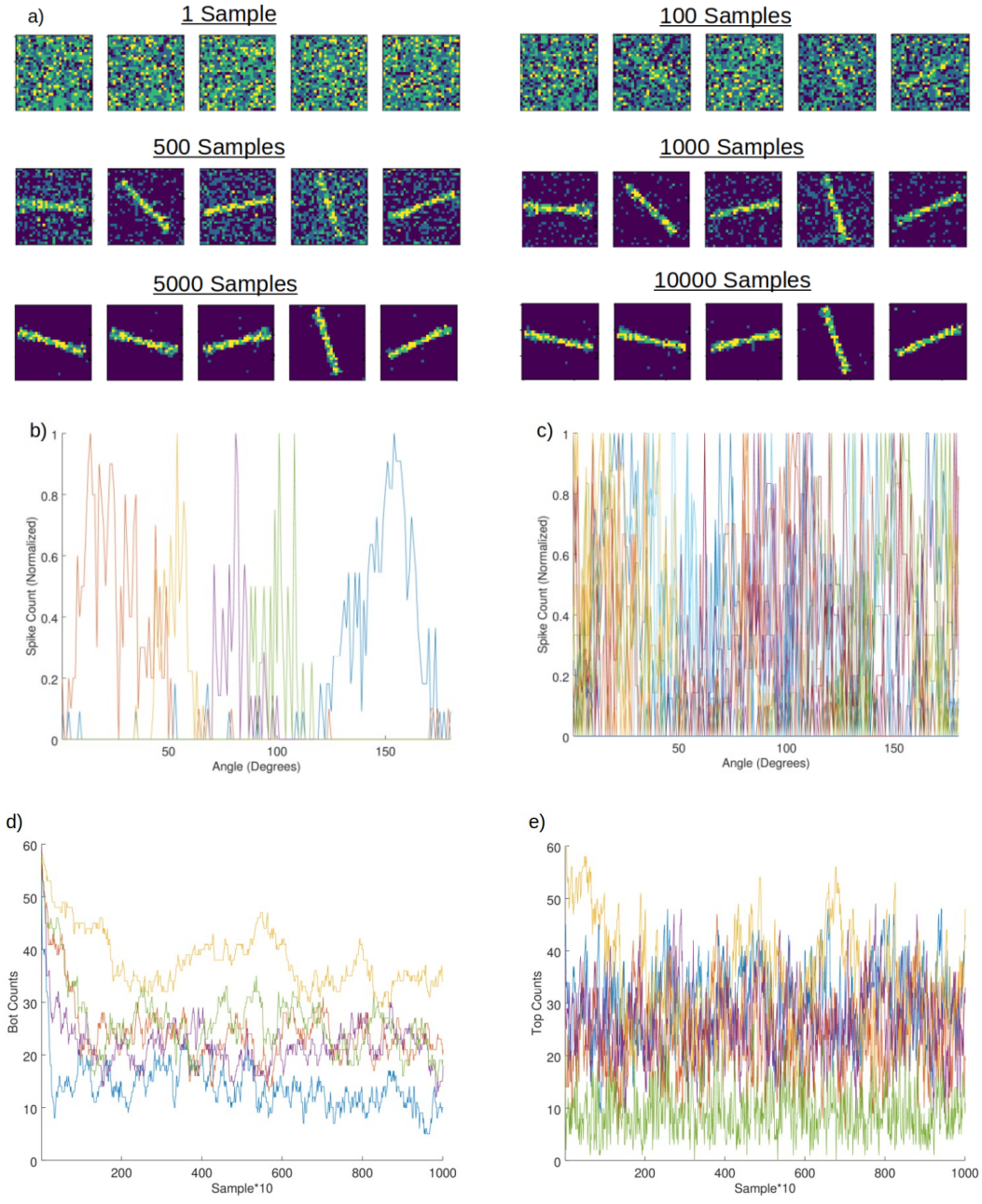


Figure 36: Relevant figures for the 30x30 test case. a) the evolution of receptive fields for a random selection of five output neurons. b) the tuning curves for a random selection five output neurons. c) the tuning curves for all 50 output neurons. d) the temporal evolution of the number of homeostatic synapses with potentiated long-term SHE-MTJs (bot) for a random selection of five output neurons. e) the temporal evolution of the number of homeostatic synapses with potentiated short-term SHE-MTJs (top) for a random selection of five output neurons.

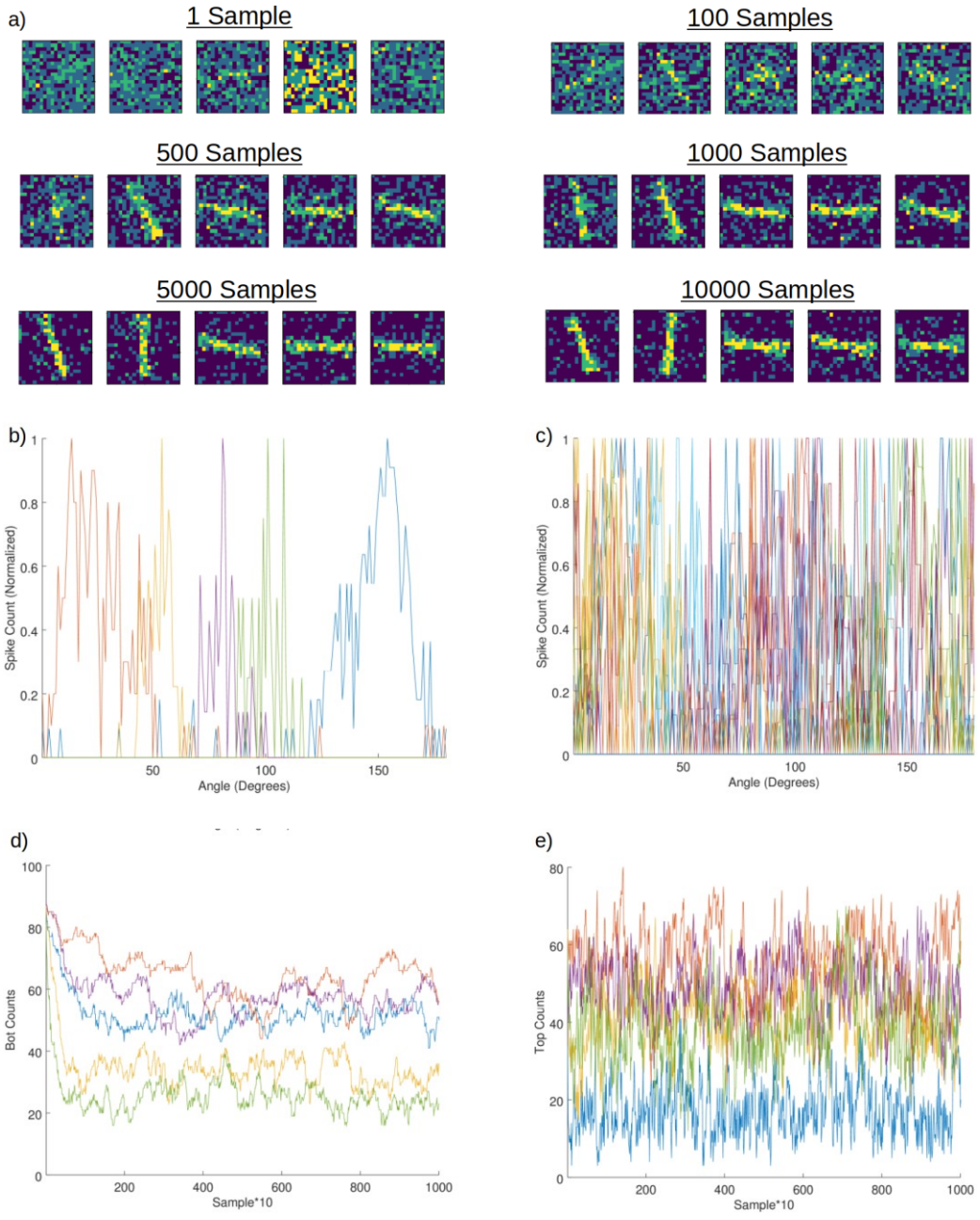


Figure 37: Relevant figures for the 20x20 test case. a) the evolution of receptive fields for a random selection of five output neurons. b) the tuning curves for a random selection five output neurons. c) the tuning curves for all 50 output neurons. d) the temporal evolution of the number of homeostatic synapses with potentiated long-term SHE-MTJs (bot) for a random selection of five output neurons. e) the temporal evolution of the number of homeostatic synapses with potentiated short-term SHE-MTJs (top) for a random selection of five output neurons.

Noise Analysis

The NSC is also quite robust to input noise and is actually able to utilize such noise for some benefit. This was tested by adding a uniformly distributed random spike rate between 0 and 7.5 spikes per 1000 clocks to each input neuron for each pixel in the stimulus window as described previously. The noise had a regulating effect, decreasing the number of homeostatic synapses required to just 30 for a 30x30 stimulus window. The NSC was still able to learn orientation selectivity with the noise, although the receptive field was qualitatively noisier and the tuning curves were on average a bit broader, as shown in Figure 38.

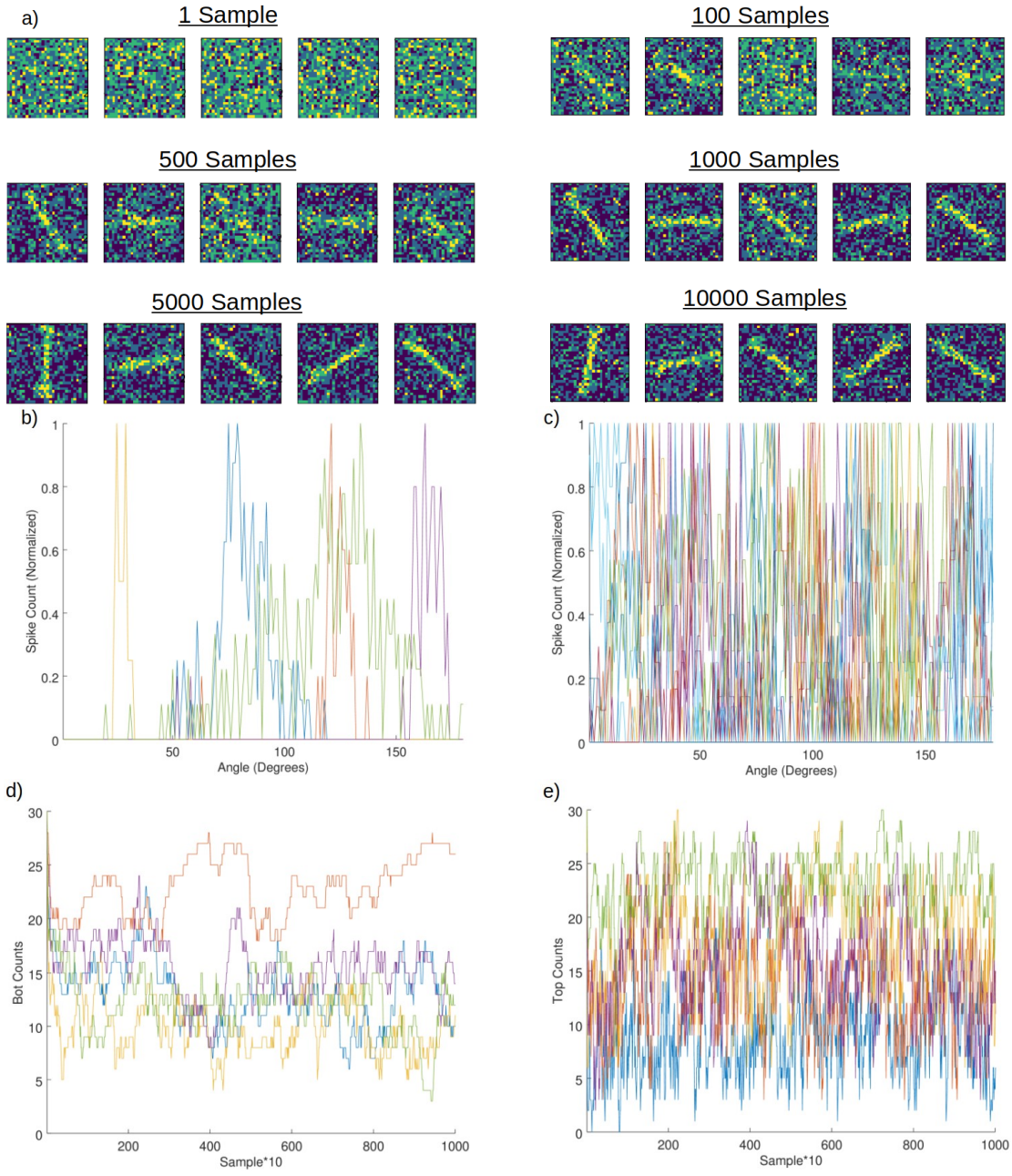


Figure 38: Relevant figures for the 30x30 test case with noise. a) the evolution of receptive fields for a random selection of five output neurons. b) the tuning curves for a random selection five output neurons. c) the tuning curves for all 50 output neurons. d) the temporal evolution of the number of homeostatic synapses with potentiated long-term SHE-MTJs (bot) for a random selection of five output neurons. e) the temporal evolution of the number of homeostatic synapses with potentiated short-term SHE-MTJs (top) for a random selection of five output neurons.

Power Analysis

The average power consumption for each of the NSC circuits was found using SPICE simulations and was determined to be 310nW for the stochastic spiking neuron with PSP circuit, 1.9-7.7nW for each of the input synapses, depending on the synaptic strength, and 1-3.4nW for each of the homeostatic synapses, depending on its strength. The average power consumption of the network during the read phase for the neurons, homeostasis mechanism, and active synapses for the 20x20, 30x30, and 30x30 with noise test cases are shown in Figure 39. The inhibitory mechanism was found to be negligible since very few output neurons are ever active at one time. As shown, the power consumption due to the output neurons are all equal since the number of neurons does not change. The power consumption due to the synapses increases from the 20x20 case to the 30x30 case since there are more inputs and synapses, and the noise increases the synaptic power consumption due to there being more active synapses as well as a higher number of higher strength synapses. The homeostasis power consumption is highest for the 20x20 case since it has the fewest active input synapses, and therefore needs on average more homeostatic input synapses to drive the neurons to spike and is lowest for the 30x30 with noise for the exact opposite reason. The power consumption of the update phase is not considered due to the NSC requiring updates only during training or re-training, and thus, is a very small fraction of the total lifetime energy usage. Additionally, the power consumption of the update phase depends heavily upon the materials, dimensions, and technology of the devices used.

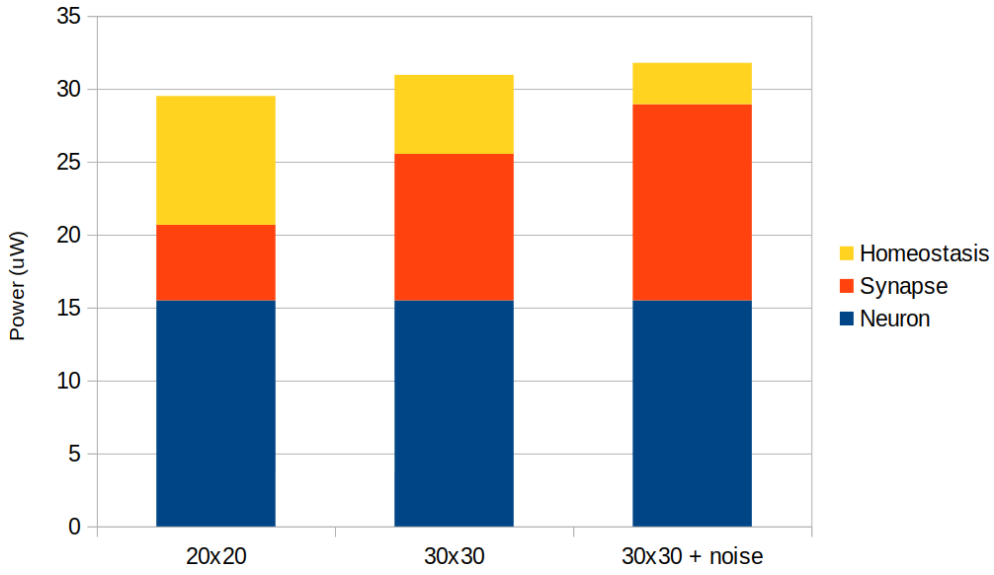


Figure 39: Average power consumption for each component of the NSC for each test case.

Summary

The NSC described in this Chapter provides several intriguing insights to realizing ultra-low-power neuromorphic circuits and architectures. Future directions for extending the NSC could be to implement recurrent connections and migrate the inhibitory mechanism to a population of inhibitory neurons, which would more closely resemble cortical network motifs, to explore how networks of NSCs could be connected together in deep or hierarchical fashions to realize greater computational ability, or to develop methodologies that can implement supervised or reinforcement learning capabilities.

CHAPTER 7: CONCLUSION

This Chapter presents a summary of the circuits and techniques that were developed in this dissertation, a selection of the limitations of the approaches with lessons learned and avenues for improvements, and finally, a discussion about future works in the directions laid out herein.

Summary of the Developed Circuits and Techniques

In order to continue attaining improvements in processor technology when Moore's Law for transistor scaling ends, the novel properties of new nanoelectronics devices must be effectively utilized for improving the capabilities of contemporary processor designs while exploring entirely new computational paradigms. Spintronic devices are well situated to achieving these goals with utilizing beneficial properties of scalability, non-volatility, and vertical integration for reducing the area of standard digital circuits while enabling those circuits to be powered off when idle without any loss of data, greatly reducing static power consumption, which is an ever-growing portion of total power consumption for highly scaled nanoelectronics circuits. Additionally, the intrinsic stochastic behaviors of spintronic devices, which is typically seen as a challenge for standard digital designs, enables whole new classes of neuromorphic paradigms that can yield compact and ultra-low-power neuronally-inspired computations. In this dissertation, the intrinsic switching properties of some promising spintronic devices have been leveraged to demonstrate these benefits.

Complementary Switching in Hybrid Spin-CMOS Digital Circuits

The hybrid spin-CMOS digital circuits developed in Chapter 3 utilized the complementary switching properties of the DWCSTT device as a voltage divider that can directly interface with CMOS without the need for sense amplifiers. This property was utilized to imbue critical components of synchronous and asynchronous datapaths with non-volatility, enabling the circuits to be power-gated without requiring storing or restoring circuitry, signals, or timing overheads, while also reducing the number of devices needed to realize the circuit. In particular, a non-volatile D F/F was developed that required 10 fewer transistors than traditional CMOS only designs and significantly fewer than that compared to alternative non-volatile D F/F circuits that required store and restore overheads. It was also shown that the width of the transistors driving the DWCSTT device could be adjusted to tune the design between speed, power, and area demands. Additionally, a non-volatile Muller C-Element was developed that used just 8 transistors with 1 DWCSTT device, which is equivalent in transistor count to traditional CMOS-only Muller C-element circuits, but the non-volatility of the developed circuit was shown to enable entire asynchronous pipelines to be power gated without loss of data or operational functionality when power was restored. The developed non-volatile Muller C-element was also shown to be faster, lower power, and had a lower device count than alternative non-volatile Muller C-element designs. It was also explored how improvements to spintronic manufacturing technology improving TMR would benefit the operational characteristics of the circuit in addition to trading off power, delay, area, and energy by adapting the driving transistor widths. Finally, the slower switching speed of the non-volatile Muller C-element when compared with pure CMOS implementations was found to

be a benefit when utilized in bundled data asynchronous protocols since it negated the need for inserting delay elements between pipeline states.

Stochastic Switching for Neuromorphic Circuits

Neuromorphic paradigms that combine large arrays of synapses with non-linear neurons are a natural fit for the dense non-volatile properties of spintronic device, and many previous works have explore this proposition. This dissertation focused on leveraging the stochastic properties of spintronic devices to improve the capabilities of contemporary machine learning approaches utilizing DNN architectures while enabling new implementations at ultra-low-power that are in accordance with some of the latest theories of brain computation from computational neuroscience. Both stochastic neurons using p-bit circuits as well as hybrid spin-CMOS synapses using the stochastic switching of non-volatile SHE-driven spintronic devices are explored

In Chapter 4, the SHE-driven p-bit is used to realize a compact and low-power PAF for computing the non-linear function of weighted summation required for BNNs in resistive crossbar and pseudo-crossbar circuits. Due to the very low current operation of the SHE-driven p-bit, highly resistive devices can be used in the crossbar network, which leads to a low power consumption of just 75 nW per active synapse with 4.98 uW per PAF. Additional exploration of error rate under the effects of process variations for on-chip vs off-chip training showed that on-chip training almost completely mitigated the reduction in accuracy induced by variations in the resistances of the resistive devices in the crossbar network.

In Chapter 5, two approaches for leveraging p-bits to implement stochastically spiking neurons that have alternative benefits are presented. The first utilized the SHE-driven p-bit to realize a high-speed spiking neuron with a tunable spike rate based on input current using just 0.6-9.8 uW of power. It was also demonstrated that such a neuron with a neuromorphic synapse implementing PSPs could implement perceptron functionality. The second circuit preferred to eschew size and speed optimizations in order to operate at as low a voltage as possible, and thus, low power, demonstrating that combining the embedded p-bit with additional CMOS circuitry could realize a stochastic spiking neuron with homeostasis and PSPs resembling biological behaviors using a maximum of just 77 nW. It was also demonstrated that process variations will not cause the circuit to fail, but simply induces heterogeneity into the circuits behavior that is similar to the biological heterogeneity found in neurons of the same type and region of the cortex.

In Chapter 6, a wholistic circuit, architecture, and algorithm co-design utilizing imprecise and stochastic components at subthreshold voltages and under the effects of process variations is developed to realize brain-inspired computational abilities at ultra-low-power. A stochastic spiking neuron utilizing the embedded p-bit was developed in combination with a digital PSP circuit to implement the neuronal behavior described by Neural Sampling, a powerful theoretical framework from computational neuroscience. A low-precision hybrid spin-CMOS synapse was designed to leverage the stochastic switching properties of SHE-driven MTJ devices for realizing the new *Probabilistic Hebbian Plasticity* learning rule along with the voltage-driven current-source properties of subthreshold CMOS for ultra-low-power. A novel non-volatile homeostasis mechanism is also developed that utilizes the stochastic switching properties of SHE-driven MTJ

devices for balancing spiking activity across multiple time-scales. All components were modeled with the effects of process variations and demonstrated to achieve unsupervised learning of orientation selectivity using just 311 nW per neuron and 1.9-7.7 nW per synapse at 200mV.

Lessons Learned and Limitations

Based on the simulations and analyses conducted as a part of this dissertation, the uncovered limitations of the approaches are discussed in conjunction with recommendations for possible improvements.

Delay and energy overheads when using spintronics in digital circuits: While the hybrid spin-CMOS circuits developed in Chapter 3 imbue the digital circuits with improved area efficiency and non-volatility, they are significantly slower compared to pure CMOS approaches and require significantly more energy per computation. Therefore, with current spintronic devices, such approaches would only be beneficial in applications that afforded long periods of idle time, since the power benefits of the hybrid spin-CMOS approach come from the fact that you can power off the circuits while idle, which you cannot do with pure-CMOS approaches without storing the data to non-volatile cells, which is power intensive in itself. Future spintronic devices offer the possibility of greater energy efficiency than CMOS, and such devices would greatly improve the hybrid spin-CMOS approaches presented herein [6].

PAFs decrease accuracy in BNNs: Although Hubara et al. [40] propose that stochastic binarization is more appealing than deterministic binarization and we showed that area and power improvements can be realized using a p-bit based PAF in Chapter 4, our experiments showed no benefits to accuracy of the BNN for CIFAR-10 dataset when using stochastic binarization over the deterministic approach, regardless of the choice of architecture. However, it is yet to be seen the extent to which such PAFs help to regularize the network across more diverse datasets and prevent overfitting, akin to Dropout [110].

Asynchronous analog spiking neurons are challenging to integrate in a wholistic neuromorphic architecture: The spintronic stochastic spiking neurons developed in Chapter 5 demonstrated how to achieve high speed and ultra-low-power compact neuron circuits, which we developed as stepping stones towards developing neuromorphic architectures that achieved the same goals. However, while attempting to design synapses and associated learning algorithms and circuits, the asynchronicity and analog output of the neurons raised many challenges regarding identifying learning conditions that circuits could realize across time-scales and ensuring that the paradigms operated appropriately when considering process variations. Thus, we chose to tradeoff some speed, power, and area to realize synchronous designs in Chapter 6 that allowed for very simple learning rules, synapses, and homeostasis mechanisms that work properly at subthreshold voltages and including process variation.

Future Work

Future work regarding hybrid spin-CMOS digital circuits would be to integrate emerging voltage-controlled spintronic devices that could be far more energy efficient and faster than the DWM or SOC devices explored herein, such as the MESO device [6]. Such devices could ultimately replace far more CMOS circuitry than explored herein for even finer-grained pipelining and power-gating strategies.

Regarding neuromorphic circuits, the promising work in Chapter 6 could be extended to recurrent neural networks consisting of separate populations of excitatory and inhibitory neurons, much like the circuits of the cortex, to realize temporal pattern recognition and hierarchical structures with top-down and bottom-up integration for higher-dimensional pattern recognition. Alterations to PHP that include a reward term could be utilized for reinforcement learning applications and benchmarking against current state of the art DNN schemes could be used to determine possible benefits in training speed, operation speed, and hardware efficiency.

APPENDIX: COPYRIGHT PERMISSIONS

On Apr 11, 2019, at 8:28 AM, M.E. Brennan <me.brennan@ieee.org> wrote:

Dear Steven,

I'm including the standard response for reusing IEEE copyrighted papers in dissertations/theses. You will need to construct the best citation you can and indicate that the papers are accepted for publication. If you have any additional questions, please let me know.

S. D. Pyle, R. Zand, S. Sheikhfaal, and R. F. DeMara, "Subthreshold Spintronic Stochastic Spiking Neural Networks with Probabilistic Hebbian Plasticity and Homeostasis," accepted to IEEE Journal on Exploratory Solid-State Computational Devices and Circuits (JXCDC) for Special Issue on Nonvolatile Memory for Efficient Implementation of Neural/Neuromorphic Computing, in-press.

S. D. Pyle, J. D. Sapp, and R. F. DeMara, "Leveraging Stochasticity for In-Situ Learning in Binarized Deep Neural Networks," accepted to IEEE Computer for Special Issue on Cognitive Computing Systems and Applications, in-press.

The IEEE does not require individuals working on a dissertation/thesis to obtain a formal reuse license however, you must follow the requirements listed below:

Textual Material

Using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © [Year of publication] IEEE.

In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.

If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Kind regards,

M.E. Brennan

Title: Compact Spintronic Muller C-Element With Near-Zero Standby Energy
Author: Steven D. Pyle
Publication: Magnetics, IEEE Transactions on
Publisher: IEEE
Date: Feb. 2018

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

From: **Gendle,Eleanor** EGendle@theiet.org
Subject: Permission to reuse content from the IET
Date: April 11, 2019 at 3:22 PM
To: stevendpyle@gmail.com
Cc: Vukmirovic,Kruna KVukmirovic@theiet.org, Newland,Samantha SamanthaNewland@theiet.org

Dear Mr. Steven Pyle,

Pyle, S.D.; Li, H.; DeMara, R.F.: 'Compact low-power instant store and restore D flip-flop using a self-complementing spintronic device', Electronics Letters, 2016, 52, (14), p. 1238-1240, DOI: 10.1049/el.2015.4114
IET Digital Library, https://digital-library.theiet.org/content/journals/10.1049/el.2015.4114

Pyle, Steven D.; Camsari, Kerem Y.; DeMara, Ronald F.: 'Hybrid spin-CMOS stochastic spiking neuron for high-speed emulation of In vivo neuron dynamics', IET Computers & Digital Techniques, 2018, 12, (4), p. 122-129, DOI: 10.1049/iet-cdt.2017.0145
IET Digital Library, https://digital-library.theiet.org/content/journals/10.1049/iet-cdt.2017.0145

I have been forwarded your request to obtain permission to reuse the content of the two above items published with the IET. I am happy to tell you that as these are for your dissertation there is no permission fee and we only ask that you give the full bibliographic reference to the version of record.

Please do not hesitate to contact me if you have any further questions.

Best regards,

Eleanor Gendle
Executive Editor

The Institution of Engineering and Technology

T: +44 (0)1438 767318

Visit our website www.theiet.org
Follow us on Twitter and LinkedIn

Michael Faraday House, Six Hills Way, Stevenage, SG1 2AY, United Kingdom

REFERENCES

[1] K. Wang, J. Alzate, and P. K. Amiri, "Low-power non-volatile spintronic memory: STT-RAM and beyond," *Journal of Physics D: Applied Physics,* vol. 46, no. 7, p. 074003, 2013.

[2] S. D. Pyle, D. Fan, and R. F. Demara, "Compact Spintronic Muller C-Element With Near-Zero Standby Energy," *IEEE Transactions on Magnetics,* vol. 54, no. 2, pp. 1-7, 2018. © 2018 IEEE

[3] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proceedings of the IEEE,* vol. 106, no. 2, pp. 260-285, 2018.

[4] M. M. Waldrop, "The chips are down for Moore's law," *Nature News,* vol. 530, no. 7589, p. 144, 2016.

[5] (2015). *2015 International Technology Roadmap for Semiconductors (ITRS)*.

[6] S. Manipatruni, D. E. Nikonov, C.-C. Lin, T. A. Gosavi, H. Liu, B. Prasad, Y.-L. Huang, E. Bonturim, R. Ramesh, and I. A. Young, "Scalable energy-efficient magnetoelectric spin–orbit logic," *Nature,* vol. 565, no. 7737, p. 35, 2019.

[7] S. D. Pyle, R. Zand, S. Sheikhfaal, and R. F. DeMara, "Subthreshold Spintronic Stochastic Spiking Neural Networks with Probabilistic Hebbian Plasticity and Homeostasis," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits (JxCDC),* in-press 2019. © 2019 IEEE

[8] S. D. Pyle, H. Li, and R. F. DeMara, "Compact low-power instant store and restore D flip-flop using a self-complementing spintronic device," *IET Electonics Letters,* vol. 52, no. 14, pp. 1238-1240, 2016.

[9] 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. 082508, 2008.

[10] S. Kanai, Y. Nakatani, M. Yamanouchi, S. Ikeda, H. Sato, F. Matsukura, and H. Ohno, "Magnetization switching in a CoFeB/MgO magnetic tunnel junction by combining spin-transfer torque and electric field-effect," *Applied Physics Letters,* vol. 104, no. 21, p. 212406, 2014.

[11] S. Kanai, M. Yamanouchi, S. Ikeda, Y. Nakatani, F. Matsukura, and H. Ohno, "Electric field-induced magnetization reversal in a perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction," *Applied Physics Letters,* vol. 101, no. 12, p. 122403, 2012.

[12] D. Suzuki, M. Natsui, A. Mochizuki, S. Miura, H. Honjo, K. Kinoshita, S. Fukami, H. Sato, S. Ikeda, and T. Endoh, "Design and fabrication of a perpendicular magnetic tunnel junction based nonvolatile programmable switch achieving 40% less area using shared-control transistor structure," *Journal of applied physics,* vol. 115, no. 17, p. 17B742, 2014.

[13] S. Angizi, Z. He, Y. Bai, J. Han, M. Lin, R. F. DeMara, and D. Fan, "Leveraging Spintronic Devices for Efficient Approximate Logic and Stochastic Neural Networks," in *Proceedings of the 2018 on Great Lakes Symposium on VLSI*, 2018, pp. 397-402: ACM.

[14] S. Angizi, H. Jiang, R. F. DeMara, J. Han, and D. Fan, "Majority-Based Spin-CMOS Primitives for Approximate Computing," *IEEE Transactions on Nanotechnology,* 2018.

[15] N. Khoshavi and R. F. DeMara, "Read-Tuned STT-RAM and eDRAM Cache Hierarchies for Throughput and Energy Optimization," *IEEE Access,* 2018.

[16] A. Roohi and R. F. DeMara, "NV-Clustering: Normally-Off Computing Using Non-Volatile Datapaths," *IEEE Transactions on Computers,* 2018.

[17] A. Roohi, R. Zand, and R. F. DeMara, "A tunable majority gate-based full adder using current-induced domain wall nanomagnets," *IEEE Transactions on Magnetics,* vol. 52, no. 8, pp. 1-7, 2016.

[18] A. Roohi, R. Zand, and R. F. DeMara, "Synthesis of normally-off boolean circuits: An evolutionary optimization approach utilizing spintronic devices," in *Quality Electronic Design (ISQED), 2018 19th International Symposium on*, 2018, pp. 49-54: IEEE.

[19] S. Salehi and R. F. DeMara, "SLIM-ADC: Spin-based Logic-In-Memory Analog to Digital Converter leveraging SHE-enabled Domain Wall Motion devices," *Microelectronics Journal,* vol. 81, pp. 137-143, 2018.

[20] S. Salehi and R. F. DeMara, "BGIM: Bit-Grained Instant-on Memory Cell for Sleep Power Critical Mobile Applications," in *2018 IEEE 36th International Conference on Computer Design (ICCD)*, 2018, pp. 342-345: IEEE.

[21] S. Salehi, N. Khoshavi, and R. F. Demara, "Mitigating Process Variability for Non-Volatile Cache Resilience and Yield," *IEEE Transactions on Emerging Topics in Computing,* 2018.

[22] S. Salehi, N. Khoshavi, R. Zand, and R. F. DeMara, "Self-Organized Sub-bank SHE-MRAM-based LLC: An energy-efficient and variation-immune read and write architecture," *Integration,* 2018.

[23] S. Salehi, M. B. Mashhadi, A. Zaeemzadeh, N. Rahnavard, and R. F. DeMara, "Energy-Aware Adaptive Rate and Resolution Sampling of Spectrally Sparse Signals Leveraging VCMA-MTJ Devices," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems,* vol. 8, no. 4, pp. 679-692, 2018.

[24] R. Zand, K. Y. Camsari, S. D. Pyle, I. Ahmed, C. H. Kim, and R. F. DeMara, "Low-Energy Deep Belief Networks Using Intrinsic Sigmoidal Spintronic-based Probabilistic Neurons," in *Proceedings of the 2018 on Great Lakes Symposium on VLSI*, 2018, pp. 15-20: ACM.

[25] R. Zand, A. Roohi, and R. F. DeMara, "Fundamentals, Modeling, and Application of Magnetic Tunnel Junctions," *Nanoscale Devices: Physics, Modeling, and Their Application,* p. 337, 2018.

[26] M. D. Stiles and J. Miltat, "Spin-transfer torque and dynamics," in *Spin dynamics in confined magnetic structures III*: Springer, 2006, pp. 225-308.

[27] M. D. Stiles and A. Zangwill, "Anatomy of spin-transfer torque," *Physical Review B,* vol. 66, no. 1, p. 014407, 2002.

[28] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, and C. Fukumoto, "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM," in *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, 2005, pp. 459-462: IEEE.

[29] A. Roohi, R. Zand, and R. DeMara, "A Tunable Majority Gate based Full Adder using Current-Induced Domain Wall Nanomagnets."

[30] Y. Seo, X. Fong, and K. Roy, "Domain wall coupling-based STT-MRAM for on-chip cache applications," *IEEE Transactions on Electron Devices,* vol. 62, no. 2, pp. 554-560, 2015.

[31] M. Cubukcu, O. Boulle, M. Drouard, K. Garello, C. O. Avci, I. M. Miron, J. Langer, B. Ocker, P. Gambardella, and G. Gaudin, "Spin-orbit torque magnetization switching of a three-terminal perpendicular magnetic tunnel junction," *Applied Physics Letters,* vol. 104, no. 4, p. 042406, 2014.

[32] S. Datta, S. Salahuddin, and B. Behin-Aein, "Non-volatile spin switch for Boolean and non-Boolean logic," *Applied Physics Letters,* vol. 101, no. 25, p. 252411, 2012.

[33] K. Ando, S. Takahashi, J. Ieda, Y. Kajiwara, H. Nakayama, T. Yoshino, K. Harii, Y. Fujikawa, M. Matsuo, and S. Maekawa, "Inverse spin-Hall effect induced by spin pumping in metallic system," *Journal of Applied Physics,* vol. 109, no. 10, p. 103913, 2011.

[34] F. Czeschka, L. Dreher, M. Brandt, M. Weiler, M. Althammer, I.-M. Imort, G. Reiss, A. Thomas, W. Schoch, and W. Limmer, "Scaling behavior of the spin pumping effect in ferromagnet-platinum bilayers," *Physical review letters,* vol. 107, no. 4, p. 046601, 2011.

[35] C. Du, H. Wang, P. C. Hammel, and F. Yang, "Y3Fe5O12 spin pumping for quantitative understanding of pure spin transport and spin Hall effect in a broad range of materials," *Journal of Applied Physics,* vol. 117, no. 17, p. 172603, 2015.

[36] M. Jamali, J. S. Lee, J. S. Jeong, F. Mahfouzi, Y. Lv, Z. Zhao, B. K. Nikolić, K. A. Mkhoyan, N. Samarth, and J.-P. Wang, "Giant Spin Pumping and Inverse Spin Hall Effect in the Presence of Surface and Bulk Spin− Orbit Coupling of Topological Insulator Bi2Se3," *Nano Letters,* vol. 15, no. 10, pp. 7126-7132, 2015.

[37] A. Mellnik, J. Lee, A. Richardella, J. Grab, P. Mintun, M. H. Fischer, A. Vaezi, A. Manchon, E.-A. Kim, and N. Samarth, "Spin Transfer Torque Generated by the Topological Insulator Bi_2Se_3," *arXiv preprint arXiv:1402.1124,* 2014.

[38] H. Wang, C. Du, Y. Pu, R. Adur, P. Hammel, and F. Yang, "Large spin pumping from epitaxial Y 3 Fe 5 O 12 thin films to Pt and W layers," *Physical Review B,* vol. 88, no. 10, p. 100406, 2013.

[39] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature,* vol. 521, no. 7553, p. 436, 2015.

[40] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Advances in neural information processing systems*, 2016, pp. 4107-4115.

[41] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in computational neuroscience,* vol. 9, p. 99, 2015.

[42] G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic tunnel junction enabled all-spin stochastic spiking neural network," in *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2017, pp. 530-535: IEEE.

[43] D. Zhang, L. Zeng, Y. Zhang, W. Zhao, and J. O. Klein, "Stochastic spintronic device based synapses and spiking neurons for neuromorphic computation," in *Nanoscale*

Architectures (NANOARCH), 2016 IEEE/ACM International Symposium on, 2016, pp. 173-178: IEEE.

[44] L. Buesing, J. Bill, B. Nessler, and W. Maass, "Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons," *PLoS computational biology,* vol. 7, no. 11, p. e1002211, 2011.

[45] G. Deco, E. T. Rolls, and R. Romo, "Stochastic dynamics as a principle of brain function," *Progress in neurobiology,* vol. 88, no. 1, pp. 1-16, 2009.

[46] R. Legenstein, Z. Jonke, S. Habenschuss, and W. Maass, "A probabilistic model for learning in cortical microcircuit motifs with data-based divisive inhibition," *arXiv preprint arXiv:.05182,* 2017.

[47] K. Ryu, J. Kim, J. Jung, J. P. Kim, S. H. Kang, and S.-O. Jung, "A magnetic tunnel junction based zero standby leakage current retention flip-flop," *IEEE Transactions on Very Large Scale Integration Systems,* vol. 20, no. 11, pp. 2044-2053, 2012.

[48] D. Suzuki, N. Sakimura, M. Natsui, A. Mochizuki, T. Sugibayashi, T. Endoh, H. Ohno, and T. Hanyu, "A compact low-power nonvolatile flip-flop using domain-wall-motion-device-based single-ended structure," *IEICE Electronics Express,* vol. 11, no. 13, pp. 20140296-20140296, 2014.

[49] D. E. Muller, "A theory of asynchronous circuits," *Report,* no. 66, 1955.

[50] E. Zianbetov, E. Beigné, and G. Di Pendina, "Non-volatility for ultra-low power asynchronous circuits in hybrid cmos/magnetic technology," in *Asynchronous Circuits and Systems (ASYNC), 2015 21st IEEE International Symposium on*, 2015, pp. 139-146: IEEE.

[51] N. Onizawai and T. Hanyu, "Soft-error tolerant transistor/magnetic-tunnel-junction hybrid non-volatile C-element," *IEICE Electronics Express,* vol. 11, no. 24, pp. 20141017-20141017, 2014.

[52] C. M. Liyanagedera, A. Sengupta, A. Jaiswal, and K. Roy, "Magnetic tunnel junction enabled stochastic spiking neural networks: From non-telegraphic to telegraphic switching regime," *arXiv preprint arXiv:.09247,* 2017.

[53] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, and Y. Nakamura, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science,* vol. 345, no. 6197, pp. 668-673, 2014.

[54] P. Wijesinghe, A. Ankit, A. Sengupta, and K. Roy, "An All-Memristor Deep Spiking Neural Computing System: A Step Toward Realizing the Low-Power Stochastic Brain," *IEEE Transactions on Emerging Topics in Computational Intelligence,* vol. 2, no. 5, pp. 345-358, 2018.

[55] M. Suri, D. Querlioz, O. Bichler, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Bio-inspired stochastic computing using binary CBRAM synapses," *IEEE Transactions on Electron Devices,* vol. 60, no. 7, pp. 2402-2409, 2013.

[56] J. Bill and R. Legenstein, "A compound memristive synapse model for statistical learning through STDP in spiking neural networks," *Frontiers in neuroscience,* vol. 8, p. 412, 2014.

[57] G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic tunnel junction based long-term short-term stochastic synapse for a spiking neural network with on-chip STDP learning," *Scientific reports,* vol. 6, p. 29545, 2016.

[58] 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. 031014, 2017.

[59] P. Debashis, R. Faria, K. Y. Camsari, and Z. Chen, "Design of Stochastic Nanomagnets for Probabilistic Spin Logic," *IEEE Magnetics Letters,* vol. 9, pp. 1-5, 2018.

[60] B. Sutton, K. Y. Camsari, B. Behin-Aein, and S. Datta, "Intrinsic optimization using stochastic nanomagnets," *Scientific Reports,* vol. 7, p. 44370, 2017.

[61] T. A. Gosavi, S. Manipatruni, S. V. Aradhya, G. E. Rowlands, D. Nikonov, I. A. Young, and S. A. Bhave, "Experimental Demonstration of Efficient Spin–Orbit Torque Switching of an MTJ With Sub-100 ns Pulses," *IEEE Transactions on Magnetics,* vol. 53, no. 9, pp. 1-7, 2017.

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

[63] N. Locatelli, A. Mizrahi, A. Accioly, R. Matsumoto, A. Fukushima, H. Kubota, S. Yuasa, V. Cros, L. G. Pereira, and D. Querlioz, "Noise-enhanced synchronization of stochastic magnetic oscillators," *Physical Review Applied,* vol. 2, no. 3, p. 034009, 2014.

[64] 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*, 2016, pp. 34.3. 1-34.3. 4: IEEE.

[65] 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.

[66] D. Bromberg, M. Moneck, V. Sokalski, J. Zhu, L. Pileggi, and J.-G. Zhu, "Experimental demonstration of four-terminal magnetic logic device with separate read-and write-paths," in *Electron Devices Meeting (IEDM), 2014 IEEE International*, 2014, pp. 33.1. 1-33.1. 4: IEEE.

[67] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm design exploration," in *Proceedings of the 7th international Symposium on Quality Electronic Design*, 2006, pp. 585-590: IEEE Computer Society.

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

[69] W. Wang and C. Chien, "Voltage-induced switching in magnetic tunnel junctions with perpendicular magnetic anisotropy," *Journal of Physics D: Applied Physics,* vol. 46, no. 7, p. 074004, 2013.

[70] S. M. Nowick and M. Singh, "High-performance asynchronous pipelines: an overview," *IEEE Design & Test of Computers,* vol. 28, no. 5, pp. 8-22, 2011.

[71] S. Roy, P. M. Mattheakis, L. Masse-Navette, and D. Z. Pan, "Clock tree resynthesis for multi-corner multi-mode timing closure," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 34, no. 4, pp. 589-602, 2015.

[72] A. Yakovlev, P. Vivet, and M. Renaudin, "Advances in asynchronous logic: From principles to GALS & NoC, recent industry applications, and commercial CAD tools," in *Proceedings of the Conference on Design, Automation and Test in Europe*, 2013, pp. 1715-1724: EDA Consortium.

[73] M. Shams, J. C. Ebergen, and M. I. Elmasry, "Modeling and comparing CMOS implementations of the C-element," *IEEE transactions on very large scale integration (VLSI) systems,* vol. 6, no. 4, pp. 563-567, 1998.

[74] 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 CoFeB/MgO/CoFeB pseudo-spin-valves annealed at high temperature," *Applied Physics Letters,* vol. 93, no. 8, p. 2508, 2008.

[75] A. Lecoutre, B. Negrevergne, and F. Yger, "Recognizing Art Style Automatically in painting with deep learning," in *Asian Conference on Machine Learning*, 2017, pp. 327-342.

[76] J. Tang, D. Sun, S. Liu, and J.-L. Gaudiot, "Enabling deep learning on iot devices," *Computer,* vol. 50, no. 10, pp. 92-96, 2017.

[77] S. Angizi and D. Fan, "IMC: energy-efficient in-memory convolver for accelerating binarized deep neural network," in *Proceedings of the Neuromorphic Computing Symposium*, 2017, p. 3: ACM.

[78] L. Ni, Z. Liu, H. Yu, and R. V. Joshi, "An energy-efficient digital ReRAM-crossbar-based CNN with bitwise parallelism," *IEEE Journal on Exploratory Solid-State Computational Devices,* vol. 3, pp. 37-46, 2017.

[79] X. Sun, X. Peng, P.-Y. Chen, R. Liu, J.-s. Seo, and S. Yu, "Fully parallel RRAM synaptic array for implementing binary neural network with (+ 1,− 1) weights and (+ 1, 0) neurons," in *Proceedings of the 23rd Asia and South Pacific Design Automation Conference*, 2018, pp. 574-579: IEEE Press.

[80] X. Sun, S. Yin, X. Peng, R. Liu, J.-s. Seo, and S. Yu, "XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks," *algorithms,* vol. 2, p. 3, 2018.

[81] S. D. Pyle, J. D. Sapp, and R. F. DeMara, "Leveraging Stochasticity for In Situ Learning in Binarized Deep Neural Networks," *Computer, In Press,* 2019. © 2019 IEEE

[82] S. Sinha, G. Yeric, V. Chandra, B. Cline, and Y. Cao, "Exploring sub-20nm FinFET design with predictive technology models," in *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*, 2012, pp. 283-288: IEEE.

[83] K. Y. Camsari, S. Ganguly, and S. Datta, "Modular approach to spintronics," *Scientific reports,* vol. 5, p. 10571, 2015.

[84] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Technical report2009, vol. 1.

[85] X. Huang, H. Wu, D. C. Sekar, S. N. Nguyen, K. Wang, and H. Qian, "Optimization of TiN/TaOx/HfO2/TiN RRAM arrays for improved switching and data retention," in *Memory Workshop (IMW), 2015 IEEE International*, 2015, pp. 1-4: IEEE.

[86] S. D. Pyle, K. Y. Camsari, and R. F. DeMara, "Hybrid spin-CMOS stochastic spiking neuron for high-speed emulation of In vivo neuron dynamics," *IET Computers and Digital Techniques,* vol. 12, no. 4, pp. 122-129, 2018.

[87] S. Wu, S. A. Cybart, D. Yi, J. M. Parker, R. Ramesh, and R. Dynes, "Full electric control of exchange bias," *Physical review letters,* vol. 110, no. 6, p. 067202, 2013.

[88] A. Aggarwal and T. Horiuchi, "Neuromorphic VLSI second-order synapse," *Electronics Letters,* vol. 51, no. 4, pp. 319-321, 2015.

[89] 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 *Electron Devices Meeting (IEDM), 2014 IEEE International*, 2014, pp. 12.5. 1-12.5. 4: IEEE.

[90] 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. 055934, 2017.

[91] 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. 055903, 2018.

[92] D. Vodenicarevic, N. Locatelli, A. Mizrahi, J. S. Friedman, A. F. Vincent, M. Romera, A. Fukushima, K. Yakushiji, H. Kubota, and S. Yuasa, "Low-energy truly random number generation with superparamagnetic tunnel junctions for unconventional computing," *Physical Review Applied,* vol. 8, no. 5, p. 054045, 2017.

[93] B. Nessler, M. Pfeiffer, L. Buesing, and W. Maass, "Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity," *PLoS computational biology,* vol. 9, no. 4, p. e1003037, 2013.

[94] D. Querlioz, O. Bichler, A. F. Vincent, and C. Gamrat, "Bioinspired programming of memory devices for implementing an inference engine," *Proceedings of the IEEE,* vol. 103, no. 8, pp. 1398-1416, 2015.

[95] A. F. Vincent, J. Larroque, N. Locatelli, N. B. Romdhane, O. Bichler, C. Gamrat, W. S. Zhao, J.-O. Klein, S. Galdin-Retailleau, and D. Querlioz, "Spin-transfer torque magnetic memory as a stochastic memristive synapse for neuromorphic systems," *IEEE transactions on biomedical circuits and systems,* vol. 9, no. 2, pp. 166-174, 2015.

[96] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H.-S. P. Wong, "Stochastic learning in oxide binary synaptic device for neuromorphic computing," *Frontiers in neuroscience,* vol. 7, p. 186, 2013.

[97] B. Ermentrout, M. Pascal, and B. Gutkin, "The effects of spike frequency adaptation and negative feedback on the synchronization of neural oscillators," *Neural Computation,* vol. 13, no. 6, pp. 1285-1310, 2001.

[98] D. A. McCormick, B. W. Connors, J. W. Lighthall, and D. A. Prince, "Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex," *Journal of neurophysiology,* vol. 54, no. 4, pp. 782-806, 1985.

[99] F. Baroni and A. Mazzoni, "Heterogeneity of heterogeneities in neuronal networks," *Frontiers in computational neuroscience,* vol. 8, p. 161, 2014.

[100] L. E. Dobrunz and C. F. Stevens, "Heterogeneity of release probability, facilitation, and depletion at central synapses," *Neuron,* vol. 18, no. 6, pp. 995-1008, 1997.

[101] S. Fisher, A. Teman, D. Vaysman, A. Gertsman, O. Yadid-Pecht, and A. Fish, "Digital subthreshold logic design-motivation and challenges," in *IEEE 25th Convention of Electrical and Electronics Engineers in Israel, 2008.*, 2008, pp. 702-706: IEEE.

[102] A. Sengupta and K. Roy, "Encoding neural and synaptic functionalities in electron spin: A pathway to efficient neuromorphic computing," *Applied Physics Reviews,* vol. 4, no. 4, p. 041105, 2017.

[103] D. Kappel, R. Legenstein, S. Habenschuss, M. Hsieh, and W. Maass, "A dynamic connectome supports the emergence of stable computational function of neural circuits through reward-based learning," *eNeuro,* vol. 5, no. 2, pp. ENEURO. 0301-17.2018, 2018.

[104] G. G. Turrigiano and S. B. Nelson, "Hebb and homeostasis in neuronal plasticity," *Current opinion in neurobiology,* vol. 10, no. 3, pp. 358-364, 2000.

[105] F. Zenke, G. Hennequin, and W. Gerstner, "Synaptic plasticity in neural networks needs homeostasis with a fast rate detector," *PLoS computational biology,* vol. 9, no. 11, p. e1003330, 2013.

[106] F. Jug, "On competition and learning in cortical structures," ETH Zurich, 2012.

[107] M. Stimberg, D. F. Goodman, V. Benichoux, and R. Brette, "Equation-oriented specification of neural models for simulations," *Frontiers in neuroinformatics,* vol. 8, p. 6, 2014.

[108] T. Devolder, J. Hayakawa, K. Ito, H. Takahashi, S. Ikeda, P. Crozat, N. Zerounian, J.-V. Kim, C. Chappert, and H. J. P. r. l. Ohno, "Single-shot time-resolved measurements of nanosecond-scale spin-transfer induced switching: Stochastic versus deterministic aspects," vol. 100, no. 5, p. 057206, 2008.

[109] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology,* vol. 148, no. 3, pp. 574-591, 1959.

[110] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research,* vol. 15, no. 1, pp. 1929-1958, 2014.